

Part No. 313197-D Rev 00
August 2004

4655 Great American Parkway
Santa Clara, CA 95054

Network Design Guidelines

Passport 8000 Series Software Release 3.7
Implementation Notes



NORTEL
NETWORKS™

Copyright © 2004 Nortel Networks

All rights reserved. August 2004.

The information in this document is subject to change without notice. The statements, configurations, technical data, and recommendations in this document are believed to be accurate and reliable, but are presented without express or implied warranty. Users must take full responsibility for their applications of any products specified in this document. The information in this document is proprietary to Nortel Networks Inc.

The software described in this document is furnished under a license agreement and may be used only in accordance with the terms of that license. The software license agreement is included in this document.

Trademarks

Nortel Networks, the Nortel Networks logo, the Globemark, Unified Networks, OPTera, and BayStack are trademarks of Nortel Networks.

Adobe and Acrobat Reader are trademarks of Adobe Systems Incorporated.

Microsoft, Windows, and Windows NT are trademarks of Microsoft Corporation.

Netscape and Navigator are trademarks of Netscape Communications Corporation.

UNIX is a trademark of X/Open Company Limited.

The asterisk after a name denotes a trademarked item.

Restricted rights legend

Use, duplication, or disclosure by the United States Government is subject to restrictions as set forth in subparagraph (c)(1)(ii) of the Rights in Technical Data and Computer Software clause at DFARS 252.227-7013.

Notwithstanding any other license agreement that may pertain to, or accompany the delivery of, this computer software, the rights of the United States Government regarding its use, reproduction, and disclosure are as set forth in the Commercial Computer Software-Restricted Rights clause at FAR 52.227-19.

Statement of conditions

In the interest of improving internal design, operational function, and/or reliability, Nortel Networks Inc. reserves the right to make changes to the products described in this document without notice.

Nortel Networks Inc. does not assume any liability that may occur due to the use or application of the product(s) or circuit layout(s) described herein.

Portions of the code in this software product may be Copyright © 1988, Regents of the University of California. All rights reserved. Redistribution and use in source and binary forms of such portions are permitted, provided that the above copyright notice and this paragraph are duplicated in all such forms and that any documentation, advertising materials, and other materials related to such distribution and use acknowledge that such portions of the software were developed by the University of California, Berkeley. The name of the University may not be used to endorse or promote products derived from such portions of the software without specific prior written permission.

SUCH PORTIONS OF THE SOFTWARE ARE PROVIDED "AS IS" AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, WITHOUT LIMITATION, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE.

In addition, the program and information contained herein are licensed only pursuant to a license agreement that contains restrictions on use and disclosure (that may incorporate by reference certain limitations and notices imposed by third parties).

Nortel Networks Inc. software license agreement

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

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

1. Licensed Use of Software. Nortel Networks grants Customer a nonexclusive license to use a copy of the Software on only one machine at any one time or to the extent of the activation or authorized usage level, whichever is applicable. To the extent Software is furnished for use with designated hardware or Customer furnished equipment (“CFE”), Customer is granted a nonexclusive license to use Software only on such hardware or CFE, as applicable. Software contains trade secrets and Customer agrees to treat Software as confidential information using the same care and discretion Customer uses with its own similar information that it does not wish to disclose, publish or disseminate. Customer will ensure that anyone who uses the Software does so only in compliance with the terms of this Agreement. Customer shall not a) use, copy, modify, transfer or distribute the Software except as expressly authorized; b) reverse assemble, reverse compile, reverse engineer or otherwise translate the Software; c) create derivative works or modifications unless expressly authorized; or d) sublicense, rent or lease the Software. Licensors of intellectual property to Nortel Networks are beneficiaries of this provision. Upon termination or breach of the license by Customer or in the event designated hardware or CFE is no longer in use, Customer will promptly return the Software to Nortel Networks or certify its destruction. Nortel Networks may audit by remote polling or other reasonable means to determine Customer’s Software activation or usage levels. If suppliers of third party software included in Software require Nortel Networks to include additional or different terms, Customer agrees to abide by such terms provided by Nortel Networks with respect to such third party software.

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

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

4. General

- a. If Customer is the United States Government, the following paragraph shall apply: All Nortel Networks Software available under this License Agreement is commercial computer software and commercial computer software documentation and, in the event Software is licensed for or on behalf of the United States Government, the respective rights to the software and software documentation are governed by Nortel Networks standard commercial license in accordance with U.S. Federal Regulations at 48 C.F.R. Sections 12.212 (for non-DoD entities) and 48 C.F.R. 227.7202 (for DoD entities).
- b. Customer may terminate the license at any time. Nortel Networks may terminate the license if Customer fails to comply with the terms and conditions of this license. In either event, upon termination, Customer must either return the Software to Nortel Networks or certify its destruction.
- c. Customer is responsible for payment of any taxes, including personal property taxes, resulting from Customer's use of the Software. Customer agrees to comply with all applicable laws including all applicable export and import laws and regulations.
- d. Neither party may bring an action, regardless of form, more than two years after the cause of the action arose.
- e. The terms and conditions of this License Agreement form the complete and exclusive agreement between Customer and Nortel Networks.
- f. This License Agreement is governed by the laws of the country in which Customer acquires the Software. If the Software is acquired in the United States, then this License Agreement is governed by the laws of the state of New York.

Contents

Preface	27
Before you begin	27
Text conventions	28
Acronyms	28
Hard-copy technical manuals	34
How to get help	34
Chapter 1	
General network design considerations	37
Hardware considerations	37
CPU memory upgrade	38
E- and M-modules	38
10 Gigabit Ethernet	39
Overview	39
10GE to 1GE comparison	40
10GE WAN	41
Design constraints	43
Hardware record optimization	46
Record reservation	46
8692SF module	48
Electrical considerations	48
Software considerations	49
Chapter 2	
Designing redundant networks	53
General considerations	53
Network reliability and availability	54
Physical layer	55

Ethernet cable distances	56
Transmission distance and optical link budget	58
IEEE 802.3ab Gigabit Ethernet- copper cabling	58
Auto-Negotiation for Ethernet 10/100 BASE Tx	59
100BASE-FX failure recognition/ far end fault indication	60
Gigabit and remote fault indication	61
Using single fiber fault detection (SFFD) for remote fault indication	62
Configuring SFFD using the CLI	64
VLACP	64
Platform redundancy	67
HA mode	69
Link redundancy	71
MLT	71
Switch-to-switch links	72
Routed links	72
MLT and STG	73
MLT traffic distribution algorithm	73
Path cost implementation notes	74
IEEE 802.3ad-based link aggregation (IEEE 802.3 2002 clause 43)	75
Overview	76
LACP	77
Link aggregation operation	78
Principles of link aggregation	79
LACP and MLT	80
LACP and spanning tree interaction	81
Link aggregation rules	81
Link aggregation examples	82
Switch-to-switch example	83
Switch-to-server MLT example	84
Client/server MLT example	85
Network redundancy	86
Basic network layouts- physical structure for redundant networks	86
Redundant network edge	90
Recommended and not recommended network edge designs	91
SMLT	92

Overview	93
IST link	94
CP-Limit considerations with SMLT IST	95
SMLT links	96
SMLT ID configuration	98
Supported SMLT links	98
Single port SMLT	98
Interaction between SMLT and IEEE 802.3ad	102
Layer 2 traffic load sharing	103
Layer 3 traffic load sharing	103
Failure scenarios	104
SMLT designs	106
SMLT and Spanning Tree	110
SMLT scalability	110
RSMLT	112
SMLT/RSMLT operation in L3 environments	112
Failure scenarios	113
Designing and configuring an RSMLT network	115
Network design examples	115
Layer 1 examples	115
Layer 2 examples	118
Layer 3 examples	121
Spanning tree protocol	124
STGs and BPDU forwarding	124
Multiple STG interoperability with single STG devices	125
The problem	125
The solution	126
Create two STGs and set MAC addresses for the STGs	127
Configure STG roots	128
Configure VLANs	128
PVST+	129
Passport 8600 PVST+ implementation and guidelines	131
Using MLT to protect against split VLANs	132
Isolated VLANs	132

Chapter 3	
Designing stacked VLAN networks	135
About stacked VLAN	135
Features	136
sVLAN operation	137
Components	138
Switch levels	139
IEEE 802.1Q tag	139
UNI port behavior	140
NNI port behavior	140
sVLAN and SMLT	141
UNI ports and SMLT	141
NNI ports and SMLT	142
Network loop detection and prevention	142
sVLAN multi-level onion architecture	144
Network level requirements	146
Independent VLAN learning limitation	146
sVLAN and network or device management	147
sVLAN restrictions	147
Chapter 4	
Designing Layer 3 switched networks	149
VRRP	150
VRRP and other routing protocols	150
VRRP and STG	151
ICMP redirect messages	152
Avoiding excessive ICMP redirect messages	152
Subnet-based VLANs	155
Subnet-based VLAN and IP routing	155
Subnet based VLAN and VRRP	155
Subnet-based VLAN and multinetting	156
Subnet-based VLAN and DHCP	156
Subnet-based VLAN scalability	156
Subnet-based VLAN and wireless terminals	156
PPPoE protocol-based VLAN design	157

Implementing bridged PPPoE and IP traffic isolation	157
Indirect connections	160
Direct connections	162
BGP	163
Overview	164
Hardware and software dependencies	164
Scaling considerations	165
Design scenarios	166
Internet peering	166
BGP applications to connect to an AS	167
Edge aggregation	167
ISP segmentation	168
Multiple regions separated by EBGp	169
Multi-homed to non-transit AS/single provider	170
Considerations	170
Interoperability	171
OSPF	172
Scalability guidelines	172
Design guidelines	173
OSPF route summarization and black hole routes	173
OSPF network design scenarios	174
Scenario 1: OSPF on one subnet in one area	174
Scenario 2: OSPF on two subnets in one area	176
Scenario 3: OSPF on two subnets in two areas	177
IPX	180
GNS	180
LLC encapsulation and translation	181
IPX RIP/SAP policies	181
IP routed interface scaling considerations	182
Chapter 5	
Enabling Layer 4-7 application services	183
Introduction	183
Layer 4-7 switching	184
Layer 4-7 switching in the Passport 8600 environment	185

WSM location	185
WSM components	186
WSM architecture	187
Passport default parameters and settings	188
WSM default parameters	190
Applications and services	192
Local server load balancing	192
Health checking metrics	194
GSLB	196
Application redirection	197
VLAN filtering	198
Application abuse protection	199
Layer 7 deny filters	200
Network problems addressed by the WSM	201
Network architectures	202
Using the Passport 8600 as a Layer 2 switch	202
Leveraging Layer 3 routing in the Passport 8600	203
Implementing L4-7 services with a single Passport 8600	204
Implementing L4-7 services with dual Passport 8600s	205
Architectural details and limitations	206
User and password management	207
Passport unknown MAC discard	209
Syslog	210
Image management	210
SNMP and MIB management	211
Console and management support	212
WAN link load balancing	212
VRRP hot standby	213
Chapter 6	
Designing multicast networks	215
Multicast handling in the Passport 8600	215
Multicast and MLT	216
DVMRP or PIM route tuning to load share streams	217
Multicast flow distribution over MLT	219

IP multicast scaling	221
DVMRP scalability	221
Interface scaling	221
Route scaling	222
Stream scaling	222
PIM-SM and PIM-SSM scalability	222
Interface scaling	223
Route scaling	223
Stream scaling	224
Improving multicast scalability	224
General IP multicast rules and considerations	226
IP multicast address ranges	226
IP to Ethernet multicast MAC mapping	227
Dynamic configuration changes	229
DMVRP IGMPv2 back-down to IGMPv1	230
TTL in IP multicast packets	230
Multicast MAC filtering	232
Multicast filtering and multicast access control	233
New release 3.5 multicast access control policies	233
Multicast access policies before release 3.5	234
Guidelines for multicast access policies	235
Split-subnet and multicast	236
IGMP and routing protocol interactions	237
IGMP and DVMRP	237
IGMP and PIM-SM	238
IGMP and PIM-SSM	239
DVMRP general design rules	240
General network design	240
Sender and receiver placement	241
DVMRP timers tuning	241
DVMRP policies	242
Announce and accept policies	242
Do not advertise self	245
Default route policies	246
DVMRP passive interface	247

General design considerations with PIM-SM	248
General requirements	249
SPT switchover	250
Recommended MBR configuration	251
Redundant MBR configuration	252
MBR and DVMRP path cost considerations	255
PIM passive interface	255
Circuitless IP for PIM-SM	255
Static RP	256
Auto-RP protocol	256
RP redundancy	257
Non-supported static RP configuration	260
RP placement	260
BSR hash algorithm	260
RP and extended VLANs	264
Receivers on interconnected VLANs	264
PIM network with non-PIM interfaces	265
Multicast and SMLT	266
Triangle designs	267
All Layer 2 IGMP snooping	267
Layer 2 and Layer 3 multicast	268
Square designs	269
Design that avoids duplicate traffic	270
DVMRP versus PIM	272
Flood and prune versus shared and source trees	272
Unicast routes for PIM versus DMVRP own routes	273
Convergence and timers	274
Traffic delay with PIM while rebooting peer SMLT switches	274
Enabling multicast on network interfaces	275
Reliable multicast specifics	275
Protocol timers	275
PGM-based designs	276
Multicast stream initialization	277
TV delivery and multimedia applications	277
Static (S,G)s with DVMRP and IGMP static receivers	278

Join/leave performance	278
Fast leave	279
LMQI tuning	280
IGAP	281
PIM-SSM and IGMPv3	284
IGMPv3 and PIM-SSM design considerations	284
PIM-SSM design considerations	285
Chapter 7	
Designing secure networks	287
Denial of service attacks	288
Malicious code	288
Attacks to resiliency and availability	289
Additional information and references	289
Implementing security measures with the Passport 8600	290
Passport 8600 DoS protection mechanisms	290
Broadcast/Multicast rate limiting	290
Directed broadcast suppression	291
Prioritization of control traffic	291
Control traffic limitation	291
ARP limitation	292
Multicast learning limitation	293
Passport 8600 damage prevention mechanisms	293
Stopping spoofed IP packets	294
Preventing the network from being used as a broadcast amplification site ..	295
High secure mode (CLI)	295
Passport 8600 security against malicious code	296
Passport 8600 security against resiliency and availability attacks	299
Passport 8600 access protection mechanisms	299
Data plane	300
Extended authentication protocol- 802.1x	300
Traffic isolation: VLANs	303
Filtering capabilities	303
Routing policies (announce/accept policies)	307
OSPF	308

BGP	308
Control plane	309
Management	309
High secure mode (bootconfig)	311
Management access control	311
Access policies	314
Authentication	314
Encryption of control plane traffic	317
Modifying the RADIUS/SNMP header network address	319
SNMPv3 support in release 3.3 and 3.7	319
SNMP community string encryption	320
Other platforms and equipment	320
Chapter 8	
Connecting Ethernet networks to WAN networks	325
Engineering considerations	325
ATM scalability	325
Performance	326
ATM resiliency	327
F5 OAM loopback request/reply	328
Feature considerations	329
ATM and MLT	329
ATM and 802.1q tags	329
ATM and DiffServ	330
ATM and IP multicast	330
Shaping	332
Applications considerations	332
ATM WAN connectivity and OE/ATM interworking	333
Point-to-point WAN connectivity	333
Service provider solutions – OE/ATM interworking	334
OE/ATM interworking- A detailed look	335
Transparent LAN services	337
Video over DSL over ATM	338
Point-to-multipoint configuration for video over DSL over ATM	339
Point-to-point configuration for video over DSL over ATM	339

ATM and voice applications	339
Design recommendations	340
ATM latency testing results	340
Chapter 9	
Provisioning QoS networks	341
Combining IP filtering and DiffServ features	342
IP filtering and ARP	342
IP filtering and forwarding decisions	343
Global filters	343
Source/destination filters	343
IP filter ID	344
Per-hop behaviors	344
Admin weights for traffic queues	344
DiffServ interoperability with Layer 2 switches	345
DiffServ access ports in drop mode	346
Quality of Service overview	346
Nortel Networks QoS strategy	348
Traffic classification	348
Class of service mapping to standards	350
Passport 8600 QoS mechanisms	350
QoS highlights	351
Internal QoS level	352
Emission priority queuing and drop precedence	352
Packet classification	355
Filtering	357
Policing and rate metering	357
Passport 8600 network QoS	358
Trusted vs. untrusted interfaces	359
Access vs. core port	359
Bridged vs. routed traffic	361
Tagged vs. untagged packets	362
QoS summary	364
QoS and filtering	366
Filtering	366

DiffServ access port (IP bridged traffic with DiffServ enabled)	366
Source MAC-based VLANs	368
Protocol-based/IP subnet-based VLANs	369
Core port (IP bridged traffic)	369
Port-based VLANs	369
Non-IP traffic (bridged or L2)	370
Port-based VLANs	370
Protocol-based VLANs	370
Source MAC-based VLANs	370
DiffServ access (IP routed traffic)	370
DiffServ core (IP routed traffic)	371
QoS flow charts	371
QoS and network congestion	375
No congestion	376
Momentary bursts of congestion	376
Severe congestion	378
QoS network scenarios	380
Scenario 1 – bridged traffic	380
Case 1 – Customer traffic is trusted	380
Case 2 – Customer traffic is untrusted	383
Case 3– RPR interworking	383
Scenario 2 – routed traffic	384
Case 1 – Customer traffic is trusted	384
Chapter 10	
Managing Passport 8000 Series switches	387
Offline switch configuration	387
Port mirroring	388
Local port mirroring	388
Identifying E-modules	388
Mirroring scalability	389
Remote mirroring	390
pcmboot.cfg	391
Default management IP address	392
Backup configuration files	392

DNS client	392
Appendix A	
QoS algorithm	393
Appendix B	
Scaling numbers	395
Appendix C	
Hardware and supporting software compatibility.....	399
Appendix D	
Tap and OctaPID assignment	403
Index	409

Figures

Figure 1	Basic WAN and MAN applications for 10GE	40
Figure 2	Hardware and software reliability	54
Figure 3	Auto-Negotiation process	60
Figure 4	100BASE-FX FEFI	61
Figure 5	Problem description (1 of 2)	65
Figure 6	Problem description (2 of 2)	66
Figure 7	Link Aggregation Sublayer example (according to IEEE 802.3ad)	77
Figure 8	Switch-to-switch MLT configuration	83
Figure 9	Switch-to-server MLT configuration	84
Figure 10	Client/Server MLT configuration	85
Figure 11	Four-tiered network layout	87
Figure 12	Three-tiered network layout	88
Figure 13	Two- or three-tiered networks with collapsed aggregation and core layer	89
Figure 14	Redundant network edge diagram	90
Figure 15	Recommended network edge design	91
Figure 16	Not recommended network edge design	92
Figure 17	SMLT configuration with 8600 switches as aggregation switches	94
Figure 18	Single port SMLT example	100
Figure 19	Changing a split trunk from MLT-based SMLT to single port SMLT	101
Figure 20	SMLT scaling design	107
Figure 21	SMLT triangle configuration	108
Figure 22	SMLT square configuration	109
Figure 23	SMLT full mesh configuration	110
Figure 24	SMLT and RSMLT in L3 environments	114
Figure 25	Layer 1 design examples	116
Figure 26	Layer 2 design examples	119
Figure 27	Layer 3 design examples	122
Figure 28	One STG between two Layer 3 devices and one Layer 2 device	125
Figure 29	Alternative configuration for STG and Layer 2 devices	127

Figure 30	VLANs on the Layer 2 switch	129
Figure 31	802.1d Spanning tree	130
Figure 32	VLAN isolation	133
Figure 33	Provider bridging / sVLAN operation	137
Figure 34	IEEE 802.1Q tag	140
Figure 35	Dual-homing of CPE to sVLAN UNI ports	141
Figure 36	SMLT full mesh core for sVLAN provider network	142
Figure 37	Customer traffic loops through a service provider core	143
Figure 38	One-level sVLAN design	144
Figure 39	Two-level sVLAN design	145
Figure 40	Multi-level onion design sVLAN with Q tags	146
Figure 41	Sharing the same IP address	150
Figure 42	VRRP and STG configurations	151
Figure 43	ICMP redirect messages diagram	152
Figure 44	Avoiding excessive ICMP redirect messages- option 1	153
Figure 45	Avoiding excessive ICMP redirect messages- option 2	154
Figure 46	Avoiding excessive ICMP redirect messages- option 3	155
Figure 47	PPPoE and IP traffic separation	159
Figure 48	Indirect PPPoE and IP configuration	161
Figure 49	Direct PPPoE and IP configuration	163
Figure 50	Internet peering	166
Figure 51	BGP's role to connect to an AS	167
Figure 52	Edge aggregation	167
Figure 53	Multiple regions separated by IBGP	168
Figure 54	Multiple regions separated by EBGP	169
Figure 55	Multiple OSPF regions peering with the Internet	170
Figure 56	Enabling OSPF on one subnet in one area	174
Figure 57	Configuring OSPF on two subnets in one area	176
Figure 58	Configuring OSPF on two subnets in two areas	177
Figure 59	WSM's role as an intelligent module	186
Figure 60	WSM ports	186
Figure 61	WSM data path architecture	188
Figure 62	Detailed WSM data path architecture	191
Figure 63	Single WSM default architecture	192
Figure 64	Metric selection process	195

Figure 65	Browser-based application redirection	198
Figure 66	VLAN filtering	199
Figure 67	Application abuse protection	200
Figure 68	The Passport 8600 as a Layer 2 switch	203
Figure 69	Layer 3 routing in the Passport 8600	204
Figure 70	Multiple WSMs using a single Passport 8600	205
Figure 71	Dual chassis high availability	206
Figure 72	Traffic distribution for multicast data	218
Figure 73	Multicast flow distribution over MLT	220
Figure 74	IP multicast sources and receivers on interconnected VLANs	225
Figure 75	Multicast IP address to MAC address mapping	228
Figure 76	Passport 8600 Switches and IP multicast traffic with low TTL	231
Figure 77	Applying IP Multicast access policies for DVMRP	236
Figure 78	IGMP interaction with DVMRP	238
Figure 79	IGMP interaction with PIM	239
Figure 80	Announce policy on a border router	243
Figure 81	Accept policy on a border router	244
Figure 82	Load balancing with announce policies	245
Figure 83	Do not advertise local route policies	246
Figure 84	Default route	247
Figure 85	MBR configuration	252
Figure 86	Redundant MBR configuration	253
Figure 87	Redundant MBR configuration with two separate VLANs	254
Figure 88	RP failover with default unicast routes	258
Figure 89	Interface address selection on the RP	259
Figure 90	Inefficient group-RP mapping	261
Figure 91	Receivers on interconnected VLANs	265
Figure 92	PIM network with non-PIM interfaces	266
Figure 93	Layer 2 IGMP snooping	268
Figure 94	Multicast routing using DVMRP or PIM	268
Figure 95	Square design- full mesh configuration	269
Figure 96	Multicast and SMLT design that avoids duplicate traffic	270
Figure 97	Avoiding an interruption of IGAP traffic	282
Figure 98	IDS server configuration	297
Figure 99	Alteon web switch family IDS server configuration	298

Figure 100	802.1x and OPS interaction	301
Figure 101	Traffic discard process	303
Figure 102	Dedicated Ethernet management link	310
Figure 103	Terminal servers/modem access	310
Figure 104	Access levels	314
Figure 105	RADIUS server as proxy for stronger authentication	316
Figure 106	Authentication encryption	318
Figure 107	Firewall load balancing configuration	321
Figure 108	Network with and without MLT	327
Figure 109	ATM network broken PVCs	328
Figure 110	Point-to-multipoint IP multicast	330
Figure 111	IP multicast traffic over ATM	331
Figure 112	Bringing remote sites into an aggregation PoP	334
Figure 113	OE/ATM interworking- using home run PVCs	335
Figure 114	OE/ATM interworking- using RFC 1483 bridge termination	336
Figure 115	OE/ATM interworking- using RFC 1483 bridge termination with cVRs	337
Figure 116	Supported TLS configuration	337
Figure 117	Configuring PVCs in different VLANs on the same ATM port	338
Figure 118	Passport 8600 core vs. access ports	352
Figure 119	Passport 8600 queue structures	353
Figure 120	QoS filtering decisions	357
Figure 121	Passport 8600 access port	360
Figure 122	Passport 8600 core port	361
Figure 123	Passport QoS summary graphic	364
Figure 124	Untagged ingress traffic on the port-based VLANs:	367
Figure 125	Tagged ingress traffic on the port-based VLANs:	368
Figure 126	DiffServ access mode- port-based VLANs	372
Figure 127	DiffServ access mode- MAC-based VLANs	373
Figure 128	DiffServ access mode- IP subnet and protocol-based VLANs	374
Figure 129	DiffServ core mode	375
Figure 130	Congestion bursts	377
Figure 131	OctaPID queue buffers	378
Figure 132	Severe congestion	379
Figure 133	Trusted bridged traffic	381
Figure 134	Passport 8600 summary on bridged access ports	382

Figure 135	Passport 8600 summary on bridged or routed core ports	383
Figure 136	Passport 8600 to RPR QoS internetworking	384
Figure 137	Trusted routed traffic	385
Figure 138	Passport 8600 QoS summary on routed access ports	386
Figure 139	Remote mirroring	391
Figure 140	QoS algorithm	394

Tables

Table 1	1GE vs. 10GE comparison	41
Table 2	Recommended 10GE WAN interface clock settings	42
Table 3	Example MAC and IP addressing for best throughput	45
Table 4	Record reservation specifications	47
Table 5	Number of power supplies to install	49
Table 6	Software/hardware feature dependencies	50
Table 7	10/100 Ethernet cable distances	56
Table 8	Gigabit Ethernet cable distances for 1000BASE-TX	56
Table 9	Gigabit Ethernet standard minimum distance ranges	57
Table 10	Recommended Auto-Negotiation setting on 10/100BASE-TX ports	60
Table 11	Ethernet switching devices that do not support Auto-Negotiation	62
Table 12	HA failover phases	70
Table 13	Path cost default values using 1993 ANSI/IEEE 802.1D	74
Table 14	SMLT components	92
Table 15	sVLAN components	138
Table 16	Passport default parameters and settings	189
Table 17	WSM default parameters	190
Table 18	Health checking metrics	195
Table 19	Application redirection types	197
Table 20	Network problems addressed by the WSM	201
Table 21	Passport 8600 and WSM user access levels	208
Table 22	Recommended CP limit values	292
Table 23	Source addresses that need to be filtered	294
Table 24	Configuration actions	306
Table 25	OSPF packet	308
Table 26	8600 Series switch management access levels	311
Table 27	Nortel Networks QoS traffic classification	348
Table 28	Class of service mapping to standards	350
Table 29	Passport 8600 QoS defaults	354

Table 30	Passport 8600 PTO settings	354
Table 31	IEEE 802.1p bits to QoS level mapping	361
Table 32	DSCP to QoS level mapping	362
Table 33	QoS level to IEEE 802.1p and DSCP mapping	363
Table 34	Passport 8600 E-modules	389
Table 35	Scaling numbers for Release 3.7 features	395
Table 36	Available module types and OctapPID ID assignments	404
Table 37	8608GBE/8608GBM/8608GTE/8608GTM, and 8608SXE modules	405
Table 38	8616SXE module	405
Table 39	8624FXE module	406
Table 40	8632TXE and 8632TZM modules	406
Table 41	8648TXE and 8648TXM modules	406
Table 42	8672ATME and 8672ATMM modules	407
Table 43	8681XLR module	407
Table 44	8681XLW module	408
Table 45	8683POSM module	408

Preface

This document describes a range of design considerations and related procedures that will help you to optimize the performance and stability of your Passport 8000 Series switch network.

Before you begin

This guide is intended for network architects and administrators with the following background:

- Knowledge of networks, Ethernet bridging, and IP routing
- Familiarity with networking concepts and terminology
- Knowledge of network topologies

Text conventions

This guide uses the following text conventions:

angle brackets (< >)	Indicate that you choose the text to enter based on the description inside the brackets. Do not type the brackets when entering the command. Example: If the command syntax is <code>ping <ip_address></code> , you enter <code>ping 192.32.10.12</code>
<i>italic text</i>	Indicates new terms, book titles, and variables in command syntax descriptions. Where a variable is two or more words, the words are connected by an underscore. Example: If the command syntax is <code>show at <valid_route></code> , <i>valid_route</i> is one variable and you substitute one value for it.
plain Courier text	Indicates command syntax and system output, for example, prompts and system messages. Example: <code>Set Trap Monitor Filters</code>

Acronyms

The following table describes the acronyms that you encounter in this guide.

ABR	area boundary router
ADM	add/drop multiplexer
ADSL	asymmetric digital subscriber line
APS	automatic protection switching
ARP	Address Resolution Protocol
ARU	address resolution unit
AS	autonomous systems
ASIC	application specific integrated circuit
ATM	asynchronous transfer mode

BDR	backup designated router
BFM	backplane fabric module
BGP	Border Gateway Protocol
BPDU	bridge protocol data unit
BSAC	BaySecure Access Control
BSR	bootstrap router
CLI	command line interface
CODEC	coder-decoder
CoS	class of service
CPE	Customer Premise Equipment
CPU	central processing unit
CRC	cyclic redundancy check
CS	Computer Security Institute
DA	destination address
DHCP	Dynamic Host Configuration Protocol
DMLT	distributed multilink trunking
DoS	denial of service
DDoS	distributed denial of service
DNS	domain name server
DR	designated router
DSCP	differentiated services code point
DSL	digital subscriber line
DSLAM	digital subscriber line access multiplexer
DVMRP	Distance Vector Multicast Routing Protocol
DWDM	dense wavelength division multiplexing
EBGP	exterior BGP
ECMP	equal cost multipath
ELAN	emulated LAN (ATM)
FEFI	far end fault indication

FTP	File Transfer Protocol
Gbps	gigabits per second
GE	gigabit Ethernet
GNS	get nearest server
GSLB	global server load balancing
GUI	graphical user interface
HA	High Availability
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol, Secured
IBGP	interior BGP
ICMP	Internet Control Message Protocol
IDS	intrusion detection system
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IGAP	Internet Group membership Authentication Protocol
IGMP	Internet Group Management Protocol
IGP	Interior Gateway Protocol
IP	Internet Protocol
IPCP	Internet Protocol Control Packet
IPSEC	IP security
IPMC	IP multicast
IPX	Internetwork Packet Exchange
ISD	integrated service director
IST	inter-switch trunk
JDM	Java Device Manager
Kbps	kilobits per second
LACP	Link Access Control Protocol
LAG	Link Access Group
LAN	local area network

L2	Layer 2
L3	Layer 3
LB	load balancing
LDAP	Lightweight Directory Access Protocol
LLC	logical link control
LMQI	last member query interval
LSA	link state advertisement
MAC	media access control
MAN	metro area network
Mbps	megabits per second
MBS	maximum burst size
MDA	media dependent adapter
MD5	message digest 5
MIB	management information base
MLT	multilink trunk
MMF	multimode fiber
MPLS	Multiprotocol Label Switching
MRDISC	multicast router discovery
NAT	network address translation
NBMA	non-broadcast multiaccess
NIC	network interface card Network Information Center
NTP	network time protocol
OAM	operation, administration, and maintenance
OE	Optical Ethernet
OOB	out of band
OSPF	open shortest path first
PCR	peak cell rate
PCMCIA	Personal Computer Memory Card International Association

PE	Provider Edge
PGM	pragmatic general multicast
PHB	per-hop behavior
PHY	physical layer
PIM	protocol independent multicast
PIM-SM	protocol independent multicast, sparse mode
PIM-SSM	protocol independent multicast, source specific multicast
PIP	proxy IP address
PoP	point of presence
POS	Packet over SONET
PPP	point-to-point
PTO	packet transmission opportunities
PVC	private virtual circuit
PVID	port VLAN ID
QoS	quality of service
RADIUS	remote authentication dial-in user service
RAM	random access memory
RDI	remote defect indication
RIP	Routing Information Protocol
RISC	reduced instruction set computer
RMON	remote monitoring
RP	rendezvous point
RPR	restore path request
RPF	reverse path forwarding
RSMLT	routed SMLT
RTSP	Real-Time Streaming Protocol
SA	source address
SANS	System Administration, Networking and Security Institute
SAP	Service Advertisement Protocol

SCR	sustainable cell rate
SDH	synchronous digital hierarchy
SF	switch fabric
SLA	service level agreement
SLB	server load balancing
SMLT	Split Multi-Link Trunking
SNMP	Simple Network Management Protocol
SONET	synchronous optical network
SPT	shortest path tree
SSH	secure shell
SSL	secure socket layer
STG	spanning tree group
STP	Spanning Tree Protocol shielded twisted pair
TCP	Transmission Control Protocol
TCP/IP	Transmission Control Protocol over IP
TDM	time-division multiplexing
TFTP	Trivial File Transfer Protocol
TLS	Transparent LAN Services
ToS	type of service
TTL	time to live
UDP	User Datagram Protocol
URL	universal resource locator
UTP	unshielded twisted pair
VBR	variable bit rate
VC	virtual connection
VCG	virtual connection gateway
VIP	virtual IP
VLAN	virtual local area network

VoIP	voice over IP
VPN	virtual private network
VR	virtual router
VRRP	Virtual Router Redundancy Protocol
WAN	wide area network
WC	wiring closet
WMI	Web management interface
WRR	weighted round robin
WSM	Web Switching Module

Hard-copy technical manuals

You can print selected technical manuals and release notes free, directly from the Internet. Go to the www.nortelnetworks.com/documentation URL. Find the product for which you need documentation. Then locate the specific category and model or version for your hardware or software product. Use Adobe* Acrobat Reader* to open the manuals and release notes, search for the sections you need, and print them on most standard printers. Go to Adobe Systems at the www.adobe.com URL to download a free copy of the Adobe Acrobat Reader.

How to get help

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

If you purchased a Nortel Networks service program, contact Nortel Networks Technical Support. To obtain contact information online, go to the www.nortelnetworks.com/cgi-bin/comments/comments.cgi URL, then click on Technical Support.

From the Technical Support page, you can open a Customer Service Request online or find the telephone number for the nearest Technical Solutions Center. If you are not connected to the Internet, you can call 1-800-4NORTEL (1-800-466-7835) to learn the telephone number for the nearest Technical Solutions Center.

An Express Routing Code (ERC) is available for many Nortel Networks products and services. When you use an ERC, your call is routed to a technical support person who specializes in supporting that product or service. To locate an ERC for your product or service, go to the <http://www.nortelnetworks.com/help/contact/erc/index.html> URL.

Chapter 1

General network design considerations

This chapter provides general guidelines you should be aware of when designing your network. It includes the following sections:

Topic	Page number
Hardware considerations	next
Electrical considerations	49
Software considerations	49

Hardware considerations

The hardware considerations that support the Passport 8000 Series software (release 3.5 and above) include the following:

- [“CPU memory upgrade,”](#) next
- [“E- and M-modules”](#) on page 38
- [“10 Gigabit Ethernet”](#) on page 39
- [“Hardware record optimization”](#) on page 46
- [“Record reservation”](#) on page 46
- [“8692SF module”](#) on page 48

CPU memory upgrade

Nortel Networks offers a 256MB CPU Upgrade Kit (Part # DS1404016) for the 8190SM, 8690SF and the 8691SF CPUs.

- For the 8190SM and 8690SF, you **must** install the 256MB upgrade to support the Passport 8000 Series software release 3.5 and above.
- For the 8691SF, Nortel Networks recommends that you install the 256MB upgrade.

E- and M-modules

In addition to non-E- and M-modules, the Passport 8000 switch Series also supports E- and M-modules. The M-modules, or extended memory modules, were introduced in the Passport 8000 Series software release 3.3. They are designed to support large Layer 2 (bridging and/or multicast) and Layer 3 (more than 20,000 route) environments, or a combination of the two.

E-modules support 32K records, while M-modules support 128K records. A *record* can include the following:

- a media access control (MAC) entry
- a virtual local area network (VLAN) entry
- a multicast entry
- an Address Resolution Protocol (ARP) entry
- an Internet Protocol (IP) route entry
- a filter rule (IP filter)
- an Internetwork Packet Exchange (IPX) network entry



Note: M-modules are based on the E-module architecture. Thus, M-modules support all E-module features and characteristics. The *only* difference between the two is the added amount of memory necessary to support 128K records.

Passport 8000 Series software supports the following M-modules:

- 10 Gigabit Ethernet (GE) modules (WAN and LAN) including:
 - 8681XLW (DS1404052)
 - 8681XLR (DS1404053)
- POSM (DS1404060)
- ATMM (DS1304009)
- 8632TXM (DS1404055)
- 8648TXM (DS1404056)
- 8608GBM (DS1404059)
- 8608GTM (DS1404061)

[Table 35](#) in [Appendix B](#) provides scaling numbers for E- and M-modules.

10 Gigabit Ethernet

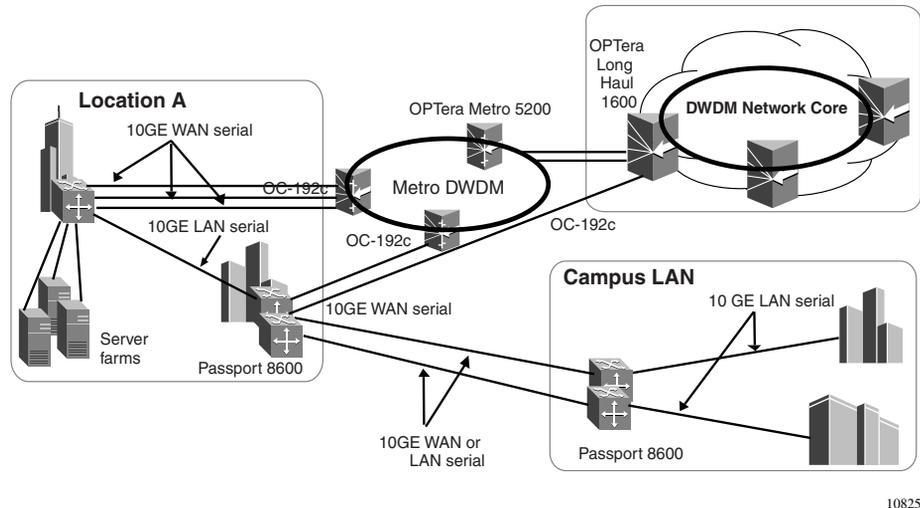
10 Gigabit Ethernet (10GE) PHY interfaces consist of both LAN and WAN interfaces. The following sections discuss 10GE in more detail:

- [“Overview,”](#) next
- [“10GE to 1GE comparison”](#) on page 40
- [“10GE WAN”](#) on page 41
- [“Design constraints”](#) on page 43

Overview

10GE provides an initial application in the point of presence (PoP), WAN, and MAN markets. [Figure 1](#) illustrates the basic applications for 10GE in areas such as:

- Intra- and Inter-PoP connectivity
- Server farm and data center connectivity

Figure 1 Basic WAN and MAN applications for 10GE

10825EB

10GE to 1GE comparison

10GE differs from 1GE in that it not only supports a much faster media speed, but also for the first time provides both WAN and LAN connectivity. A synchronous optical network (SONET)/Synchronous Digital Hierarchy (SDH) payload encloses WAN Ethernet frames travelling across a fiber-optic link. Embedding Ethernet packets inside SONET frames requires support for SONET-like management, configuration, and statistics.

Unlike the WAN 10GE, 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 10GE module use different module IDs and are fixed in one mode of operation.

Another key difference is that unlike 1GE, 10GE supports only full duplex mode. As per IEEE 802.3ae, Auto-Negotiation is not supported on 10GE. [Table 1](#) provides additional details on the differences between 1GE and 10GE.

Table 1 1GE vs. 10GE comparison

1GE	10GE
<ul style="list-style-type: none"> • CSMA/CD and full duplex • 802.3 Ethernet frame format (includes Min/max frame size) • Carrier extension • One physical interface • Optical/copper Media • Leverage FC PMAs • 8B/10B encoding 	<ul style="list-style-type: none"> • Full duplex only, no auto-negotiation • 802.3 Ethernet frame format (includes Min/max frame size) • Throttle MAC speed (rate adapt) • LAN and WAN physical (PHY) interfaces • Optical media ONLY • Define new PMDs • New coding schemes (64B/66B)

10GE WAN

The following 10GE WAN interface components are explained in the subsections that follow:

- WAN PHY clocking
- WAN PHY budget loss considerations
- MMF usage

For more information about the WAN interfaces see *Using the 10 Gigabit Ethernet Modules: 8681XLR and 8681XLW* in the Passport 8000 Series documentation set.

WAN PHY clocking and other product internetworking

Whether you use internal or line clocking depends on the application and configuration. Typically, you should find the default internal clocking sufficient for most applications, while you should use line clocking on both ends of the 10GE WAN connection (line-line) when connecting through a WAN cloud using SONET/SDH ADM products such as Nortel Networks OPTera Connect DX.

This allows the 10GE WAN modules to synchronize to a WAN timing hierarchy and minimize any timing slips. Also, note that interworking 10GE WAN across a SONET/SDH ADM requires the use of an OC-192c/VC-4-64c payload cross-connection type.

When connecting either back to back using dark fiber, or through metro (OM5200) or long haul (LH 1600G) DWDM equipment, you may use the timing combinations of internal-internal, line-internal, or internal-line on both ends of the 10GE WAN connection. In those scenarios, at least one of the modules provides the reference clock, while DWDM equipment does not typically provide sources for timing synchronization. It is recommended then that you avoid using a line-line combination since it causes an undesired timing loop.

[Table 2](#) presents the recommended clock source settings for 10GE WAN interfaces connected back to back, via dark fiber or DWDM, or across a SONET/SDH WAN. Be sure to select the best clock settings to ensure accurate data recovery and minimize SONET-layer errors.

Table 2 Recommended 10GE WAN interface clock settings

Clock source at both ends of the 10GE 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

Although the 10GE WAN module uses a 1310nm transmitter, it also uses a wideband Rx that allows it to interwork with products using 1550nm 10G interfaces. Such products include Nortel's OPTera Connect DX and LH 1600G that also use a wideband Rx to receive at 1310nm. Nortel Networks OM5200 10G Optical transponder utilizes a 1310nm client side transmitter.

WAN PHY budget loss

When connecting to co-located equipment, such as the OPTera Metro 5200, you should ensure there is enough optical attenuation to avoid overloading the optical receivers of each device. Typically, this may be on the order of approximately 3 to 5db. However, it is not necessary to do so when using the 10GE WAN in an optically-protected configuration with two OM5200 10G transponders.

In such a configuration, you should use an optical splitter that provides a few dB loss. Also, take care here not to attenuate the signal below the Rx sensitivity of the OM5200 10G transponder, which is approximately -11dBm. Other WAN equipment, such as the OPTera Connect DX and LH 1600G, have transmitters that allow you to change the Tx power level. By default, they are typically set around -10dBm, thus requiring no Rx attenuation into the 10GE WAN module. Refer to the *Using the 10 Gigabit Ethernet Modules: 8681XLR and 8681XLW* for the optical specifications for the 10GE modules.

Although distances of up to 10km are supported by the IEEE 802.3ae standard, it is possible to achieve longer distances depending on the fiber characteristics and loss budget of the single mode fiber (SMF) you use.

MMF usage

Nortel Networks does not support multimode fiber (MMF) with 10GE LAN/WAN modules due to limited testing. However, if needed, it is highly recommended to use connections that would be within 100 meters in length.

Design constraints

You should be aware of the following design constraints for 10GE:

- Dual-switch fabrics (SFs)
- Internal multilink trunk (MLT) and load balancing

Each of these is explained in the subsections that follow.

Dual-switch fabric use

Since 10GE modules are M-modules, Nortel Networks strongly recommends you use the 8691SF in a chassis. Due to the internal architecture, you should utilize dual SFs for load balancing and redundancy in any configuration based on 10GE modules.

Based on the hashing algorithm, and on the internal architecture (internal MLT), the best throughput that you can achieve is 9.18 Gb/s (Jumbo Frames). [Chapter 2, “Designing redundant networks”](#) contains more information on the internal Passport architecture, while the hashing algorithm is explained in the subsection that follows.

Internal MLT and load balancing

Every 10GE module uses one MLT ID of the 32 available in the Passport 8600. The MLT ID is configured automatically when a 10GE board is detected in the chassis.

To ensure maximum utilization of the 10GE modules, ingress data is distributed from one 10GE receiver across eight forwarding engines. By hashing the data traffic of the MAC source and destination addresses or the IP source and destination address in the case of IP traffic, traffic is distributed among the eight forwarding engines. This results in a single flow having up to 1.12 Gbps of throughput. [Table 3](#) shows an example of MAC and IP addressing for best throughput through a 10GE interface.

Using eight consecutive sets of addresses from this example table results in aggregating 8 Gbps streams over a 10GE link. You can use any 4 consecutive sets of addresses from this table if you are aggregating 4 Gbps streams over a 10GE link. Note that in normal network scenarios where you have many parallel flows, the load distribution algorithm over the 10GE module ensures that full capacity is used.

Table 3 Example MAC and IP addressing for best throughput

Src MAC	Dest MAC	Src IP	Dest IP
00:01:00:02:00:00	00:01:00:00:02:00	10.1.1.2	10.1.2.2
00:01:00:03:00:00	00:01:00:00:03:00	10.1.1.3	10.1.2.10
00:01:00:04:00:00	00:01:00:00:04:00	10.1.1.4	10.1.2.6
00:01:00:05:00:00	00:01:00:00:05:00	10.1.1.5	10.1.2.14
00:01:00:06:00:00	00:01:00:00:06:00	10.1.1.6	10.1.2.26
00:01:00:07:00:00	00:01:00:00:07:00	10.1.1.7	10.1.2.18
00:01:00:08:00:00	00:01:00:00:08:00	10.1.1.8	10.1.2.22
00:01:00:09:00:00	00:01:00:00:09:00	10.1.1.9	10.1.2.38
00:01:00:0A:00:00	00:01:00:00:0A:00	10.1.1.66	10.1.2.66
00:01:00:0B:00:00	00:01:00:00:0B:00	10.1.1.67	10.1.2.74
00:01:00:0C:00:00	00:01:00:00:0C:00	10.1.1.68	10.1.2.70
00:01:00:0D:00:00	00:01:00:00:0D:00	10.1.1.69	10.1.2.78
00:01:00:0E:00:00	00:01:00:00:0E:00	10.1.1.70	10.1.2.90
00:01:00:0F:00:00	00:01:00:00:0F:00	10.1.1.71	10.1.2.82
00:01:00:10:00:00	00:01:00:00:10:00	10.1.1.72	10.1.2.86
00:01:00:11:00:00	00:01:00:00:11:00	10.1.1.73	10.1.2.102
00:01:00:12:00:00	00:01:00:00:12:00	10.1.1.130	10.1.2.130
00:01:00:13:00:00	00:01:00:00:13:00	10.1.1.131	10.1.2.138
00:01:00:14:00:00	00:01:00:00:14:00	10.1.1.132	10.1.2.134
00:01:00:15:00:00	00:01:00:00:15:00	10.1.1.133	10.1.2.142
00:01:00:16:00:00	00:01:00:00:16:00	10.1.1.134	10.1.2.154
00:01:00:17:00:00	00:01:00:00:17:00	10.1.1.135	10.1.2.146
00:01:00:18:00:00	00:01:00:00:18:00	10.1.1.136	10.1.2.150
00:01:00:19:00:00	00:01:00:00:19:00	10.1.1.137	10.1.2.166

Hardware record optimization

You can optimize control record utilization and achieve a faster boot time in a switch with a high number of interfaces configured by enabling the control record optimization feature.

The 8600 Series switch creates hardware records for routing protocol destination multicast addresses. Frames received for protocols that are not enabled, are dropped at the hardware level. Records are created for RIP, OSPF, VRRP, DVMRP and PIM on all VLANs. These records are used only when routing is not enabled on the interface. In scaled environments, you can optimize record utilization by not programming these records. When control record optimization is enabled, these records are not created and the switch can achieve higher record scaling as well as a faster boot time.



Note: This feature is not supported in HA mode.

The following command is used to enable control record optimization:

```
config bootconfig flags control-record-optimization [true/ false]
save bootconfig
```

Because this is a bootconfig command, remember to save the configuration and reboot the switch after enabling or disabling control record optimization.

Record reservation

Hardware resources or records are shared in a Passport 8600 switch between MAC addresses, local IP interfaces, ARP entries, IP routes, static routes, and IP multicast records and filters. In certain network scenarios, the total number of hardware records required may exceed the available amount. In order to guarantee network stability, you can pre-reserve a minimum set of records.

The default record reservation values for 8600 Series switches are shown in [Table 4](#). These values indicate the preconfigured reserved space record space for the listed protocols. Each protocol can use additional records from the total available set on an as-needed basis.

Table 4 Record reservation specifications

Record type	Default	Range
MAC	2k	0-100k
IP/ARP local	2k	0-6k
Static route	200	0-500
IPMC	500	0-4k
Filter	4k	1k-4k
Total	8.7k	



Note: Be aware that reserved records cannot be overwritten by other types of records. Thus, if you reserve 5k for MAC entries, 5k for ARP entries, 500 for static routes, 500 for IPMC, and 4k for filters, BGP will not be able to use more than 17k records for IP routes on E-modules with a total of 32k records.

8692SF module

Release 3.7 of the 8600 Series switch introduces the Passport 8692SF module. Dual 8692SF switch fabric modules enable a maximum switch bandwidth of 512 Gb/s. Using SMLT in the core, a redundant Passport 8600 switch with two 8692SF modules can provide over 1 Tb/s of core switching capacity.



Note: You can install the 8692SF module in slots 5 or 6 of the 8006, 8010, or 8010co chassis. The 8692SF module is not supported in the 8003 chassis with Release 3.7 software.



Note: The Passport 8600 Series software does not support configurations of the Passport 8692SF module, and Passport 8690SF or Passport 8691SF module installed in the same chassis.

To upgrade to the Passport 8692SF module, see *Installing Passport 8600 Switch Modules* (part number 312749-H)

Electrical considerations

Each Passport 8000 Series chassis provides redundant power options, depending on the chassis and the number of modules installed. A single 8004PS power supply model can support up to five modules in both the Passport 8006 and 8010 chassis. [Table 5](#) shows the power supply matrix.

Table 5 Number of power supplies to install

Chassis	Number of modules ¹	Number of power supplies	
		Required	Redundant configuration
8003	1—3	1	2
8006	1—5	1	2
	6	2	3
8010	1—5	1	2
	6—10	2	3
8010co	1—10	2	3

¹ Includes 1 CPU module for the 8003 chassis; 1 or 2 CPU modules for the 8006, 8010, or 8010co chassis.

Unlike the 8001PS, the 8004PS can provide more output power (850Watts for the 8004 vs. 780Watts), which translates into the ability to support one additional module in most configurations with a single power supply (non-redundant configuration). Check your product installation guides for watts-consumed per modules or contact your Nortel Networks representative.

Software considerations

[Table 6](#) lists the dependencies for several hardware-related features. To ensure proper behavior here, you have access to two modes, enhanced operational, and M mode. Enhanced operational mode allows increased VLAN scalability, while M mode allows increased record scalability.

You enable enhanced operational mode by using the CLI command **config sys set flags enhanced-operational-mode true**, while you enable M mode by entering the **config sys set flags m-mode true** CLI command.

For M-modules, Nortel Networks strongly recommends that you use 8691SFs. (Otherwise, the chassis operates in legacy mode). Based on the internal hardware architecture, it is further recommended that you employ two 8691SFs for traffic balancing and redundancy when using 10GE modules. Additional dependencies are detailed in [Table 6](#).

Table 6 Software/hardware feature dependencies

Mode	Software/ Hardware features	Dependencies
Enhanced operational- allows a higher combination of VLANs and MLT groups	MGID Optimization (See “SMLT and Spanning Tree” on page 110).	E- or M-modules A board is recognized and is taken offline when you set up enhanced operational mode before rebooting.
M mode supports the new M-modules	Increased scalability	M-modules and 8691SF/8692SF A board is recognized and is taken offline when you set up M mode before rebooting.
N/A	10GE modules	E- or M-modules
N/A	BGP	E- or M-modules and 8691SF. For more information on BGP dependencies, see Table 35 on page 395 containing the scaling numbers for E- and M-modules ¹ .
N/A	Layer 3 redundancy	With release 3.7, the Passport 8600 supports the synchronization of the High Availability (HA) mode parameters: <ul style="list-style-type: none"> • L2 parameters- See the Passport 8000 Series documentation for a complete description • L3 parameters- ARP entries, Static routes, RIPv1/v2, OSPF, VRRP, Route Redistribution, and Filters

Table 6 Software/hardware feature dependencies (continued)

Mode	Software/ Hardware features	Dependencies
N/A	SNMPv3/ SSH	Encryption modules. For SNMPv3 and SSH, the encryption modules are: <ul style="list-style-type: none"> • SSH (since 3.2.1): p80c3xxx.img • SNMPv3 (with 3.3): p80c3xxx.des For 10GE, SSH requires E- or M-modules and 8691SF/8692SF
N/A	Jumbo frames	Specific modules. Note: Since release 3.3, jumbo frames (9600 data frames) are supported with the following conditions: <ul style="list-style-type: none"> • The jumbo data frames are forwarded/routed by the Passport 8600. The jumbo control frames are blocked in the initial phase • Only the 8608SX(E), 8608GT(E), 8608GB(E), 10GE I/O modules and the 2 Gig ports of the 8632TXE board have the ability to forward 9.6K jumbo frames. For all other modules, the 8648, 8616SX(E), 8616GT(E) and 10/100 Mbps ports (8632TXE), the maximum transmission unit (MTU) is 1950 bytes, which means that jumbo frames are dropped at the hardware level ingress and egress. At ingress, packets dropped are counted using the <i>Packet Too Long</i> counter. • You configure jumbo frames by setting the MTU to 9600 bytes • During boot time, the value is loaded that you specified in the configuration file. I/O modules that do not support jumbo frames retain the default value (1950 bytes). If you insert a new I/O module and the MTU is set to 9600 for the chassis, the MTU is set to 9600 for jumbo frame compatible I/O modules/ports. The MTU of other I/O modules/ports is then set to the default MTU (1950 bytes).

- 1 Nortel Networks recommends that you use the 8691SF/8692SF in a BGP environment. Non-E and E-modules are recommended for small BGP environments of less than 20K routes. M-modules are required when you have a configuration with more than 20K routes.

Chapter 2

Designing redundant networks

This chapter provides guidelines that help you design redundant networks. It includes the following sections:

Topic	Page number
General considerations	next
Physical layer	55
Platform redundancy	67
Link redundancy	71
Network redundancy	86
Network design examples	115
Spanning tree protocol	124
Using MLT to protect against split VLANs	132
Isolated VLANs	132

General considerations

A number of general factors need to be considered when designing redundant networks, including:

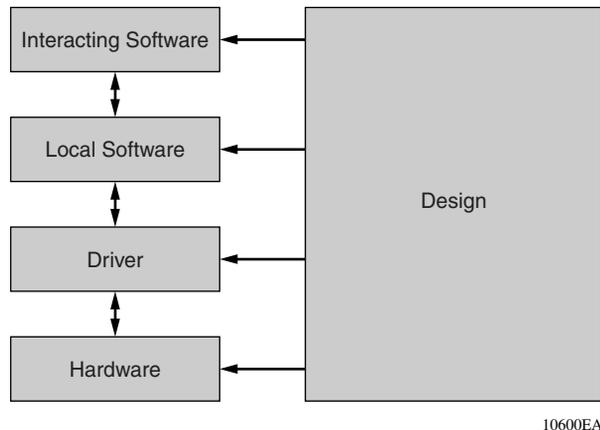
- Reliability and availability
- Platform redundancy
- Desired level of redundancy

This section includes a number of basic network examples to help you in organizing the structure of your network.

Network reliability and availability

A robust data network system depends on system hardware and software interacting together. In the case of the software, you can divide it into three different levels as shown in [Figure 2](#).

Figure 2 Hardware and software reliability



10600EA

These levels are based on the actual functions of the software. For example:

- You can view **Drivers** as lowest level of software that actually performs any functions. Drivers reside on a single module without interacting with other modules, or even external devices. Therefore, you can regard them as being very stable.
- You can view MLT as a prime example of **Local Software** since functionally it may have to interact with several modules, but still in the same device. You can test its functions in an easy way since no external interaction is needed.
- Finally, you can view the **Interacting Software** as the most complex of the levels since it depends on interaction with external devices. OSPF is a good example of this software level. The interaction here may happen with other devices of the same type running a different software version, or even with the devices of other vendors, running a completely different implementation.

Based upon network problem tracking statistics, the following rough stability estimation model of these components has been developed:

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

Based on this model, you may rightly conclude that it makes sense for the network design to off-load the interacting software by putting as much as possible on the other components, especially at the hardware level. Given that reality, Nortel Networks recommends that you follow these generic rules when designing networks:

- 1 Design networks as simply as possible
- 2 Provide redundancy, but do not over-engineer your network
- 3 Use a toolbox to design your network
- 4 Design according to the product capabilities described in the *Release Notes for the Passport 8000 Series Switch Release 3.3*
- 5 Follow the design rules that are provided here in this document and also in the in the various configuration guides for the Passport 8000 Series switch.

Physical layer

The physical layer includes:

- [“Ethernet cable distances,”](#) next
- [“Auto-Negotiation for Ethernet 10/100 BASE Tx”](#) on page 59
- [“100BASE-FX failure recognition/ far end fault indication”](#) on page 60
- [“Gigabit and remote fault indication”](#) on page 61
- [“Using single fiber fault detection \(SFFD\) for remote fault indication”](#) on page 62
- [“VLACP”](#) on page 64

Each of these topics is explained in more detail in the sections that follow.

Ethernet cable distances

Table 7 and Table 8 list distances for 10/100 Ethernet and 1000BASE-TX Gigabit Ethernet cables. Table 9 presents the standard minimum distance ranges for 1000BASE-SX, LX, XD, and ZX Gigabit Ethernet cables. Note that Table 9 represents the minimum distances attainable on *high* quality fiber. You may find it possible to run Gigabit Ethernet cable significantly farther, however, assuming that the loss budget is not exceeded and dispersion is well-controlled.

Table 7 10/100 Ethernet cable distances

	Ethernet 10BASE-T	Fast Ethernet 100BASE-TX	Fast Ethernet 100BASE-FX
IEEE standard	802.3 Clause 14	802.3 Clause 21	802.3 Clause 26
Data rate	10 Mbps	100 Mbps	100 Mbps
Multimode fiber distance	N/A	N/A	412 m (half-duplex) 2 km (full duplex)
Cat 5 UTP distance	100 m	100 m	N/A
STP/Coax distance	500 m	100 m	N/A

Table 8 Gigabit Ethernet cable distances for 1000BASE-TX

	1000BASE-T
IEEE Standard	802.3 Clause 40
Data Rate	1000 Mbps
Optical Wavelength (nominal)	N/A
Multimode Fiber (50 μ m) distance	N/A
Multimode Fiber (62.5 μ m) distance	N/A
Singlemode Fiber (10 μ m) distance	N/A
UTP-5 100 ohm distance	100 m
STP 150 ohm distance	N/A
Number of Wire Pairs/Fiber	4 pairs
Connector Type	RJ-45

Note: Distances are for full duplex. In most cases, this is the expected mode of operation.

Table 9 Gigabit Ethernet standard minimum distance ranges

Trnscv	Fibr typ ¹	Diam (Mcrs)	Modl Bndwd (MHz- km)	Min. Rng (Mtr)	Ave. Optcl TX Pwr	Ave. Rcvr. Snsitiv	Optcl Wvleng	Flux Bdgt (dB)	Patch Loss (dB)	Rmng Flux Bdgt (dB)	Fibr Loss (dB/kM)	Max Fbr Len. (kM)	Sugg. Safe Margn	Flux Bdgt w/Safe Margn (dB)	Sugg Max Fber Len. (kM)
1000 BASE-SX ³	MMF	62.5	160	2 to 220 ²	-9.5 to -4 dBm	-17 dBm (min)	850 nm	7.5	1.0	6.5	3.5	1.9	3.0	3.5	1.0
1000 BASE-SX	MMF	62.5	200	2 to 2753	-9.5 to -4 dBm	-17 dBm (min)	850 nm	7.5	1.0	6.5	3.5	1.9	3.0	3.5	1.0
1000 BASE-SX	MMF	50	400	2 to 500	-9.5 to -4 dBm	-17 dBm (min)	850 nm	7.5	1.0	6.5	3.5	1.9	3.0	3.5	1.0
1000 BASE-SX	MMF	50	500	2 to 5504	-9.5 to -4 dBm	-17 dBm (min)	850 nm	7.5	1.0	6.5	3.5	1.9	3.0	3.5	1.0
1000 BASE-LX ⁴	MMF	62.5	500	2 to 5505	-5.2 to 0 dBm	-22 dBm (min)	1300 nm	16.8	1.0	15.8	1.0	15.8	3.0	12.8	12.8
1000 BASE-LX	MMF	50	400	2 to 5505	-5.2 to 0 dBm	-22 dBm (min)	1300 nm	16.8	1.0	15.8	1.5	10.5	3.0	12.8	8.5
1000 BASE-LX	MMF	50	500	2 to 5505	-5.2 to 0 dBm	-22 dBm (min)	1300 nm	16.8	1.0	15.8	1.5	10.5	3.0	12.8	8.5
1000 BASE-LX	SMF	9	N/A	2 to 10000	-5.2 to 0 dBm	-22 dBm (min)	1300 nm	16.8	1.0	15.8	0.4	39.5	3.0	12.8	32.0
1000 BASE-XD ²	SMF	9	N/A	Up to 50 km	-5.2 to 0 dBm	-24 dBm (min)	1550 nm	18.8	1.0	17.8	0.4	44.5	3.0	14.8	37.0
1000 BASE-ZX	SMF	9	N/A	Up to 70 km	0 to 5.2 dBm	-24 dBm (min)	1550 nm	22	1.0	21.0	0.3	70.0	3.0	18.0	60.0
10GE WAN and LAN ⁵	SMF	9	N/A	Up to 10 km	-5 to -1 dBm	-12.4 dBm	1310 nm	7.4	1.0	6.4	0.4	16.0	2.4	4.0	10.0

1: Multimode fiber = MMF; single-mode fiber = SMF.

- The TIA 568 building wiring standard calls for 160/500 MHz-km multimode fiber.
- The international ISO/EC 11801 building wiring standard calls for 200/500 MHx-km multimode fiber.
- The ANSI Fibre channel specification calls for 500/500 MHx-km 50 micron multimode fiber and 500/500 fiber will be proposed for addition to ISO/EC 11801.
- Using LX optics on multimode fiber may require the use of DMD-compensating patchcords.

2: This is a Bay Networks product.

3: The IEEE standard for 1000BASE-SX is 802.3 Clause 38.3

4: The IEEE standard for 1000BASE-LX is 802.3 Clause 38.4. Note that 1000BASE-XD and 1000BASE-ZX are non-IEEE standard.

5: When the OM5200 10GE and Passport 8600 10GE interfaces are connected to each other in a co-located environment, you may need to attenuate the input power levels by 5 dB to avoid overloading the 10GE Rx. Note that this recommendation is especially valid for the OM5200. It is not exclusively restricted to that device, however.

Transmission distance and optical link budget

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. You can calculate the optical link budget for a proposed network configuration by:

- 1 Identifying all points where signal strength will be lost
- 2 Calculating the expected loss for each point
and
- 3 Adding the expected losses together

By calculating the optical link budget, you can then determine the link's transmission distance, or amount of usable signal strength for a connection between the point where it originates and the point where it terminates.

The absorption of light by molecules in an optical fiber causes the signal to lose some of the light's intensity. This is an area where you should expect loss of signal strength (attenuation) and which you must consider when planning an optical network.

Factors that affect optical signal strength include:

- fiber optic cable (typically .25 dB - .3 dB per kilometer)
- network devices the signal passes through
- connectors
- repair margin (user-determined)

IEEE 802.3ab Gigabit Ethernet- copper cabling

The Institute of Electrical and Electronics Engineers (IEEE) Standards Board approved a specification, known as IEEE 802.3ab, for GE over copper cabling in June 1999. This standard specifies the operation of GE over distances up to 100m using 4-pair 100 ohm Category 5 balanced unshielded twisted pair copper cabling. It is also known as the 1000BASE-T specification since it allows deployment of GE in the wiring closets (WCs) and even right to the desktop if required. It does so without changing the unshielded twisted pair (UTP)-5 copper cabling that is installed in many buildings today.

Auto-Negotiation for Ethernet 10/100 BASE Tx

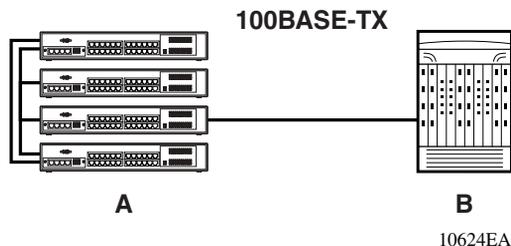
Auto-Negotiation lets devices that share a link segment and automatically configures both devices to take maximum advantage of their abilities. Auto-Negotiation uses a modified 10BASE-T link integrity test pulse sequence, such that no packet or upper layer protocol overhead is added to the network devices.

Auto-Negotiation allows the devices at both ends of a link segment to advertise abilities, acknowledge receipt and understanding of the common mode(s) of operation that both devices share, and to reject the use of operational modes, that both devices do not share. Where more than one common mode exists between the two devices, a mechanism is provided to allow the devices to resolve to a single mode of operation using a predetermined priority resolution function.

The Auto-Negotiation function allows the devices to switch between the various operational modes in an ordered fashion, permits management to disable or enable the Auto-Negotiation function, and allows management to select a specific operational mode. The Auto-Negotiation function also provides a Parallel Detection (so-called auto sensing) function to allow 10BASE-T, 100BASE-TX, and 100BASE-T4 compatible devices to be recognized, even though they may not provide Auto-Negotiation. In this case only the speed can be sensed but not the duplex mode. Nortel Networks recommends the Auto-Negotiation setting on 10/100BASE-TX ports shown in [Table 10](#).

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

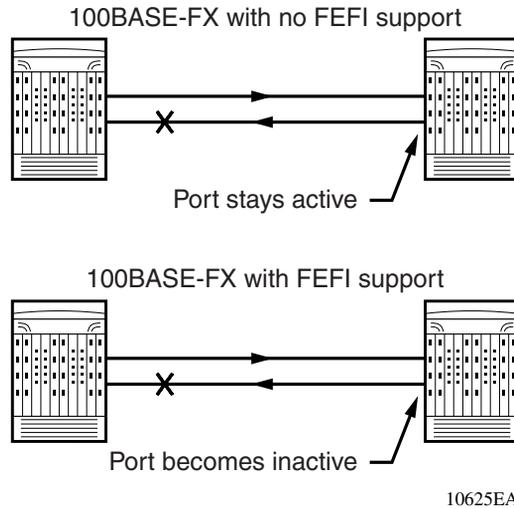
Port on A (Figure 3)	Port on B (Figure 3)	Remarks	Recommendations
AUTO-NEGOTIATION	AUTO-NEGOTIATION	Ports negate on highest supported mode on both sides.	Recommended setting if both ports support Auto-Negotiation mode.
Fixed setting: Full Duplex	Fixed setting: Full Duplex	Both sides require the same mode	Recommended setting if full duplex is required, but Auto-Negotiation is not supported.
Fixed setting: Half Duplex	AUTO-NEGOTIATION	Mode should be set to half-duplex since Auto-Negotiation port cannot detect duplex mode. Speed can be sensed. Auto-Negotiation ports default to half.	10 half duplex recommended on fixed side.

Figure 3 Auto-Negotiation process

100BASE-FX failure recognition/ far end fault indication

Be aware that not all 100BASE-FX drivers support Far End Fault Indication (FEFI). The Passport 8624 supports FEFI. Without FEFI support, if one of two unidirectional fibers forming the connection between the two switches fail, the transmitting side has no mechanism to determine that the link is broken in one direction (Figure 4).

This can lead to network connectivity problems, because the transmitting switch keeps the link active since it still sees signals from the far end. However, the outgoing packets are dropped because of the failure. To avoid this loss of connectivity, Nortel Networks recommends that you use higher layer protocols like OSPF, or a similar protocol.

Figure 4 100BASE-FX FEFI

Gigabit and remote fault indication

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



Note: See [“Using single fiber fault detection \(SFFD\) for remote fault indication,”](#) next, for information about Ethernet switching devices that do not support Auto-Negotiation.

For information on the asynchronous transfer mode (ATM) remote fault indication mechanism F5 and OA&M, see [“F5 OAM loopback request/reply”](#) on page 328.

Using single fiber fault detection (SFFD) for remote fault indication



Note: This information applies to 8600 modules only.

The Ethernet switching devices listed in [Table 11](#) do not support Auto-Negotiation on fiber-based Gigabit Ethernet ports.

Table 11 Ethernet switching devices that do not support Auto-Negotiation

Switch name / Part number	Port or MDA type / Part number
BayStack 470-48T (AL2012x34)	SX GBIC (AA1419001)
	LX GBIC (AA1419002)
	XD GBIC (AA1419003)
	ZX GBIC (AA1419004)
BayStack 470-24T (AL2012x37)	SX GBIC (AA1419001)
	LX GBIC (AA1419002)
	XD GBIC (AA1419003)
	ZX GBIC (AA1419004)
BayStack 460-24T-PWR (AL20012x20)	2 port SFP GBIC MDA (AL2033016)
BPS2000 (AL2001x15)	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)

The port types listed in [Table 11](#) are unable to participate in remote fault indication (RFI), which is a part of the Auto-Negotiation specification. Without RFI, and in the event of a single fiber strand break, there is a possibility that one of the two devices will not detect a fault and will continue to transmit data even though the far end device is not receiving it.

SFFD is an alternative method of providing RFI that must be used when one of the devices listed in [Table 11](#) is present at one or both sides of a Gigabit Ethernet fiber connection. For SFFD to work properly, both ends of the fiber connection must have SFFD enabled, and Auto-Negotiation disabled.



Note: Consult the technical documents for the products in [Table 11](#) to determine if the installed software supports SFFD.

Since Auto-Negotiation works on the 8600 Series switch, it is not necessary to enable SFFD on fiber-based links with an 8600 Series switch at both ends. In this case, Auto-Negotiation should be enabled (and SFFD disabled) on both switches.

When SFFD is enabled on the 8600 Series switch, it detects single fiber faults, and brings the link down 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.



Note: On the BayStack or BPS2000 devices, it may take up to 50 seconds to drop link once a single fiber fault is detected. BayStack or BPS2000 devices may flap the links 4 times during that 50 seconds. Once the fault is corrected, the link is brought up within 12 seconds.

SFFD is supported on the following 8600 Series switch modules:

- 8608SX, 8608SX-E and 8608SX-M
- 8608GBIC, 8608GBIC-E and 8608GBIC-M
- 8616SX, 8616SX-E and 8616SX-M
- 8632TX, 8632TX-E and 8632TX-M (GBIC port only when a fiber GBIC is used)



Note: SFFD is disabled by default since Nortel Networks recommends that you use RFI through Auto-Negotiation whenever it is supported by the devices on both ends of a fiber link.

Configuring SFFD using the CLI



Note: This information applies to 8600 modules only.

SFFD configuration is supported through the CLI. It is not supported in Device Manager.

Since Nortel Networks recommends that, if it is possible, you use RFI through Auto-Negotiation, SFFD is disabled by default. To determine if SFFD is required for a fiber-based connection on your 8600 Series switch, see [Table 11 on page 62](#).

SFFD configuration rules

To make sure that SFFD works properly, use the following rules:

- Use the default setting (disabled) for SFFD whenever Auto-Negotiation is supported on both ends of a fiber link.
- Configure both ends of a fiber connection with the same setting. If a port at one end of a fiber link is configured for SFFD, the port at the other end must also be configured for SFFD.
- Enable only one option per port—either SFFD or Auto-Negotiation—not both. If you enable SFFD on a port, you must disable Auto-Negotiation. If you enable Auto-Negotiation for a port, you must disable SFFD.
- Configure all ports in an MLT with the same option. If you enable SFFD for one port in an MLT, all ports in the MLT must have SFFD enabled and Auto-Negotiation disabled. If you enable Auto-Negotiation for one port in an MLT, all ports in the MLT must have Auto-Negotiation enabled and SFFD disabled.

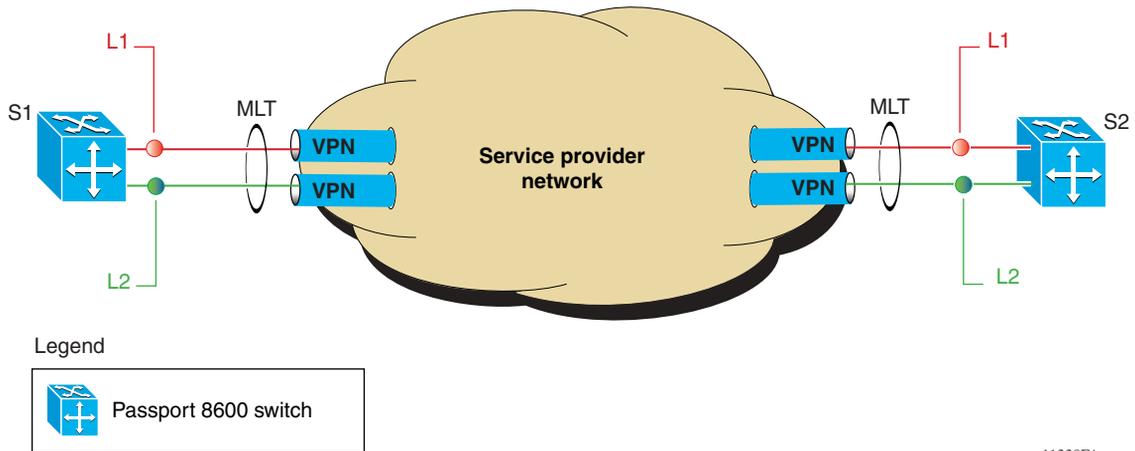
VLACP

Ethernet has been extended to detect remote link failures through functions such as *Remote fault indication* or *Far-end fault indication* mechanisms. A major limitation of these functions, however, is that they terminate at the next Ethernet hop. Therefore, failures cannot be determined on an end-to-end basis over multiple hops.

For example, as shown in [Figure 5](#), when Enterprise networks connect their 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 cloud.

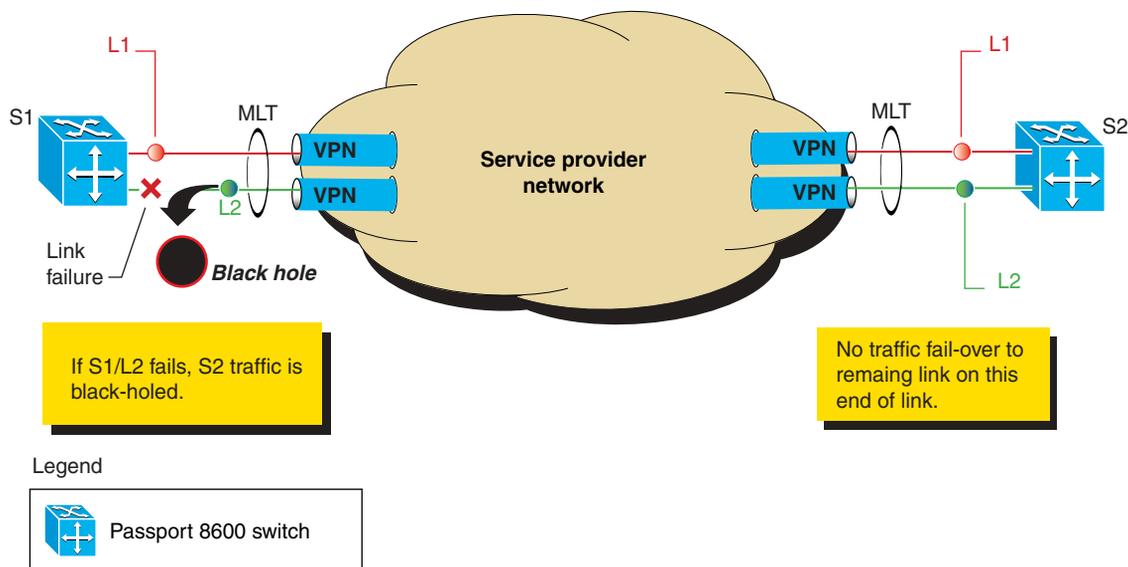
For this example, the MLT (between Enterprise switches S1 and S2) extends through the service provider (SP) network.

Figure 5 Problem description (1 of 2)



As shown in [Figure 6](#), if the L2 link on S1 (S1/L2) fails, the link-down failure is not propagated over the SP network to S2. Thus, S2 continues to send traffic over the S2/L2 link, which is black-holed because the S1/L2 link has failed.

Figure 6 Problem description (2 of 2)



11339FA

As defined by IEEE, the Link Aggregation Control Protocol (LACP) is a protocol that exists between 2 bridge end-points. Therefore, the LACPDUs are terminated at the next (SP) interface. For more information, see “[LACP](#)” on page 77.

Nortel Networks* has developed an extension to LACP called *Virtual LACP (VLACP)* that provides an end-to-end failure detection mechanism. With VLACP, far-end failures can be detected. This allows MLT to properly failover when end-to-end connectivity is not guaranteed for certain links in an aggregation group. Thus, VLACP prevents the failure scenario shown in [Figure 6](#).

When used in conjunction with SMLT, VLACP allows you to switch traffic around entire network devices before L3 protocols detect a network failure, thus minimizing network outages.



Note: The fast periodic time value of 200 ms is not supported for release 3.7 of the Passport 8600 software. The minimum supported fast periodic time value is 400 ms.

Platform redundancy

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

- Redundant power supplies

You should employ $N + 1$ power supply redundancy. (N is the number of required power supplies to power the chassis and its modules). You should also connect the power supplies to an additional power supply line to protect against supply problems.



Note: The Passport 8000 Series switches have two fan trays each with 8 individual fans. Sensors are used to monitor board health.

- 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. For information, see the [“MLT traffic distribution algorithm” on page 73](#).



Note: Nortel Networks recommends you enable Auto-Negotiation on Gigabit interfaces to protect against uni-directional cable faults. Auto Negotiation is part of the IEEE 802.3u spec, while Auto-Negotiation on twisted pair is part of the 802.3 Clause 28 spec. Remote fault indication is part of the Gigabit IEEE 802.3 Clause 37 spec.

- Switch fabric redundancy

Nortel Networks recommends that you use two switch fabrics (SFs) to protect against switch fabric failures. The two SFs load share and also provide backup for each other. For more information about High Availability (HA) mode, see [“HA mode” on page 69](#).

- Central processing unit (CPU) redundancy

The CPU is the control plane of the switch. It controls all learning, calculates the routing protocols, and maintains all port states. If the *last* CPU in a system fails, I/O port status does not change. Instead, the information that has been programmed into the forwarding ASICs is used to make forwarding decisions. There is no active routing protocol update calculation, so network convergence depends on routing protocol time outs.



Note: For SMLT, it is always recommended that you use two CPU modules in the SMLT aggregation switches to avoid packet forwarding to the switch with a single failed CPU board.

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

- Warm standby mode

In this mode, the secondary CPU is waiting with the system image loaded.

- High Availability (HA) mode, often called Hot Standby

For more information, see [“HA mode” on page 69](#).

- Configuration and image redundancy:

The Passport 8000 Series lets you define a primary, secondary and tertiary configuration and system image file path. This protects against system flash failures. For example, the primary path may point to /flash, the secondary to /PCMCIA and the tertiary to a network path.

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



Note: Passport 8000 Series software (release 3.3 and above) does not support using mixed configurations of Passport 8100 modules and Passport 8600 modules simultaneously within the same chassis. Mixed configurations require the concurrent use of one Passport 8190SM and one Passport 8691SF in the system.

Due to a lack of redundancy with a single switch management module (8190SM) for Layer 2 modules, and a single switch fabric/CPU module (8691SF) for Layer 3-7 modules, Nortel Networks recommends that you do not use such configurations. Mixed configurations have not been verified under all conditions.

HA mode

HA mode activates two CPUs simultaneously. These CPUs exchange topology data so that, if a failure occurs, either CPU can take precedence in less than one second with the most recent topology data.

In HA mode, two CPUs are active and exchanging topology data through an internal and dedicated bus. This allows for a complete separation of the traffic since the bus is not used by the regular data path, nor by the data exchange between the CPU and the I/O modules. To guarantee total security, users cannot access this bus.

Depending on the protocols and data exchanged (Layer 2, Layer 3, or platform), the CPUs perform different tasks. This ensures that any time there is a failure, the backup CPU can take precedence with the most recently updated topology data.

Table 12 shows that, because of the amount of work required to perform a failover, regardless of protocol, this task is divided into several phases.

Table 12 HA failover phases

Type of data synchronized	Release 3.2	Release 3.3	Release 3.5	Release 3.7
L1/Port configuration parameters	x	x	x	x
RMON ¹ , Syslog	x	x	x	x
L2/VLAN parameters	x	x	x	x
SMLT	x	x	x	x
802.3ad/802.1x	Not applicable	Not applicable	Not applicable	x
ARP entries	Unavailable	x	x	x
Static and default routes	Unavailable	x	x	x
VRRP	Unavailable	Unavailable	Unavailable	x
RIP	Unavailable	Unavailable	Unavailable	x
OSPF	Unavailable	Unavailable	Unavailable	x
BGP	Unavailable	Unavailable	Unavailable	Unavailable ²
Filters	Unavailable	Unavailable	Unavailable	x
L2 multicast (IGMP)	x	x	x	x
L3 multicast protocols	Unavailable	Unavailable	Unavailable	Unavailable ²

1 Available in the Passport 8000 Series 3.7.1 release.

2 Under investigation for subsequent releases.

For a complete list of limitations, see the release notes that accompany your software.



Note: In HA mode, you cannot configure protocols that are not supported by HA at this time. For example, in HA Layer 3 (release 3.7), BGP and multicast routing protocols (i.e., DVMRP and PIM-SM/PIM-SSM) cannot be enabled.

HA mode is enabled from the CLI using the following command:

```
config bootconfig flags ha-cpu <true|false>
save boot
```

Remember to save the configuration and reboot the switch after enabling or disabling HA mode.

For more information about configuring HA, see *Managing Platform Operations and Using Diagnostic Tools*.

Link redundancy

The sections that follow explain the design steps that you should follow in order to achieve link redundancy.

MLT

When you configure MLT links consider the following MLT guidelines:

- On the Passport 8600 switch up to 32 MLT groups can be created on a switch
- On the Passport 8100 switch up to 6 MLT groups can be created on a switch
- On the Passport 8600 switch up to eight same type ports can belong to a single MLT group
- On the Passport 8100 switch up to four same type ports can belong to a single MLT group
- Same port type means that the ports operate on the same physical media, at the same speed, and in the same duplex mode
- MLT is interoperable with 802.3ad (static, where LACP is disabled)

Switch-to-switch links

In the Passport 8000 Series switch, Nortel Networks recommends for link management and troubleshooting purposes that physical connections in switch-to-switch MLT links follow 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. 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

Routed links

In the Passport 8000 Series switch, brouter ports do not support MLTs. An alternative to using brouter ports to connect two switches with an MLT for routed links is to use VLANs. This configuration 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 for STGx.
- 4 On each of the ports in STGx, disable the Spanning Tree Protocol (STP). By disabling STP per port, you ensure that all BPDUs are discarded at the ingress port, preventing bridging loops.
- 5 Create a VLAN on switch A and switch B (VLAN AB) using STGx. Do not add any other VLANs to STGx because to do so could potentially create a loop.
- 6 Add an IP address to both switches in VLAN AB.

MLT and STG

When you combine MLTs and STGs, note that the spanning tree protocol treats MLTs as another link that could 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.

MLT traffic distribution algorithm

The MLT traffic distribution algorithm is as follows:

- Any bridged packet except IP distribution is based on:
 $\text{MOD} (\text{DestMAC}[5:0] \text{ XOR } \text{SrcMAC}[5:0], \# \text{ of active links})$
- Bridged and routed IP or routed Internetwork Packet Exchange (IPX) distribution is based on:
 $\text{MOD} (\text{DestIP}(X)[5:0] \text{ XOR } \text{SrcIP}(X)[5:0], \# \text{ of active links})$
- Multicast flow distribution over MLT is based on source-subnet and group addresses. To determine the 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. This feature was introduced in release 3.5. The feature is not enabled by default and has to be enabled in order for IP multicast streams to be distributed.

For example, consider:

Group address G[0].G[1].G[2].G[3], Group Mask
 GM[0].GM[1].GM[2].GM[3], Source Subnet address S[0].S[1].S[2].S[3],
 Source Mask SM[0].SM[1].SM[2].SM[3]

Then, the Port =:

$$\begin{aligned} &(((((((G[0] \text{ AND } GM[0]) \text{ xor } (S[0] \text{ AND } SM[0])) \text{ xor } ((G[1] \text{ AND } GM[0] \\ &) \text{ xor } (S[1] \text{ AND } SM[1]))) \text{ xor } ((G[2] \text{ AND } GM[2]) \text{ xor } (S[2] \text{ AND } SM[2] \\ &))) \text{ xor } ((G[3] \text{ AND } GM[3]) \text{ xor } (S[3] \text{ AND } SM[3]))) \text{ MOD } (\text{active ports} \\ & \text{ of the MLT}) \end{aligned}$$

Path cost implementation notes

Passport 8000 Series switches use the following formulas, which are based on the 1993 ANSI/IEEE 802.1D Std, to calculate path cost defaults:

- Bridge Path_Cost = $1000/\text{Attached_LAN_speed_in_Mb/s}$
- MLT Path_Cost = $1000/(\text{Sum of LAN_speed_in_Mb/s of all Active MLT ports})$

Table 13 lists the calculated values.

Table 13 Path cost default values using 1993 ANSI/IEEE 802.1D

Bridge Port defaults	MLT default
<ul style="list-style-type: none"> • 100 for a 10 Mb/s LAN • 10 for a 100 Mb/s LAN • 1 for a 1000 Mb/s LAN. 	<ul style="list-style-type: none"> • 1 for a 4 * 1000 Mb/s LAN (with 4 active links)

The bridge port and MLT path cost defaults for both the single 1000Mb/s link and the aggregate 4000 Mb/s link is 1. Since the root selection algorithm chooses the link with the lowest port ID as its root port, ignoring the aggregate rate of the links, it is recommended that the following methods be used to define path cost:

- Use lower port numbers for MLT so that the MLT with the highest number of active links gets the lowest port ID.
- Modify the default path cost so that non-MLT ports, or the MLT with the lesser number of active links, has a higher value than the MLT link with a larger number of active ports.

You can change a port's path cost from the CLI (`config ethernet <ports> stg <sid> pathcost <intval>`) or JDM (`Edit > Port > STG > PathCost`).

Path cost configuration example 1

For this example, assume the following:

- Two redundant links between two 8600 Series switches
- one MLT link with 4 gigabit ports
- one non-MLT gigabit link port in slot/port 2/1

- a path cost of 4 on the non-MLT link

To configure the path cost for the non-MLT port, enter the following command:

```
config ethernet 2/1 stg 1 pathcost 4
```

Path cost configuration example 2

For this example, assume the following:

- 2 MLT links between two 8600 Series switches
- MLT 2 has four active gigabit links
- MLT 1 has two active gigabit links and is in slot/port 2/1
- a path cost of 4 on each of the links in MLT 1

To configure the port path cost for MLT 1, enter the following command:

```
config ethernet 2/1 stg 1 pathcost 4
```

IEEE 802.3ad-based link aggregation (IEEE 802.3 2002 clause 43)

IEEE 802.3ad-based link aggregation allows you to aggregate one or more links together to form Link Aggregation Groups, thus allowing a MAC client to treat the Link Aggregation Group as if it were a single link.

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

This section includes the following topics:

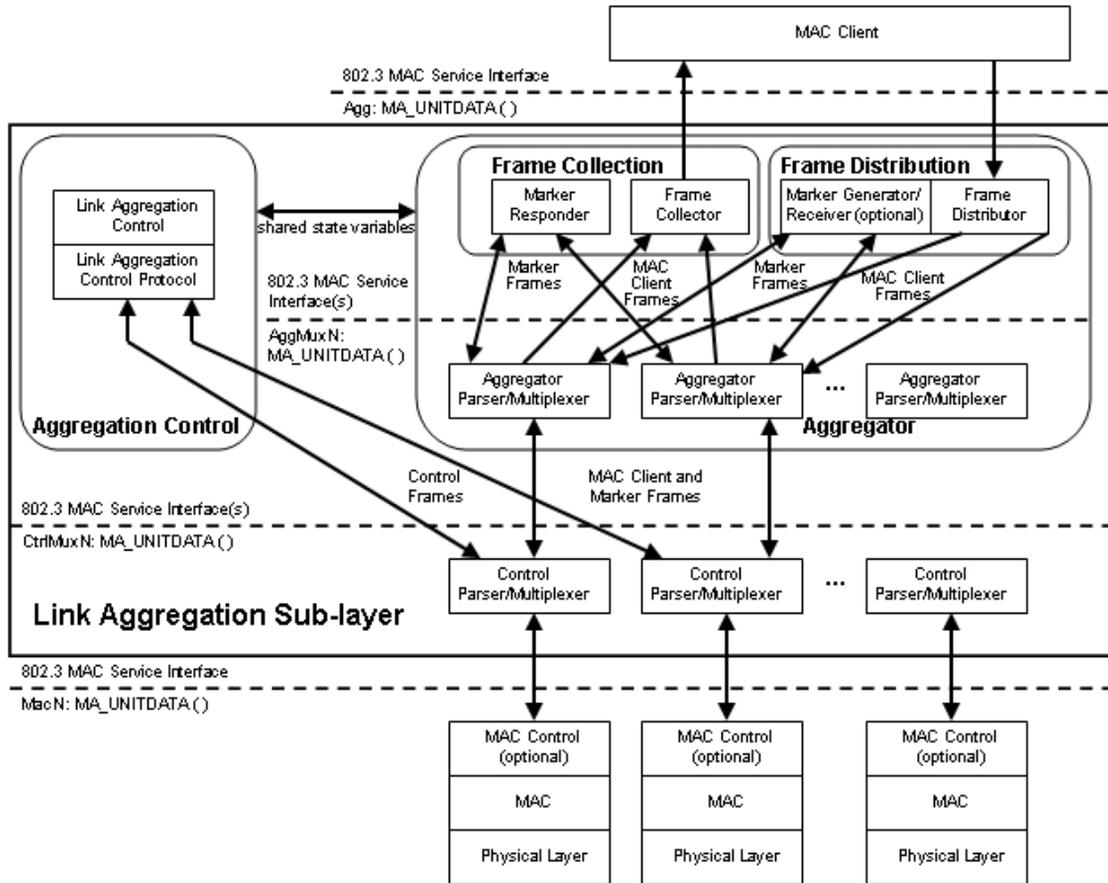
- [“Overview](#)
- [“LACP” on page 77](#)
- [“Link aggregation operation” on page 78](#)
- [“Principles of link aggregation” on page 79](#)
- [“LACP and MLT” on page 80](#)
- [“LACP and spanning tree interaction” on page 81](#)
- [“Link aggregation rules” on page 81](#)

Overview

The IEEE 802.3ad standard comprises service interfaces, LACP, the Marker Protocol, link Aggregation selection logic, parser/multiplexer, frame distribution, and Frame collection functions.

[Figure 7](#) shows the major functions of IEEE 802.3ad defined as Multiple Links Aggregation.

Figure 7 Link Aggregation Sublayer example (according to IEEE 802.3ad)



LACP

The main purpose of LACP is to manage switch ports and their port memberships to link aggregation trunk groups (LAGs). LACP can dynamically add or remove LAG ports, depending on their availability and states.

The interfaces between the LACP module and the other modules is shown in [Figure 7 on page 77](#)

Link aggregation operation

As shown in [Figure 7 on page 77](#), the Link Aggregation sublayer comprises the following functions:

- **Frame Distribution:**

This block is responsible for taking frames submitted by the MAC Client and submitting them for transmission on the appropriate port, based on a frame distribution algorithm employed by the Frame Distributor.

Frame Distribution also includes an optional Marker Generator/Receiver used for the Marker protocol. For the Passport 8600 switch, the Marker Receiver function only is implemented.
- **Frame Collection:**

This block is responsible for passing frames received from the various ports to the MAC Client. Frame Collection also includes a Marker Responder, used for the Marker protocol.
- **Aggregator Parser/Multiplexers:**
 - During transmission operations, these blocks pass frame transmission requests from the Distributor, Marker Generator, and/or Marker Responder to the appropriate port.
 - During receive operations, these blocks distinguish among Marker Request, Marker Response, and MAC Client PDUs, and pass each to the appropriate entity (Marker Responder, Marker Receiver, and Collector, respectively).
- **Aggregator:**

The combination of Frame Distribution and Collection, along with the Aggregator. Parser/Multiplexers, is referred to as the Aggregator.
- **Aggregation Control:**

This block is responsible for the configuration and control of Link Aggregation. It incorporates a Link Aggregation Control Protocol (LACP) that can be used for automatic communication of aggregation capabilities between Systems and automatic configuration of Link Aggregation.
- **Control Parser/Multiplexers:**
 - During transmission operations, these blocks pass frame transmission requests from the Aggregator and Control entities to the appropriate port.

- During receive operations, these blocks distinguish Link Aggregation Control PDUs from other frames, passing the LACPDU to the appropriate sublayer entity, and all other frames to the Aggregator.

Principles of link aggregation

Link aggregation allows you to group switch ports together to form a link group to another switch or server, thus increasing aggregate throughput of the interconnection between the devices while providing link redundancy.

Link aggregation employs the following principles and concepts:

- A MAC Client communicates with a set of ports through an Aggregator, which presents a standard IEEE 802.3 service interface to the MAC Client. The Aggregator binds to one or more ports within a System.
- It is the responsibility of the Aggregator to distribute frame transmissions from the MAC Client to the various ports, and to collect received frames from the ports and pass them to the MAC Client transparently.
- A System may contain multiple aggregators, serving multiple MAC Clients. A given port will bind to (at most) a single Aggregator at any time. A MAC Client is served by a single Aggregator at a time.
- The binding of ports to aggregators within a System is managed by the Link Aggregation Control function for that System, which is responsible for determining which links may be aggregated, aggregating them, binding the ports within the System to an appropriate Aggregator, and monitoring conditions to determine when a change in aggregation is needed.
- Such determination and binding may be under manual control through direct manipulation of the state variables of Link Aggregation (for example, Keys) by a network manager.

In addition, automatic determination, configuration, binding, and monitoring may occur through the use of a Link Aggregation Control Protocol (LACP).

The LACP uses peer exchanges across the links to determine, on an ongoing basis, the aggregation capability of the various links, and continuously provides the maximum level of aggregation capability achievable between a given pair of Systems.

- Frame ordering must be maintained for certain sequences of frame exchanges between MAC Clients.

The Distributor ensures that all frames of a given conversation are passed to a single port. For any given port, the Collector is required to pass frames to the MAC Client in the order that they are received from that port. The Collector is otherwise free to select frames received from the aggregated ports in any order. Since there are no means for frames to be mis-ordered on a single link, this guarantees that frame ordering is maintained for any conversation.

- Conversations may be moved among ports within an aggregation, both for load balancing and to maintain availability in the event of link failures.
- The standard does not impose any particular distribution algorithm on the Distributor. Whatever algorithm is used should be appropriate for the MAC Client being supported.
- Each port is assigned a unique, globally administered MAC address.

The MAC address is used as the source address for frame exchanges that are initiated by entities within the Link Aggregation sublayer itself (for example, LACP and Marker protocol exchanges).

- Each Aggregator is assigned a unique, globally administered MAC address, which is used as the MAC address of the aggregation from the perspective of the MAC Client, both as a source address for transmitted frames and as the destination address for received frames.

The MAC address of the Aggregator may be one of the MAC addresses of a port in the associated Link Aggregation Group

LACP and MLT

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

For example, when you configure a link aggregation group (LAG), use the following steps:

- 1 Assign a numeric key to the ports you want to include in the LAG.
- 2 Configure the LAG to be *aggregatable*.
- 3 Enable LACP on the port.
- 4 Create an MLT and assign the same key to that MLT.

The MLT/LAG will only aggregate those ports whose key match its own.

The newly created MLT/LAG adopts its member ports' VLAN membership when the first port is attached to the aggregator associated with this Link Aggregation Group (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 LAG, first disable LACP on the port, then enable tagging on the port and enable LACP.

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 will be notified. The STP forwarding state does not affect the operation of LACP module. LACPDU can be sent even if the port is in STP blocking state.

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

In contrast to MLT, IEEE 802.3ad-based link aggregation does not expect BPDUs to be replicated over all ports in the trunk group, therefore you must enter the following command to disable the parameter on the spanning tree group for LACP-based link aggregation:

```
#config/stg/x/ntstg disable
```

Be aware that this parameter is applicable to all trunk groups that are members of this spanning tree group. This is necessary when interworking with devices that only send BPDUs out one port of the LAG.

Link aggregation rules

Passport 8600 switch link aggregation groups operate under the following rules:

- All ports in a link aggregation group must be operating in full-duplex mode.

- All ports in a link aggregation group must be running same data rate.
- All ports in a link aggregation group must be in the same VLAN(s).
- Link aggregation is compatible with the Spanning Tree Protocol (STP).
- Link aggregation group(s) must be in the same STP group(s).
- If the `NTSTG` parameter is set to false, STP BPDU transmits only on one link.
- Ports in a link aggregation group can exist on different modules.
- Link aggregation groups are formed using LACP.
- A maximum of 32 link aggregation groups are supported.
- A maximum of 8 active links are supported per LAG.
- A maximum of 8 standby links are supported per LAG.
- Up to 16 ports can be configured in a LAG (8 active and 8 standby ports).

Link aggregation examples

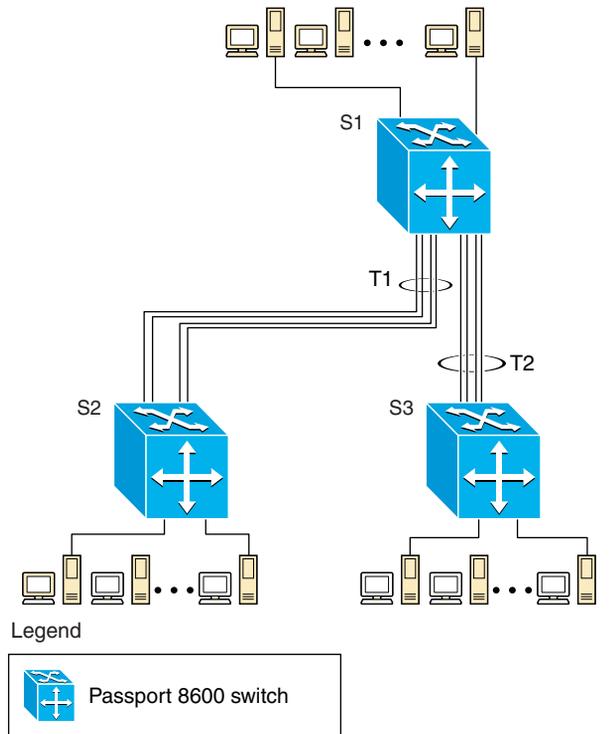
This section provides three link aggregation examples and includes the following topics:

- [“Switch-to-switch example,”](#) next
- [“Switch-to-server MLT example”](#) on page 84
- [“Client/server MLT example”](#) on page 85

Switch-to-switch example

Figure 8 shows two MLTs (T1 and T2) connecting switch S1 to switches S2 and S3.

Figure 8 Switch-to-switch MLT configuration



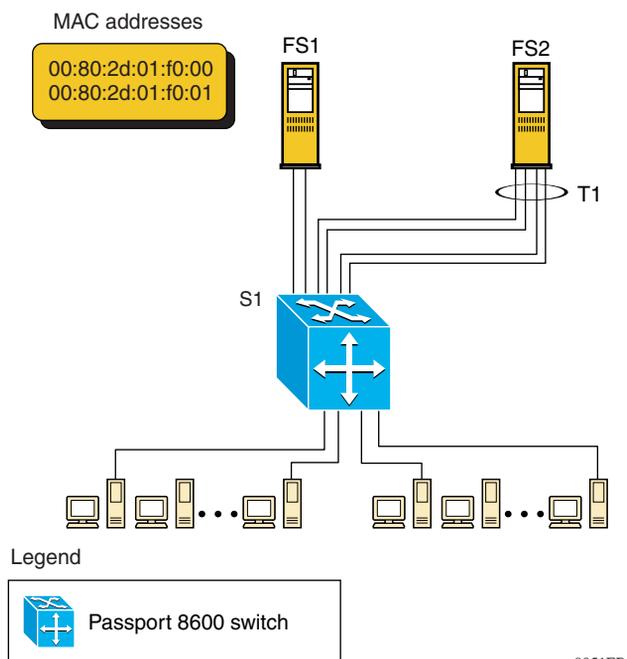
Each of the trunks shown in Figure 8 can be configured with multiple switch ports to increase bandwidth and redundancy. When traffic between switch-to-switch connections approaches single port bandwidth limitations, you can create a MultiLink Trunk to supply the additional bandwidth required to improve performance.

Switch-to-server MLT example

Figure 9 shows a typical switch-to-server trunk configuration. In this example, file server FS1 utilizes dual MAC addresses, using one MAC address for each network interface card (NIC). No MLT is configured on FS1. FS2 is a single MAC server (with a 4-port NIC) and is configured as MLT configuration, T1.

As shown in this example, One port on FS1 is blocked, thus unused; where FS2 benefits from having aggregated bandwidth on MLT T1.

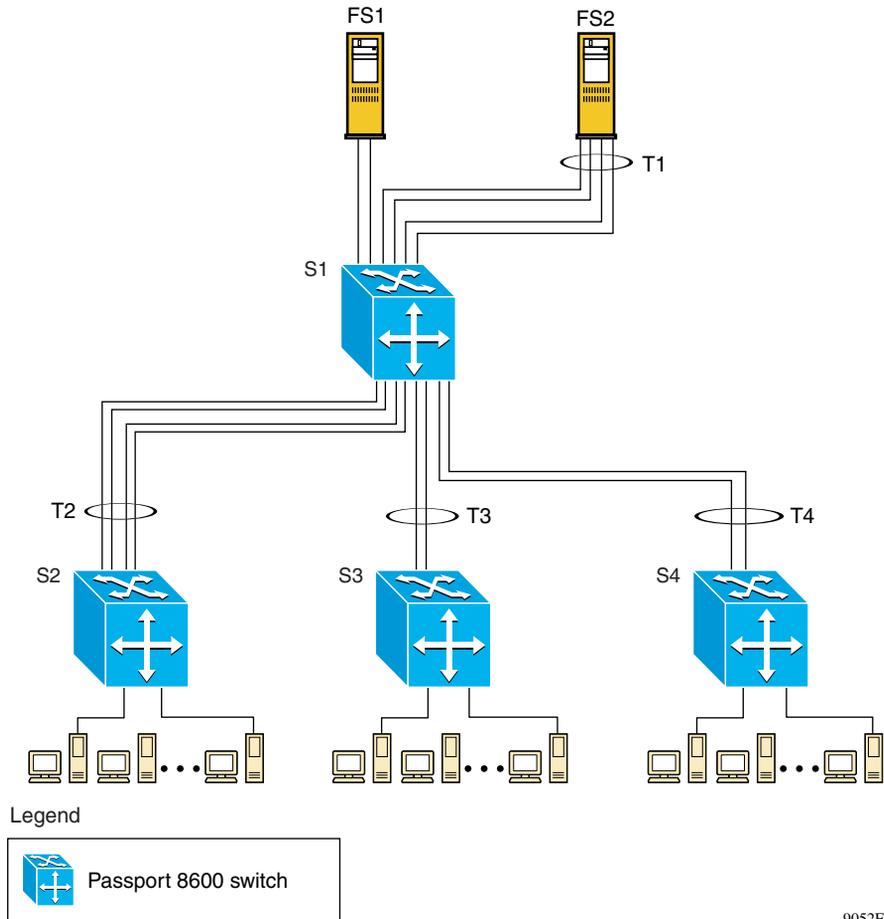
Figure 9 Switch-to-server MLT configuration



Client/server MLT example

Figure 10 shows an example of how MultiLink Trunks can be used in a client/server configuration. In this example, both servers are connected directly to Passport 8600 switch S1. FS2 is connected through a MLT configuration (T1). The switch-to-switch connections are through MLT T2, T3, and T4. Clients accessing data from the servers (FS1 and FS2) are provided with maximized bandwidth through T1, T2, T3, and T4. On Passport 8600 switches, trunk members (the ports that comprise each MLT) do not have to be consecutive switch ports; they can be selected across different modules for module redundancy.

Figure 10 Client/Server MLT configuration



9052EB

With spanning tree enabled, ports that belong to the same MultiLink Trunk operate as follows:

- All ports in the MLT must belong to the same spanning tree group if spanning tree is enabled.
- Identical bridge protocol data units (BPDUs) are sent out of each port.
- The MLT port ID is the ID of the lowest numbered port.
- If identical BPDUs are received on all ports, the MLT mode is forwarding.



Note: You can disable ntstg (ntstg <enable | disable>) if you do not want to receive BPDUs on all ports.

If no BPDU is received on a port or if BPDU tagging and port tagging do not match, the individual port is taken offline.

- Path cost is inversely proportional to the active MLT bandwidth.

Network redundancy

The sections that follow explain the design steps that you should follow in order to achieve network redundancy.

Basic network layouts- physical structure for redundant networks

When designing networks, Nortel Networks recommends that you take a modular approach. This means that you should break the design into different sections, which can then be replicated as needed, using a recursive model.

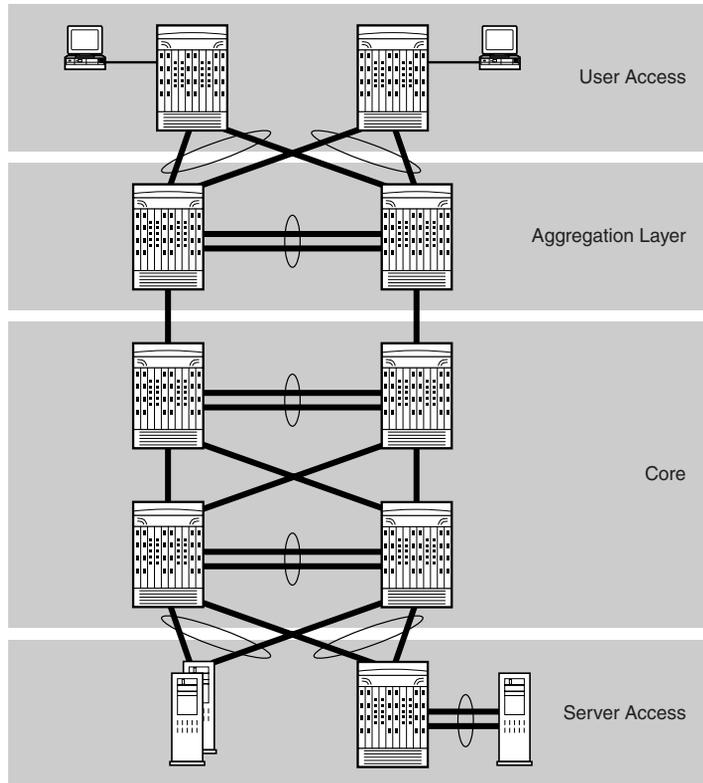
You need to consider several functional entities here, including user access, aggregation, core and server access.

- **User Access Layer-** port switched user access. Normally this layer covers the wiring closet.
- **Aggregation Layer-** aggregation of many user access or wiring closet (WC) switches, this layer is often also called distribution layer, since it involves distribution to the floor/wiring closets.

- **Core-** interconnection between different aggregation points and server farms.
- **Server Access Layer-** server farm connectivity, resource layer.

Note that the design of your network normally depends on the physical layout of your campus and its fiber and copper cable layout (Figure 11).

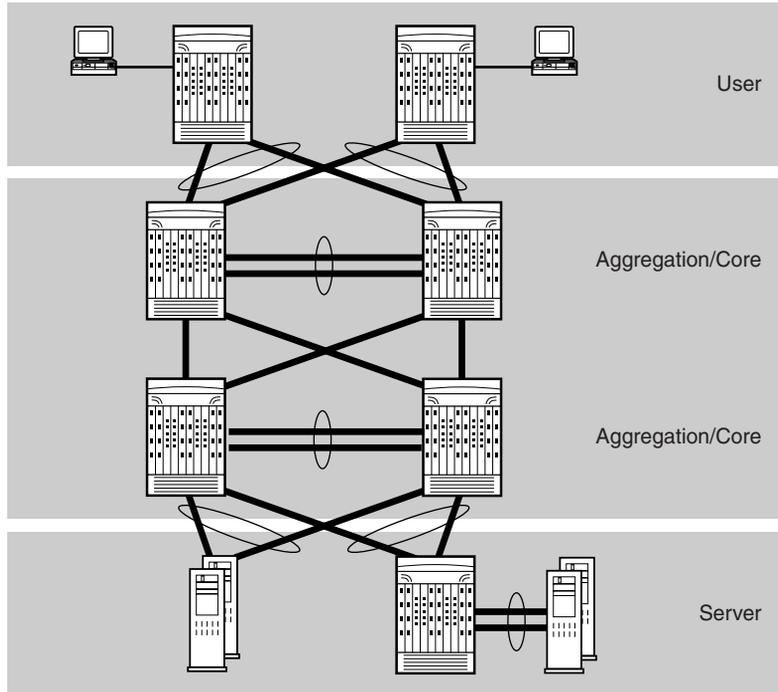
Figure 11 Four-tiered network layout



10601EA

In many cases, you can unify the different layers in one switch maintaining the functionality, but decreasing cost, complexity and network latency (Figure 12).

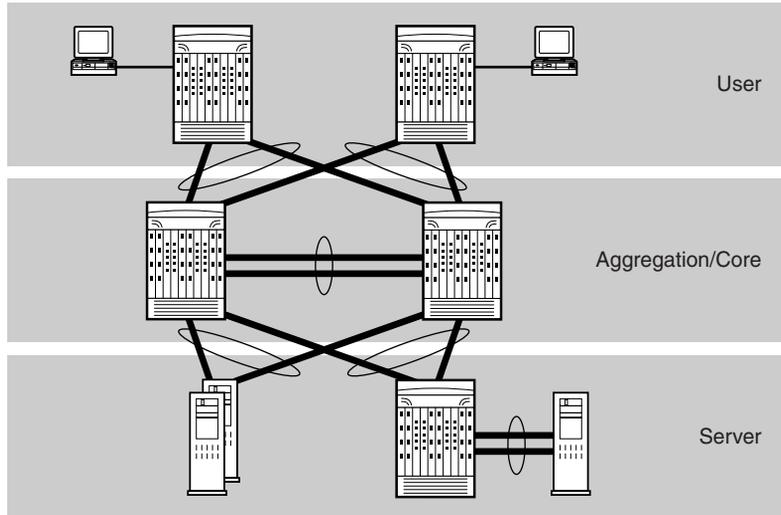
Figure 12 Three-tiered network layout



10602EA

Depending upon the physical fiber layout and the port density requirements, the Server Access and Core can be implemented by the same switch (Figure 13).

Figure 13 Two- or three-tiered networks with collapsed aggregation and core layer

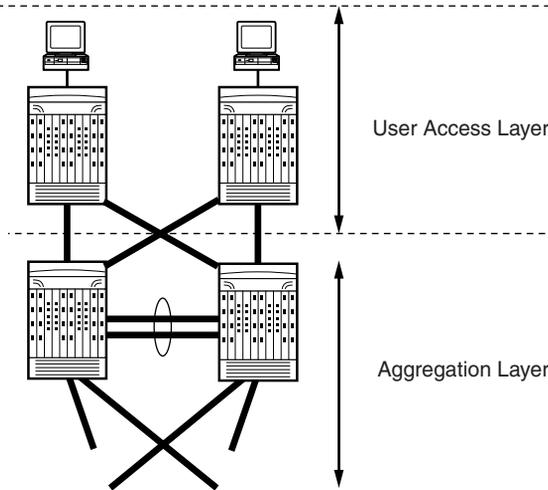


10603EA

Redundant network edge

Figure 14 depicts an aggregation switch pair distributing riser links to wiring closets.

Figure 14 Redundant network edge diagram

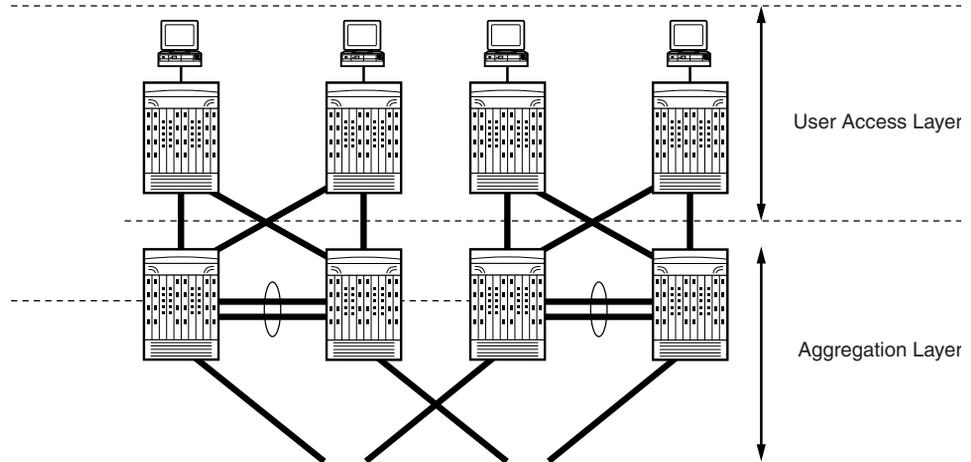


10604EA

Recommended and not recommended network edge designs

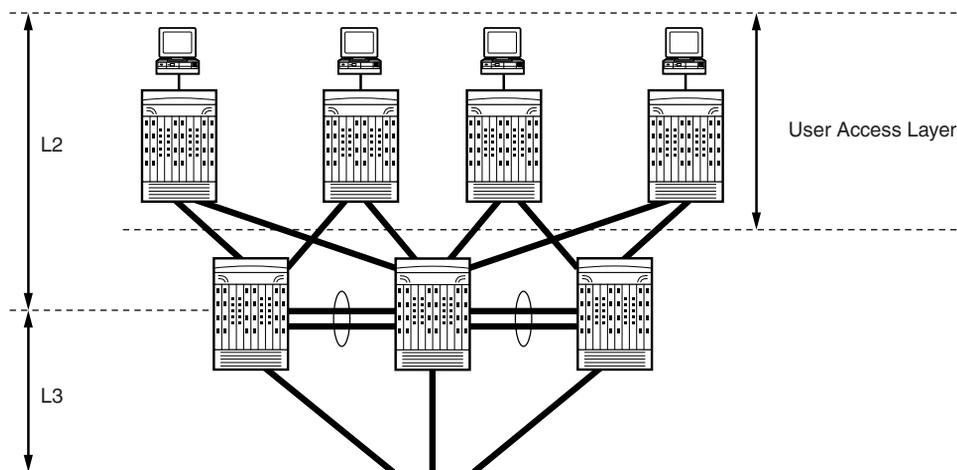
Nortel Networks recommends the network edge setup shown in [Figure 15](#).

Figure 15 Recommended network edge design



Nortel Networks recommends that you do not dual-home edge switches to a set of three aggregation switches.

[Figure 16](#) shows a network setup that Nortel Networks recommends against due to its complexity on one side. On the other side, Nortel Networks SMLT feature provides an optimal solution for a two switch pair network layout. See [“SMLT” on page 92](#) for more information on SMLT and its advantages. A discussion of MLT follows.

Figure 16 Not recommended network edge design

10606EA

SMLT

Split multilink trunking (SMLT) is defined as an MLT with one end split between two aggregation switches.

In addition, single port SMLT lets you configure a split multilink trunk using a single port. This permits scaling the number of split multilink trunks on a switch to the maximum number of available ports. For more information about single port SMLT, see [“Single port SMLT” on page 98](#).

[Table 14](#) defines the components used in SMLT.

Table 14 SMLT components

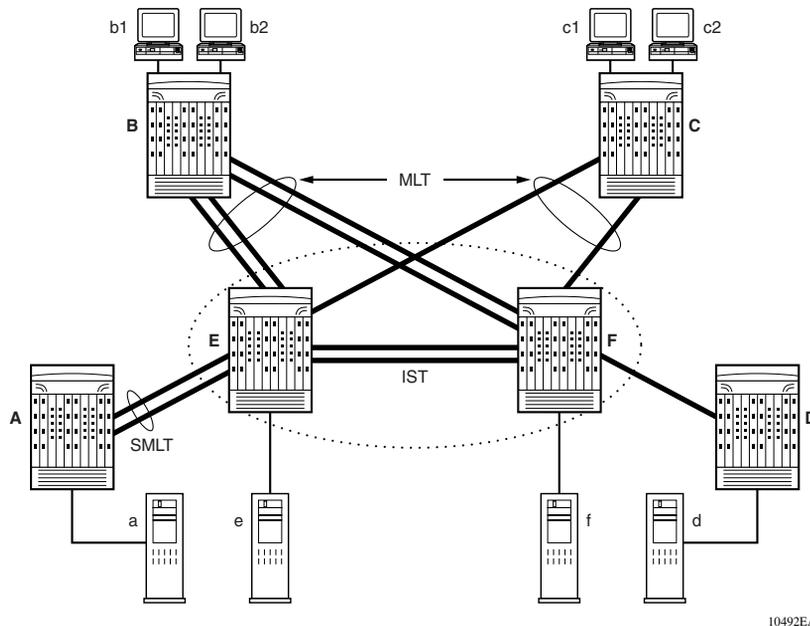
Component	Definition
SMLT aggregation switch	A switch that connects to multiple wiring closet switches, edge switches or Customer Premise Equipment (CPE) devices.
IST (Inter Switch Trunk)	One or more parallel point-to-point links that connect two Aggregation switches together. The two Aggregation switches use this channel to share information so that they may operate as a single logical switch. There can be only one IST per SMLT aggregation switch.

Table 14 SMLT components (continued)

Component	Definition
MLT	A method of link aggregation that allows multiple Ethernet trunks to be aggregated together in order to provide a single logical trunk. An MLT provides the combined bandwidth of the multiple links, as well as the physical layer protection against the failure of any single link.
SMLT Client	A switch located at the edge of the network, such as in a wiring closet or CPE. An SMLT Client switch must be able to perform link aggregation (such as with MLT or some other compatible method) but does not require any SMLT intelligence.

Overview

[Figure 17](#) shows a configuration with a pair of 8600 Series switches as aggregation switches E and F. Four separate wiring closet switches are labeled A, B, C, and D (i.e., Passport 8100s, BayStack 450s, Business Policy Switches or any other MLT-compatible device.)

Figure 17 SMLT configuration with 8600 switches as aggregation switches

Wiring closet switches B and C are connected to the aggregation switches via multilink trunks that are split between the two aggregation switches. For example, SMLT client switch B may use two parallel links for its connection to E, and two additional parallel links for its connection to F.

SMLT client switch C may have only a single link to both E and F. As shown in [Figure 17](#), switch A is also 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 the displayed configuration.

IST link

[Figure 17](#) shows that SMLT only requires two SMLT-capable aggregation switches connected via an IST (Inter Switch Trunk.) The aggregation switches use the IST link to:

- Confirm that each switch is alive and exchanging MAC address information. Thus, the link must be reliable and not exhibit a single point of failure itself.

- Forward flooded packets or packets destined for non-SMLT connected switches, or servers physically connected to the other aggregation switch.

The amount of traffic from a single SMLT wiring-closet which requires forwarding across the IST is likely to be small. However, if the aggregation switches are terminating connections to single-home devices, or if there are SMLT uplink failures, the IST traffic volume may be significant. Because of this, Nortel Networks recommends that the IST be a multi-gigabit MLT with connections across different line cards on both aggregation switches in order to ensure that there is no single point of failure in the IST.

CP-Limit considerations with SMLT IST

Control packet rate limit (CP-Limit) controls the amount of multicast and/or broadcast traffic that can be sent to the CPU from a physical port. It protects the CPU from being flooded by traffic from a single, unstable port. The CP-Limit default settings are:

- default state = enabled
- default multicast packets-per-second (pps) value = 15,000
- default broadcast pps value = 10,000

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.

Disabling IST ports in this way could impair network traffic flow, as this is a critical port for SMLT configurations.

Nortel Networks recommends that an IST MLT contain at least 2 physical ports, although this is not a requirement. Nortel Networks also recommends that CP-Limit be disabled on all physical ports that are members of an IST MLT.

Disabling CP-Limit on IST MLT ports forces another, less-critical port to be disabled if the defined CP-Limits are exceeded. In doing so, you preserve network stability should a protection condition (CP-Limit) arise. Please note that, although it is likely that one of the SMLT MLT ports (risers) would be disabled in such a condition, traffic would continue to flow uninterrupted through the remaining SMLT ports.

The command syntax to disable CP-limit is:

```
config ethernet <slot/port> cp-limit <enable|disable>
```

IST VLAN and peer IP configuration



Note: Nortel Networks recommends that you use an independent VLAN for the IST peer session.

The IST session is established between the peering Passport 8600 SMLT aggregation switches. The basis for this connection is a common VLAN and the knowledge about the peer IP addressing for the common VLAN. Nortel Networks recommends that you use an independent VLAN for this IST peer session. You can do so only by including the IST ports in the VLAN since only the IST port is a member of the IST VLAN.

You should choose the IP subnet addresses from a valid address set. You can enable a routing protocol on the IST VLAN IP interface if you wish. However, it is not necessary to do so.

Supported IST links

In the case of Gigabit Ethernet, Nortel Networks recommends that you use the non-blocking Gigabit modules 8608 or 8632 as IST connections.

SMLT links

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 is called the SMLT links.

Figure 17 also includes end stations connected to each of the switches, 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 may use any method for determining which link of their multilink trunk connections to use for forwarding a packet. This is true as long as the same link is used for a given Source/Destination (SA/DA) pair, regardless of whether or not the DA is known by B or C.

This requirement ensures that there will be no out-of-sequence packets between any pair of communicating devices. Aggregation switches will always send traffic directly to an SMLT client switch and only use the IST for traffic that they cannot forward in another more direct way.

The examples that follow explain the process in more detail.

Example 1- Traffic flow from a to b1 and/or b2

Assuming a and b1/b2 are communicating via Layer 2, traffic goes from switch A to switch E and is then forwarded up its direct link to switch B. Traffic coming down from b1 or b2 to a is sent by switch B on one of its MLT ports. Since it does not attach any special significance to the MLT, it sends traffic from b1 to a on the link to switch E, and the traffic from b2 to a on the link to switch F. In the case of traffic from b1, switch E forwards the traffic directly to switch A, while traffic from b2, which arrived at switch F, is forwarded across the IST to switch E and then to switch A.

Example 2- Traffic flow from b1/b2 to c1/c2

Traffic from b1/b2 to c1/c2 is always sent by switch B down its MLT to the core. No matter which switch (E or F) it arrives at, it is then sent directly to C through the local link. This is the reason why it is necessary for you to dual-home all client switches to the SMLT aggregation pair. By taking such a step, you reduce the amount of traffic on the IST link. Thus, a single IST failure (all SMLT links active) does not result in any traffic interruptions and your risk of your network downtime is minimized even further.

Example 3- Traffic flow from a to d

Traffic from a to d and vice versa is forwarded across the IST because it is the shortest path. This is treated purely as a standard link with no account taken of the SMLT and the fact that it is also an IST.

Example 4- Traffic flow from f to c1/c2

Traffic from f to c1/c2 is sent out directly from F. Return traffic from c1/c2 is then passed across the IST if switch C sends it down the link to E.

SMLT ID configuration

SMLT links on both aggregation switches share an SMLT link ID: SmltId. The SmltId identifies all members of a split trunk group. Therefore, it is mandatory that you terminate both sides of each SMLT having the same SmltId at the same SMLT client switch.



Note: Refer to the [“SMLT square configuration” on page 108](#) and [“SMLT full mesh configuration” on page 109](#) for the exceptions to this rule.

The SMLT IDs can be identical to the MLT IDs. However, be aware that they do not have to be. SmltId ranges are:

- 1-32 for MLT-based SMLTs
- 1-512 for single port SMLTs

Supported SMLT links

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

Single port SMLT

Single port SMLT lets you configure a split multilink trunk using a single port. The single port SMLT behaves just like an MLT-based SMLT and can coexist with SMLTs in the same system; however, an SMLT ID can belong to either an MLT-SMLT or a single-port SMLT per chassis. Single port SMLT lets you scale the number of split multilink trunks on a switch to a maximum number of available ports.

Split MLT links may exist in the following combinations on the SMLT aggregation switch pair:

- MLT-based SMLT + MLT-based SMLT
- MLT-based SMLT + single link SMLT
- single link SMLT + single link SMLT

Rules for configuring single port SMLT:

- The dual-homed device connecting to the aggregation switches must be capable of supporting MLT.
- Single port SMLT is supported on Ethernet, POS, and ATM ports.

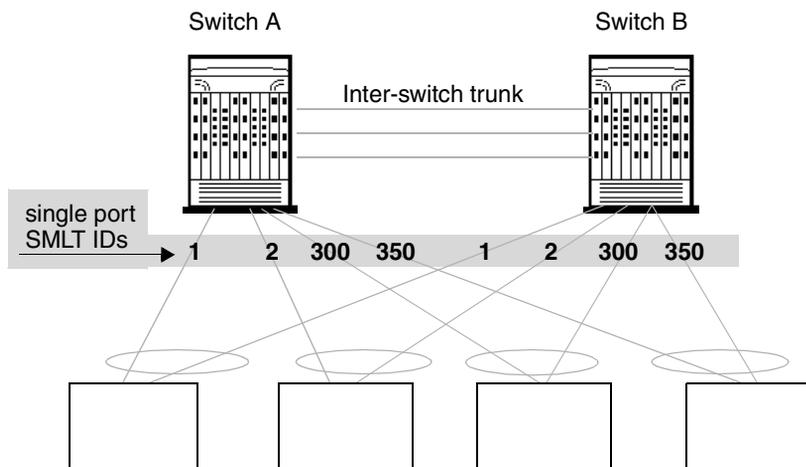


Note: Single port SMLT is not supported on 10 Gig Ethernet ports with release 3.5.

- Each single port SMLT is assigned an SMLT ID from 1 to 512.
- Single port SMLT ports can be designated as Access or Trunk (that is, IEEE 802.1Q tagged or not), and changing the type does not affect their behavior.
- You cannot change a single port SMLT into an MLT-based SMLT by adding more ports. You must delete the single port SMLT, and then reconfigure the port as SMLT/MLT.
- You cannot change an MLT-based SMLT into a single port SMLT by deleting all ports but one. You must first remove the SMLT/MLT and then reconfigure the port as single port SMLT.
- A port cannot be configured as MLT-based SMLT and as single port SMLT at the same time.

Figure 18 shows a configuration in which both aggregation switches have single port SMLTs with the same IDs. This configuration allows as many single port SMLTs as there are available ports on the switch.

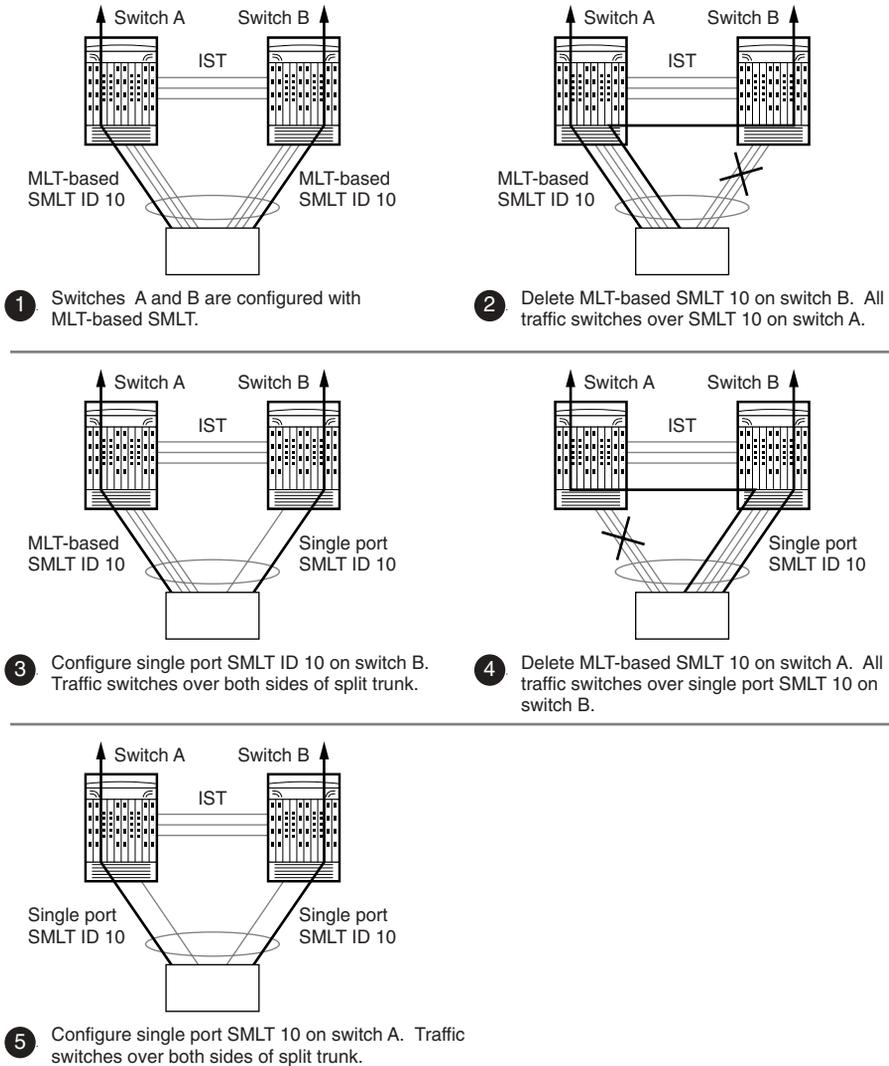
Figure 18 Single port SMLT example



Using MLT-based SMLT with single port SMLT

You can configure a split trunk with a single port SMLT on one side and an MLT-based SMLT on the other. Both must have the same SMLT ID. In addition to general use, Figure 19 shows how this configuration can be used for upgrading an MLT-based SMLT to a single port SMLT without taking down the split trunk.

Figure 19 Changing a split trunk from MLT-based SMLT to single port SMLT



11099EA

For information about configuring single port SMLT, see the publication, *Configuring Layer 2 Operations: VLANs, Spanning Tree and Multilink Trunking*.

Interaction between SMLT and IEEE 802.3ad

With this release the Passport 8600 switch fully supports the IEEE 802.3ad Link aggregation control protocol; not only on MLT and DMLT links, but also extended to a pair of SMLT switches.

With this extension, the Passport 8600 switch now provides a standardized external link aggregation interface to third party vendor IEEE 802.3ad implementations. With previous software versions, interoperability was provided through a static configuration; now a dynamic link aggregation mechanism is provided.

- MLT peers and SMLT client devices can be network switches, and can also be 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.
- Interaction between SMLT and IEEE 802.3ad:

Nortel Networks tightly coupled the IEEE link aggregation standard with the SMLT solution in order to provide seamless configuration integration while also detecting failure scenarios during network setup or operations.

Supported scenarios:

SMLT/IEEE Link aggregation interaction 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.

Failure scenarios:

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

SMLT switches can detect if SMLT IDs are not consistent. The SMLT aggregation switch, which has the lower IP address, does not allow the SMLT port to become a member of the aggregation, thus avoiding bad configurations.

- SMLT client switch does not have automatic aggregation enabled (LACP disabled):

SMLT aggregation switches can detect that aggregation is not enabled on the SMLT client, thus no automatic link aggregation is established until the configuration is resolved.

- Single CPU failures

In the case of a CPU failure in a system with only one switch fabric, the link aggregation control protocol on the other switch (or switches) detects the remote failure and triggers all links connected to the failed system to be removed out of the link aggregation group. This process allows failure recovery for the network along a different network path.



Note: Only dual-homed devices will benefit from this enhancement.

Layer 2 traffic load sharing

From the perspective of the SMLT, you achieve load sharing by the MLT path selection algorithm used on the edge switch. Usually, you do so on an SRC/DST MAC and/or SRC/DST IP address basis. However, this is not required.

From the perspective of the aggregation switch, you achieve load sharing by sending all traffic destined for the SMLT client switch directly and not over the IST trunk. The IST trunk is never used for cross traffic to and from an SMLT dual-homed wiring closet. Traffic received on the IST by an aggregation switch is never forwarded on SMLT links because the other aggregation switch performs that job, thus eliminating the possibility of a network loop.

Layer 3 traffic load sharing

You can also route VLANs that are part of an SMLT network on the SMLT aggregation switches. This enables the network to connect to an L3 core and utilize SMLT functionally as an edge collector. In addition, an extension to the Virtual Router Redundancy Protocol (VRRP), the VRRP backup master concept, has been implemented that improves the Layer 3 capabilities of VRRP in conjunction with SMLT.

Typically, only one of the VRRP switches (Master) forwards traffic for a given subnet. Using the proprietary VRRP extension (BackupMaster) on the SMLT aggregation switch, the backup VRRP switch also routes traffic if it has a destination routing table entry. The VRRP BackupMaster uses the VRRP standardized backup switch state-machine. Thus, it is compatible with the VRRP protocol.

This capability is provided in order to prevent 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 switch in the BackupMaster state routes all traffic received on the BackupMaster IP interface according to its routing table. It does not L2 switch the traffic to the VRRP master.

You must ensure that both SMLT aggregation switches can reach the same destinations through a routing protocol (i.e., OSPF); therefore Nortel Networks recommends that you configure IP addresses per VLAN that you want to route on both SMLT aggregation switches. Then, Nortel Networks recommends that you introduce an additional subnet on the IST with the shortest route path to avoid having any Internet Control Message Protocol (ICMP) redirect messages issued on the VRRP subnets. (To reach the destination, ICMP redirect messages will be issued if the router sends a packet back out through the same subnet it received it on). Refer to [“ICMP redirect messages” on page 152](#) for more details.

Failure scenarios

You should be aware of the following failure scenarios with SMLT. See [Figure 17](#) for a graphic representation of these scenarios.

- Loss of SMLT link

In this scenario, the SMLT client switch detects link failures based on link loss and sends traffic on the other SMLT link(s), 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. If the link is the only one to the aggregation switch, however, the switch informs the other aggregation switch of SMLT trunk loss on failure detection. The other aggregation switch then treats the SMLT trunk as a regular MLT trunk. In this case, the MLT port type changes from splitMLT to normalMLT.

If the link is reestablished, the aggregation switches detect this and move the trunk back to regular SMLT operation. The operation then changes from normalMLT back to splitMLT.

- Loss of aggregation switch

In this scenario, the SMLT client switch detects link failure and sends traffic on the other SMLT link(s), as it does with standard MLT.

The operational aggregation switch detects loss of partner. IST and keep alive packets are lost. The SMLT trunks are changed to regular MLT trunks, and the operation mode is changed to normalMLT. If the partner returns, the operational aggregation switch detects this. The IST then becomes active and once full connectivity is reestablished, the trunks are moved back to regular SMLT operation.

- Loss of one IST Link

In this case, the SMLT client switches do not detect a failure and communicate as usual. In normal use, there will be more than one link in the IST (as it is itself a distributed MLT). Thus, IST traffic resumes over the remaining links in the IST.

- Loss of all IST Links between an aggregation switch pair

Again, the goal of providing connectivity only after a single failure has been exceeded here, since for this to happen, multiple failures must be present.

In the event that all links in the IST fail, the aggregation switches no longer see each other. (Keep alive is lost). Both assume that their partner is dead. For the most part, there are no ill effects in the network if all SMLT client switches are dual-homed to the SMLT aggregation switches. However, traffic which is coming from single attached switches or devices no longer reaches the destination predictably.

There may be a problem for IP forwarding since both switches will try to become master for all VRRPs. Since the wiring closets have no knowledge of the failure, the network will provide intermittent connectivity for devices attached to only one aggregation switch. Finally, data forwarding, while functional, may not be optimal since the aggregation switches may never learn some MAC addresses. Thus, the aggregation switches will flood traffic that would not normally be flooded.

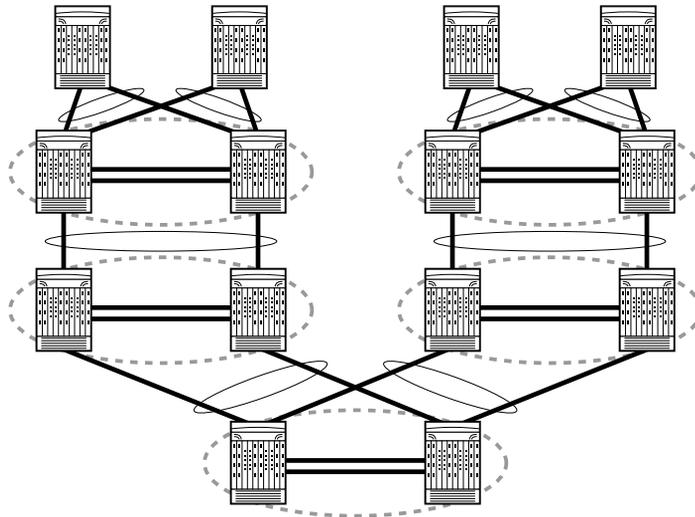
SMLT designs

SMLT designs include the elements described in the following sections:

- [“SMLT scaling,” next](#)
- [“SMLT triangle configuration” on page 107](#)
- [“SMLT square configuration” on page 108](#)
- [“SMLT full mesh configuration” on page 109](#)

SMLT scaling

Within the core of the network, you can configure SMLT groups as shown in [Figure 20](#). In this case, however, both sides of the link are configured for SMLT.

Figure 20 SMLT scaling design

10672EA

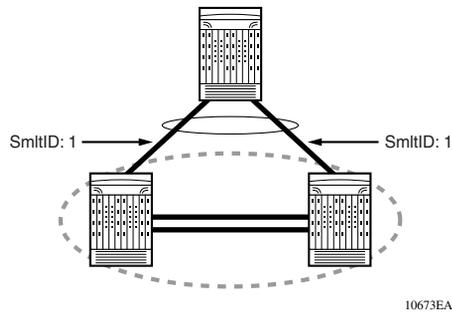
It is possible to use this configuration because there is no state information passed across the MLT link. Thus, both ends believe that the other is a single switch. The result is that no loop is introduced into the network. Any of the core switches or any of the connecting links between them may fail, but the network will recover rapidly.

SMLT triangle configuration

You configure this SMLT configuration in the shape of a triangle (Figure 21), and connect the following to the SMLT aggregation switch pair:

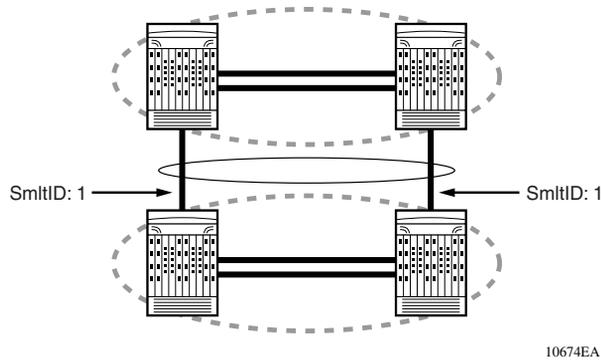
- up to 31 SMLT client switches
- up to 512 single port SMLTs

Figure 21 SMLT triangle configuration



SMLT square configuration

You configure an SMLT square configuration as shown in [Figure 22](#). In this case, all the links facing each other on an SMLT aggregation pair must use the same SmltIds (shown through the MLT ring).

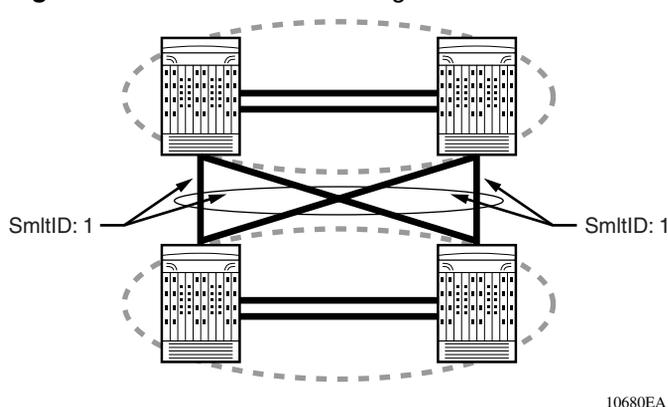
Figure 22 SMLT square configuration

SMLT full mesh configuration

You configure an SMLT full mesh configuration as shown in [Figure 23](#). Note that in this configuration all SMLT ports use the same SmltId (shown through the MLT ring).



Note: Since the full mesh configuration requires MLT-based SMLT, you cannot configure single port SMLTs in a full mesh. In [Figure 23](#), the vertical and diagonal links emanating from any switch are part of an MLT.

Figure 23 SMLT full mesh configuration

SMLT and Spanning Tree

When you configure an SMLT/IST, Spanning Tree is disabled on all the ports that belong to the SMLT/IST. As of release 3.3 of the Passport 8000 Series software, it is not possible for you to have one link on the IST where STP is enabled, even if this link is tagged and belongs to other STGs.

When you connect a VLAN to both SMLT aggregation switches with non-SMLT links, it introduces a loop and is thus, not a supported configuration. You must ensure that the connections from the SMLT aggregation switch pair are done through SMLT links, or through routed VLANs.

SMLT scalability

SMLT scalability is discussed in the following subsections:

- [“VLAN scalability on MLT and SMLT links,” next](#)
- [“IST/SMLT scalability” on page 111](#)
- [“MAC address scalability” on page 111](#)
- [“SMLT and multicast” on page 111](#)

VLAN scalability on MLT and SMLT links

With release 3.3 and above, you can use the following formula to determine the maximum number of VLANs supported per device on an MLT/SMLT:

Without E- or M-modules, you can have:

$$1980 = (\text{\# of VLANs on regular ports}) + (8 * \text{\# of VLANs on MLT ports}) + (16 * \text{\# of VLANs on SMLT ports})$$

The Enhanced Operational Mode feature allows you to exceed these limits by programming the hardware differently because of the capabilities of the E- and M-modules. Specifically, they allow you to have:

$$1980 = (\text{\# of VLANs on regular ports}) + (\text{\# of VLANs on MLT ports}) + (2 * \text{\# of VLANs on SMLT ports})$$

IST/SMLT scalability

There is one IST link per Passport 8600. SMLT IDs can be either MLT or port based. You can have a total of 31 MLT/SMLT groups (32 MLT groups minus 1 MLT group for the IST). With release 3.5, the switch supports port-based SMLT IDs (Port/SMLT). The maximum amount of Port/SMLT IDs is 512, but it is in practice limited by the amount of available ports on the switch.

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

MAC address scalability

When you use SMLT, the total number of supported MAC addresses is 12k. (With M-modules, this limit increases to 50k). This is true if all records are available for MAC address learning.

SMLT and multicast

Refer to [Chapter 6, “Designing multicast networks,”](#) on page 215 for more information on SMLT and multicast.

RSMLT

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

The Nortel Networks RSMLT feature allows rapid failover for core topologies by providing an *active-active* router concept to core SMLT networks. Supported scenarios are: SMLT triangles, squares, and SMLT full mesh topologies, with routing enabled on the core VLANs.

Routing protocols can be any of the following protocol types: IP Unicast Static Routes, RIP1, RIP2, OSPF, BGP and IPX RIP.

In the case of core router failures, RSMLT takes care of packet forwarding, thus eliminating dropped packets during the routing protocol convergence.

SMLT/RSMLT operation in L3 environments

[Figure 24 on page 114](#) shows a typical redundant network example 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 the loop-free topology and enables all links to be forwarding for VLAN 1, IP Subnet A.

The aggregation layer switches are configured with routing enabled and provide active-active default gateway functions through RSMLT.

In this case, routers R1 and R2 are forwarding 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 will failover to R1 as well.

For IP subnet A, VRRP with a Backup-Master could provide the same functions as RSMLT, as long as no additional router is connected to IP subnet A.

RSMLT provides superior router redundancy in core networks (IP subnet B), where OSPF is used for the routing protocol. Routers R1 and R2 are providing 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 are providing router redundancy for IP subnet C and also for core IP subnet B.

Failure scenarios

Please refer to [Figure 24 on page 114](#) for the following failure scenarios.

Router R1 failure:

For example, R3 and R4 are using both R1 as their next hop to reach IP subnet A. Even though R4 sends the packets to R2, they will be routed directly at R2 into subnet A. R3 sends its packets towards R1 and they are also sent directly into subnet A. When R1 fails, all packets will be directed to R2, with the help of SMLT. R2 still routes for R2 and R1.

After OSPF convergences, the routing tables in R3 and R4 change their next hop to R2 in order to reach IP subnet A. The network administrator can choose to set the hold-up timer (i.e., for the amount of time R2 will route for R1 in a failure case) for a time period greater than the routing protocol convergence, or set it as indefinite (i.e., the pair always routes for each other).

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

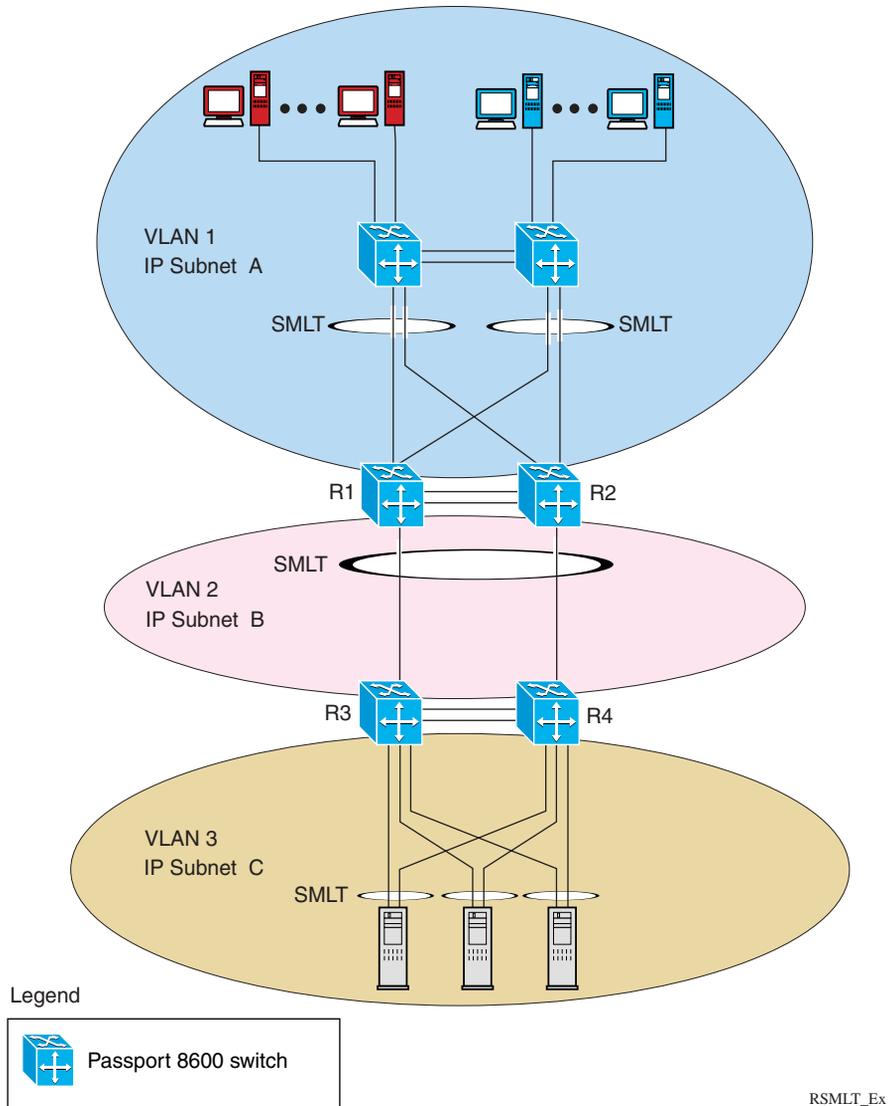
Router R1 recovery

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

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, it is recommended that you add another routed VLAN to the ISTs. This additional routed VLAN should have lower routing protocol metrics as a traversal VLAN/subnet in order to avoid unnecessary ICMP redirect generation messages. This recommendation also applies to VRRP implementations.

Figure 24 SMLT and RSMLT in L3 environments



Designing and configuring an RSMLT network

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 has to be a member of SMLT links and the IST trunk.

The VLAN also must be routable (IP address configured). On all four routers, an Interior Routing Protocol (IGP) such as OSPF has to be configured, although it is independent from RSMLT. (See [Figure 24 on page 114](#)).

There are no changes to any IGP state machines and any routing protocol, even static routes, can be used 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 R1/R2 and R3/R4.

Network design examples

Following are a series of examples to help you design all the relevant layers of your network:

- The Layer 1 examples deal with the physical network layouts
- The Layer 2 examples map VLANs on top of the physical layouts
- The Layer 3 examples show the routing instances Nortel Networks recommends to optimize IP and IPX for network redundancy

Layer 1 examples

[Figure 25](#) contains a series of Layer 1 design examples that illustrate the physical network layout.

Figure 25 Layer 1 design examples

Example 1

HA mode to cover CPU faults

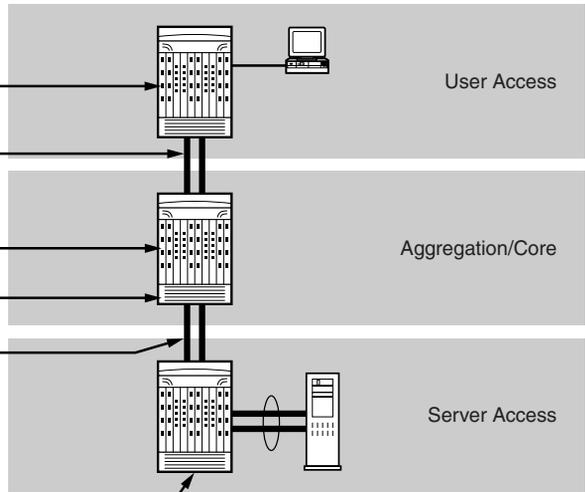
DMLT to cover complete module

Redundant switch fabrics to cover switch fabric faults

HA mode to cover CPU faults

GIG-Autonegotiation or 100FX FEF1 to cover single cable faults

HA mode and switch fabric redundancy for slot 5/6 protection



10607EA

Example 2

HA mode to cover CPU faults

SMLT/DMLT to cover complete switch failures

Distributed MLT to cover module failures

Redundant switch fabrics to cover switch fabric faults

HA mode to cover CPU faults

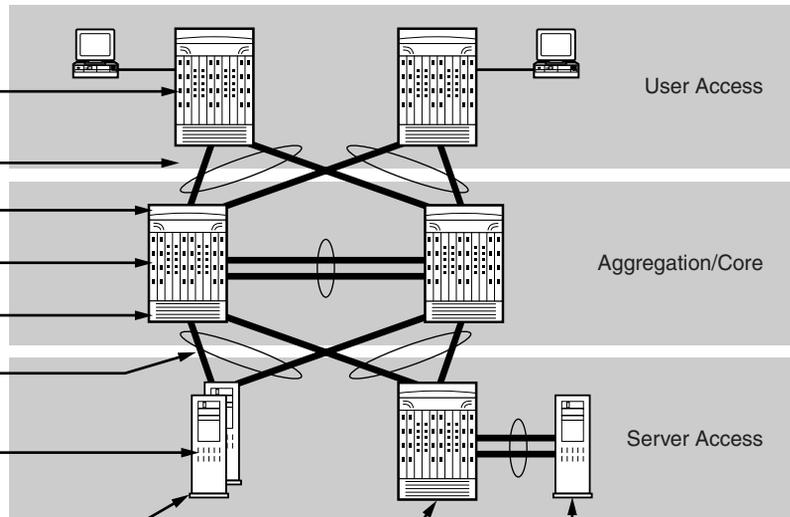
GIG-Autonegotiation or 100FX FEF1 to cover single cable faults

Server dual home through SMLT to cover complete switch failures

Server using MLT/802.3ad to protect against server NIC faults

HA mode and switch fabric redundancy for slot 5/6 protection

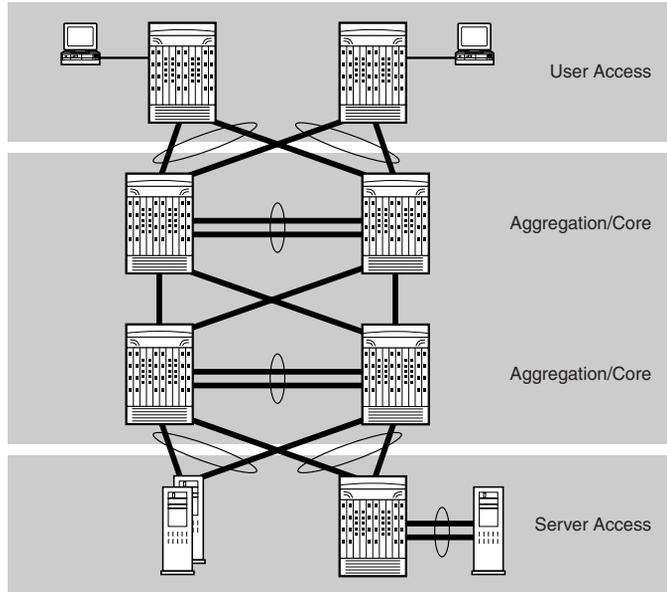
Server using MLT/802.3ad to protect against server NIC faults



10610EA

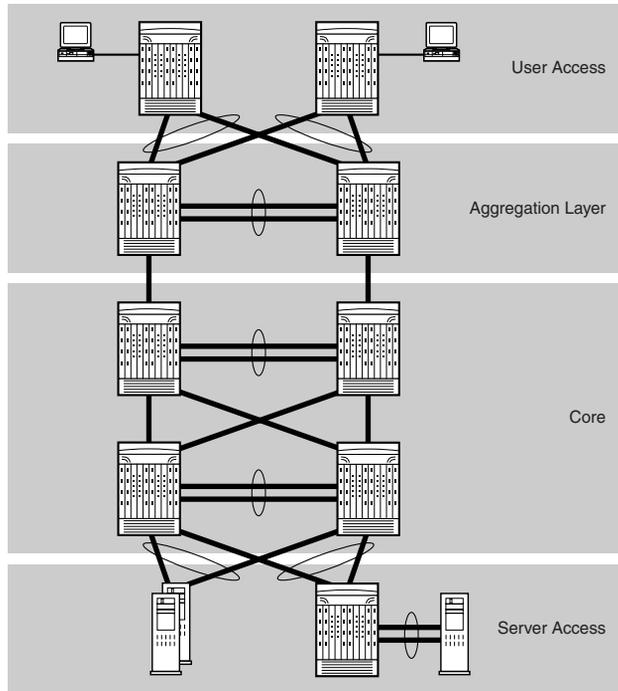
Based on Example 2, all the Layer 1 redundancy mechanisms are described.

Example 3



10612EA

Example 4



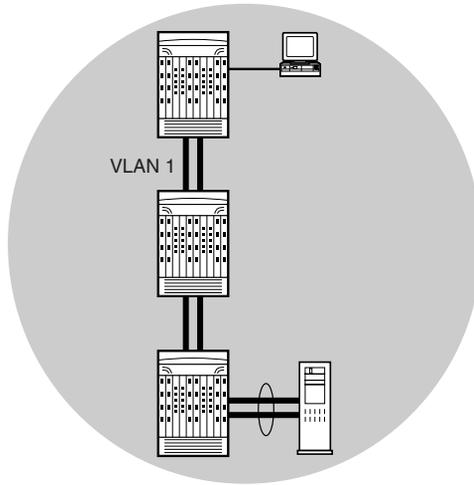
10601EA

Layer 2 examples

Figure 26 contains a series of Layer 2 network design examples that map VLANs on the top of the physical network layout.

Figure 26 Layer 2 design examples

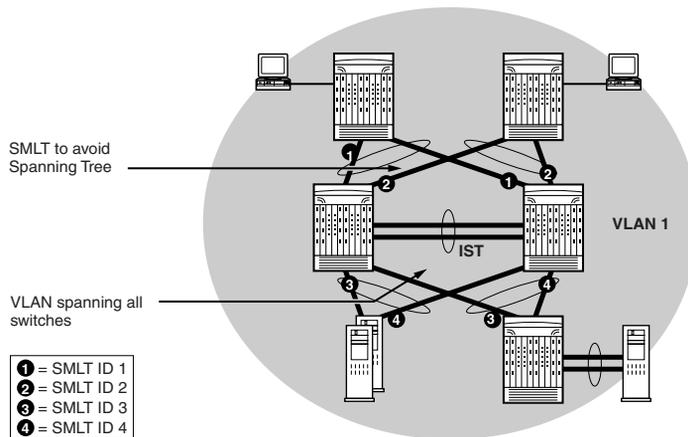
Example 1



10608EA

Example 1 shows a device redundant network using one VLAN on all switches. To support multiple VLANs, 802.1Q tagging is required on the links with trunks.

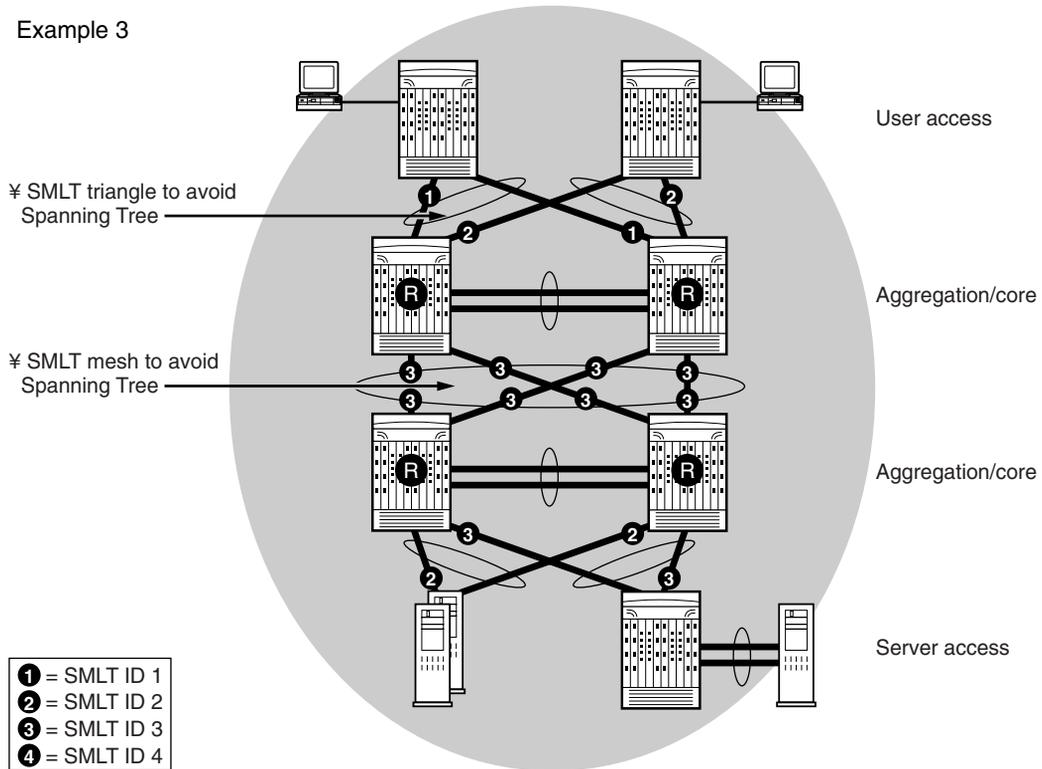
Example 2- Using SMLT



10613EA

Example 2 depicts a redundant network using SMLT. This layout does not require STP. SMLT removes the loops, but still ensures that all paths are actively used. Each wiring closet (WC) can have up to 8 Gigabytes worth of bandwidth to the core. Note that this SMLT configuration example is based on a three stage network.

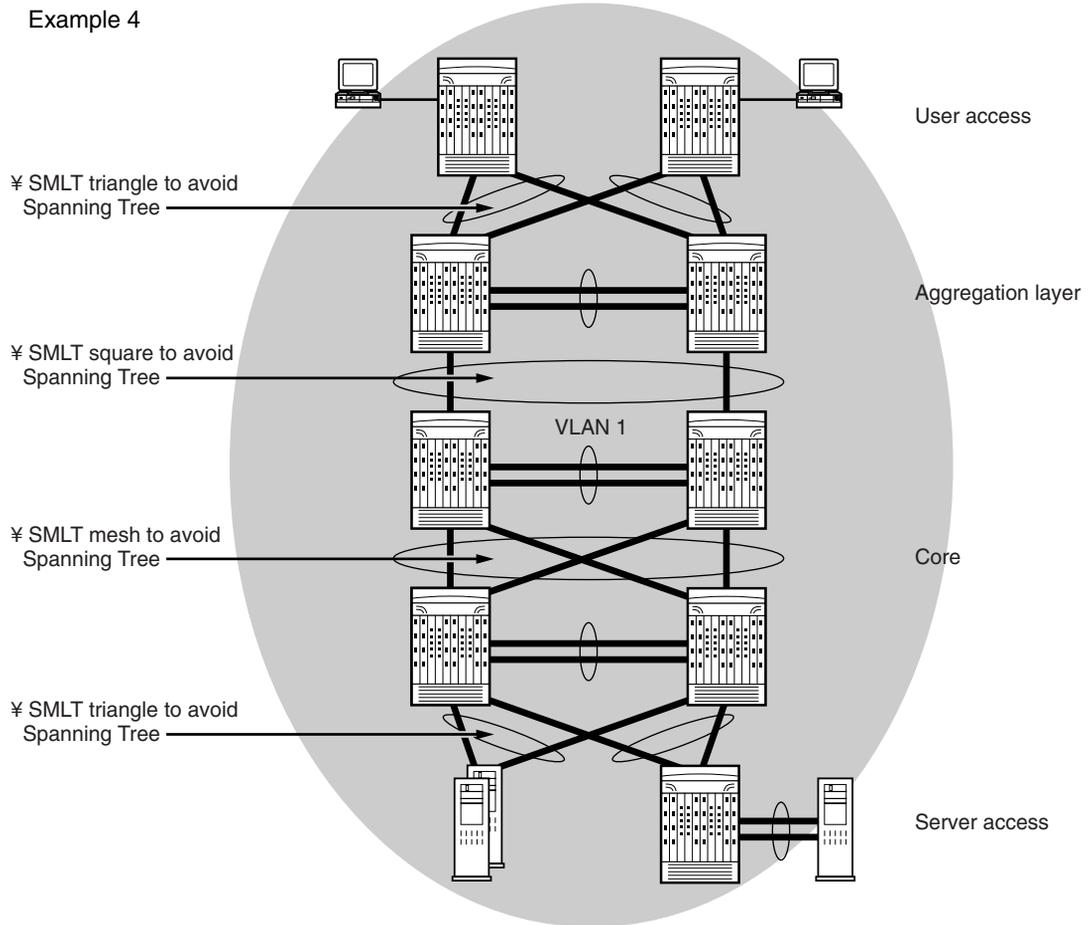
Example 3



10616EB

In Example 3, a typical SMLT ID setup is shown. (Note that SMLT is part of MLT. Therefore, all SMLT links also have an MLT ID. The SMLT and MLT ID can be the same number, but do not necessarily have to be).

Example 4



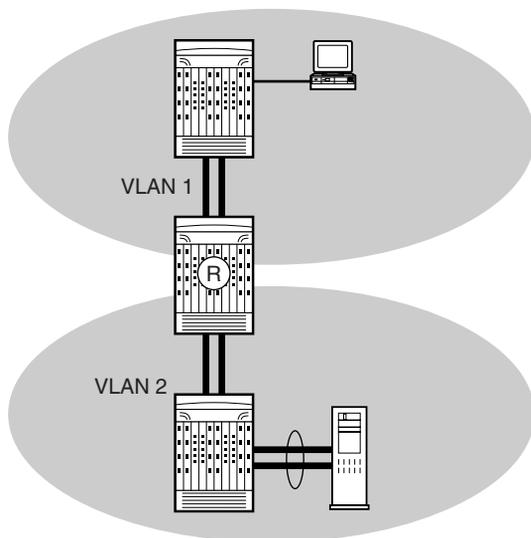
10617EB

Layer 3 examples

Figure 27 contains a series of Layer 3 network design examples that display the routing instances Nortel Networks recommends to optimize IP and IPX for network redundancy.

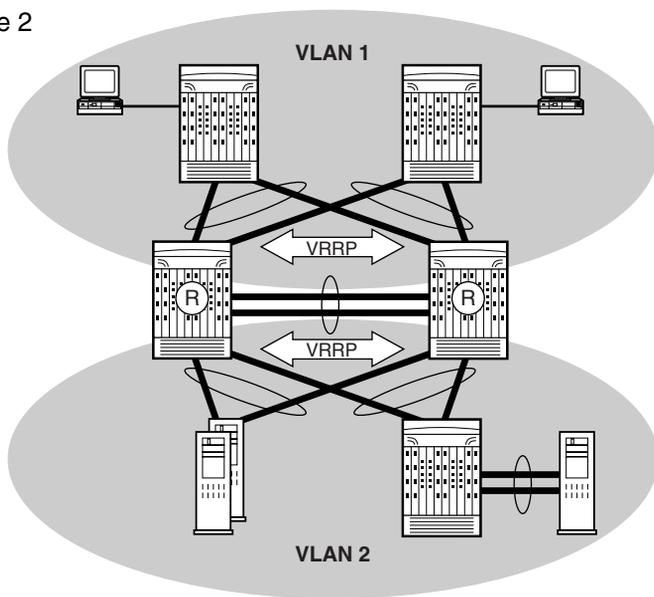
Figure 27 Layer 3 design examples

Example 1



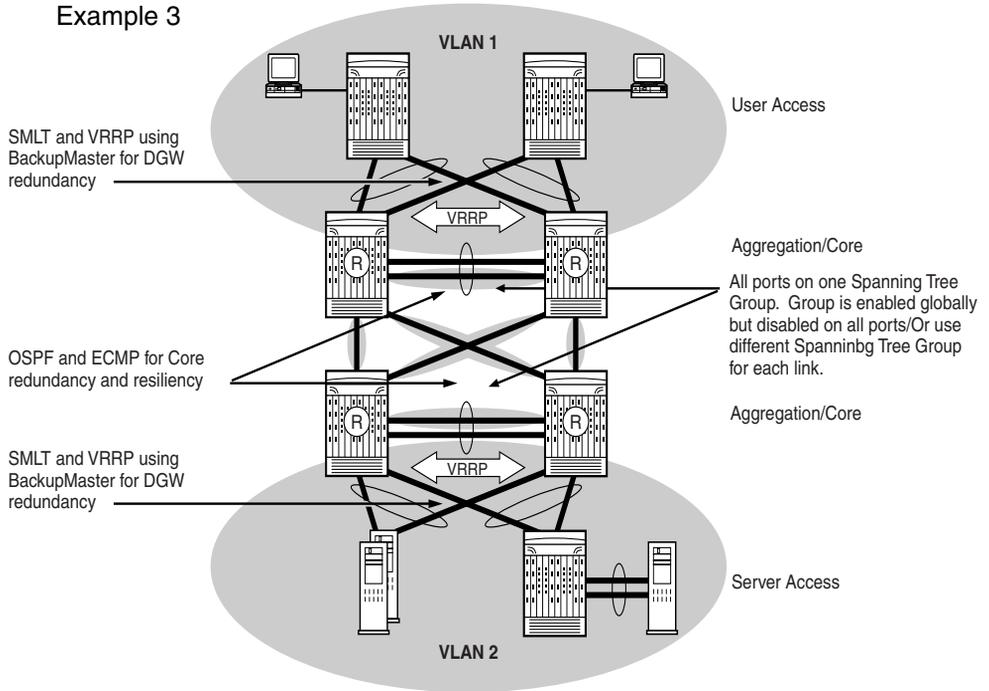
10609EA

Example 2

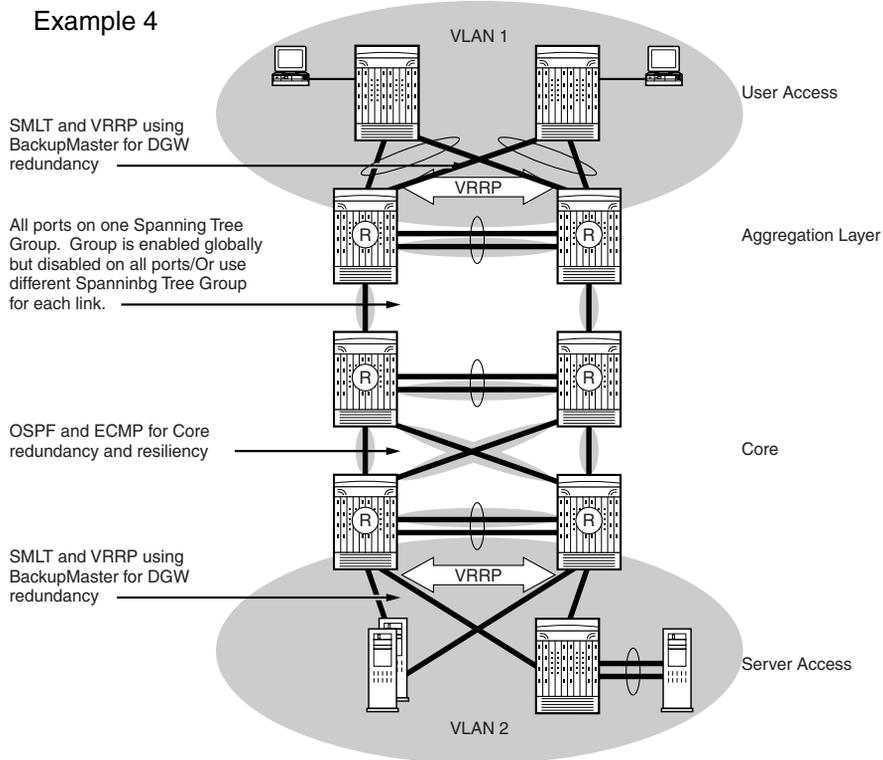


10614EA

Example 3



10615EA



Spanning tree protocol

This section describes some designs you should considering when configuring the spanning tree protocol (STP) on the Passport 8000 Series switch.

STGs and BPDU forwarding

You can enable or disable STP at port or at spanning tree group (STG) level. If you disable the protocol at STG level, BPDUs 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.

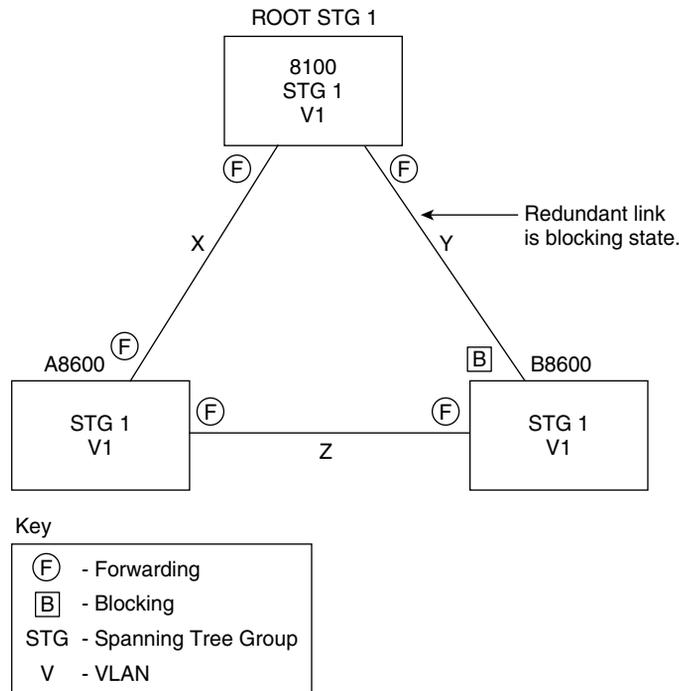
Multiple STG interoperability with single STG devices

Nortel Networks provides multiple STG interoperability with single STG devices. When you connect the Passport 8600 switch with Layer 2 switches, such as the Passport 8100 switch or the BayStack 450 switch, be aware of the differences in STG support between two types of devices. The Passport 8100 switch and the BayStack 450 switch support only one STG, while the Passport 8600 switch supports 25 STGs.

The problem

In [Figure 28](#), all three devices (8100, A8600, and B8600) are members of STG1 and VLAN1. Link Y is in blocking state to prevent a loop and links X and Z are in forwarding state. With this configuration, congestion on link X is possible since it is the only link forwarding traffic from the Passport 8600 switches to the Passport 8100 switch.

Figure 28 One STG between two Layer 3 devices and one Layer 2 device



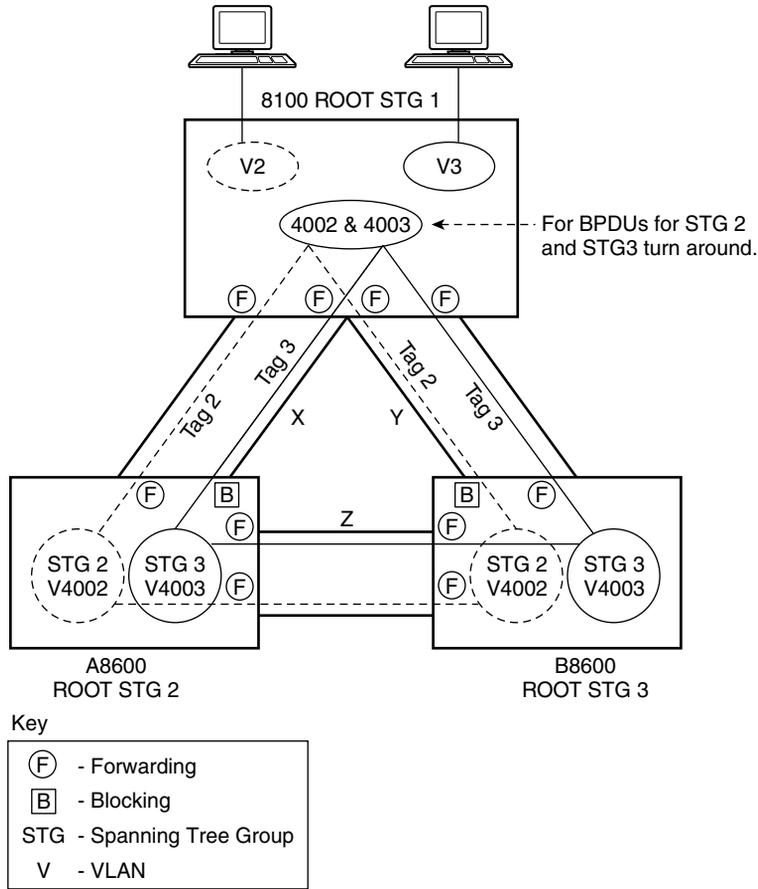
9921EA

The solution

To provide load sharing over links X and Y, create a configuration with multiple STGs that are transparent to the Layer 2 device and 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 Passport 8100 switch as regular traffic not BPDUs.

In the configuration in [Figure 29](#), the BPDUs generated by the two STGs (STG2 and STG3) are forwarded by the Passport 8100 switch. To create this configuration, you must configure STGs on the two Passport 8600 switches, 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 29 Alternative configuration for STG and Layer 2 devices



9920EB

Create two STGs and set MAC addresses for the STGs

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 will not process the STG2 and STG3 BPDUs as BPDUs, but forward them as a 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 in the CLI by using the following command:

```
config stg <stgid> create [vlan <value>] [mac <value>]
```

To change the MAC address in the Java Device Manager (JDM), select VLAN > STG > Insert.

Configure STG roots

On the Passport 8600 switches (A8600 and B8600), configure A8600 as the root of STG2 and B8600 as the root of STG3. On the Layer 2 device, the Passport 8100 switch, configure it as the root of STG1. You configure a switch to be the root of an STG by giving it the lowest root bridge priority.

To set a switch as root in an STG, you can use the CLI or the JDM. When you are connected to the switch, do one of the following:

- In the CLI, enter this command

```
config stg <id> priority 100
```

where `id` is the STG ID.
- From the JDM menu bar, choose VLAN > STG > Configuration. Double click in the Priority field of the STG you want, and enter, for example, 100. Click Apply and Refresh.

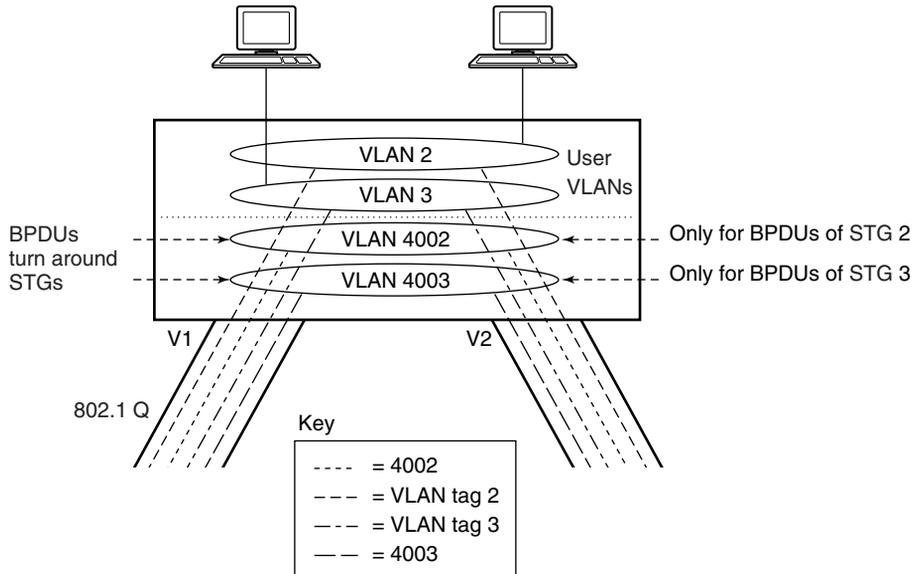
Make sure that the STG ports have tagging enabled on them and the same ports are members of STG2 and STG3.

Configure VLANs

Configure four VLANs on the Layer 2 switch to include the tagged ports connected to the Passport 8600 switches. 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, you must give two of the VLANs the IDs: “4002” and “4003.” [Figure 30](#) illustrates the four VLANs configured on the Passport 8100 switch and the traffic associated with each VLAN.

After you configure the Passport 8100 switch, configure VLAN 2 and VLAN 3 on the Passport 8600 switches.

Figure 30 VLANs on the Layer 2 switch



9933EA

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 will be tagged with a “TaggedBpduVlanId” that is derived from adding 4,000 to the STG ID number. For example, for STG3 the TaggedBpduVlanId is 4003. For more information about tagging in VLANs, refer to the *Configuring Layer 2 Operations: VLANs, Spanning Tree, Multilink Trunking* document in the Passport 8000 Series 3.3 documentation set.

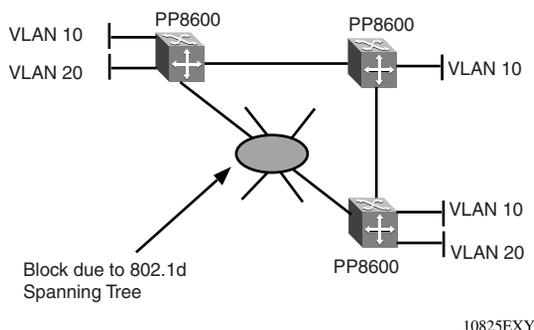
PVST+

Per-VLAN Spanning Tree Plus (PVST+) is Cisco System’s proprietary spanning tree mechanism that uses a spanning tree instance per VLAN. PVST+ is an extension of Cisco’s PVST with support for the IEEE 802.1Q standard. It is the default spanning tree protocol for Cisco switches and uses a separate spanning tree instance for each configured VLAN. In addition, it supports IEEE 802.1Q STP for support across IEEE 802.1Q regions.

Nortel Networks' Passport 8600 and Cisco both support standards-based 802.1d spanning tree. In addition they both support proprietary mechanisms for multiple instances of spanning tree. Be aware, however, that using 802.1d spanning tree provides only one instance of spanning and may lead to incomplete connectivity for certain VLANs depending on network topology. (See the previous section, "Isolated VLANs" on page 132, for more information).

In a network where one or more VLANs span only a segment of the switches (Figure 31), 802.1d spanning may block a path used by a VLAN that does not span all switches.

Figure 31 802.1d Spanning tree



The workaround here is to use multiple spanning tree instances. Specifically, the Passport 8600 uses a tagged BPDU address associated with a VLAN tag ID. This ID is applied to one or more VLANs and is used among Passport 8600 switches to prevent loops. The tagged BPDU address is unique for each STG ID. However, you must ensure that it is configured in the same way for all the Passport 8600s in the network.

With release 3.7.0, you can configure the Passport 8600 using either tagged BPDUs or PVST+. By default, when you configure PVST+, it uses IEEE 802.1Q single STP BPDUs on VLAN 1 and PVST BPDUs for other VLANs. This allows a PVST+ switch to connect to a switch using IEEE 802.1Q spanning tree as a tunnel for PVST.

PVST+ BPDUs tunnel across the 802.1Q VLAN region as multicast data. The single STP is addressed to the well-known STP MAC address 01-80-C2-00-00-00. The PVST BPDUs for other VLANs are addressed to multicast address 01-00-0C-CC-CC-CD. You can use PVST+ to load balance the VLANs by changing the VLAN bridge priority.



Note: Release 3.7.0 software implements PVST+ and not PVST

Passport 8600 PVST+ implementation and guidelines

You choose the Spanning Tree group type during the creation of the group. Choices here are either the Nortel STG (the default), or Cisco's PVST+. The guidelines for using PVST+ are similar to those when using the regular Nortel tagged BPDU method. As a result, the same recommendations or limitations apply. For example, 25 STP groups are officially supported. (The CLI allows up to 64).



Note: Adding STP groups (specifically Cisco PVST+) puts more pressure (utilization/memory) on the CPU. Thus, you should ensure that you use PVST+ only when there is no other option available (i.e., SMLT).

It is highly recommended here that you:

- Control the location of the root bridge by changing (lowering) the default value.
- Modify the path costs to optimize the traffic distribution, specifically when you have aggregation groups (MLT/802.3ad).

Using MLT to protect against split VLANs

Consider link redundancy when you create distributed VLANs. Split subnets or separated VLANs disrupt packet forwarding to the destinations in case of a link failure.

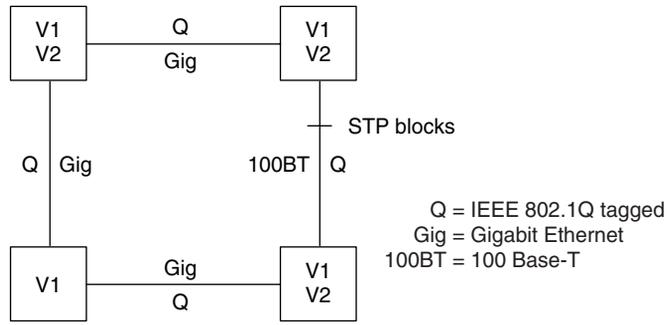
The split subnet VLAN problem can occur when a VLAN carrying IP or IPX traffic is extended across multiple switches and a link between the switches fails or is blocked by the Spanning Tree Protocol. The result is a broadcast domain that is divided into two noncontiguous parts. This problem can cause 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 in such a way that ports that are blocking do not divide your VLANs into two noncontiguous parts. Set up your VLANs in such a way that device failures do not lead to the split subnet VLAN problem. Analyze your network designs for such failure modes.

Isolated VLANs

Similar to the split VLAN issue is VLAN isolation. [Figure 32](#) shows four devices connected by two VLANs (V1 and V2) and both VLANs are in the same STG. V2 includes three of the four devices, while V1 includes all four devices. When the Spanning Tree Protocol detects a loop, it blocks the link with the highest link cost. In the case of the devices in [Figure 32](#), the 100 MB/s link is blocked, thus isolating a device in V2. To avoid this problem, either configure V2 on all devices or use a different STG for each VLAN.

Figure 32 VLAN isolation



9896EA

Chapter 3

Designing stacked VLAN networks

This section provides guidelines to help you design a stacked VLAN network. It includes the following topics:

Topic	Page number
About stacked VLAN	next
sVLAN operation	137
Network loop detection and prevention	142
sVLAN multi-level onion architecture	144
sVLAN and network or device management	147
sVLAN restrictions	147

About stacked VLAN

Stacked VLAN (sVLAN), also referred to as “Q-in-Q”, allows packets to have multiple tags, or stacked tags, so that service providers can transparently bridge tagged or untagged *customer* traffic through a core network. The current sVLAN implementation is proprietary; however, there is an IEEE draft in progress, Provider Bridges, to standardize stacked VLAN implementations.

The current provider bridging project in IEEE standard 802.1ad acts to:

- Provision multiple Virtual Bridged LANs using the common LAN equipment of a single organization.
- Use a common infrastructure of Bridges and LANs to offer independent customer organizations the equivalent of separate LANs, Bridged LANs, or Virtual Bridged LANs.

Features

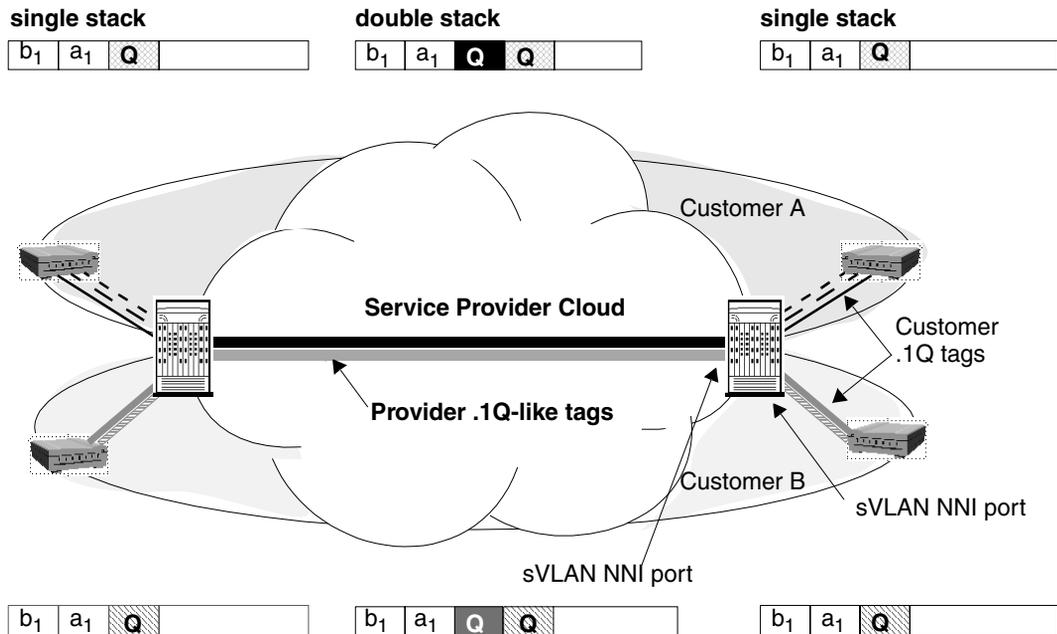
sVLAN provides the following features:

- VLAN tunneling of 802.1q tagged or untagged traffic through service provider core networks, allowing overlapping customer VLAN configurations.
- Improved VLAN scalability by summarizing customer VLANs into core VLANs.
- Improved VLAN scalability by using a layered architecture.
- Loop detection mechanism for customer-introduced loops.

sVLAN operation

Figure 33 illustrates sVLAN operation. Customer tags are encapsulated into provider frames. The original MAC Source and Destination MAC addresses are NOT altered. The switching in the provider cloud is based on MAC addresses as members of provider sVLANs.

Figure 33 Provider bridging / sVLAN operation



The following sections describe sVLAN operation:

- “Components,” next
- “Switch levels” on page 139
- “UNI port behavior” on page 140
- “NNI port behavior” on page 140
- “sVLAN and SMLT” on page 141

Components

sVLAN uses a User-to-Network Interface (UNI) for user access—that is, the ports to which the customer routers/switches connect; and a Network-to-Network Interface (NNI) in the core—that is, the links which interconnect core switches together within the sVLAN network.

Table 15 lists and describes the components used in sVLAN operation.

Table 15 sVLAN components

Component	Definition
User-to-Network (UNI) interface	Customer-facing sVLAN port that accepts any frame type (802.1Q tagged or untagged) and switches it transparently through an sVLAN. This concept is very similar to an untagged port-based VLAN port, except that tagged and untagged packets are bridged transparently within the sVLAN.
Network-to-Network (NNI) interface	Service provider core port that interconnects switches by adding a NEW .1Q-like 4 byte tag after the Dst/Src MAC pair – and in front of the .1Q tag which may have already been inserted.
Switch levels	<p>Allows stacking of multiple .1Q tags, in an <i>onion</i> architecture.</p> <ul style="list-style-type: none"> Level 0 (normal port): 802.1Q frames are classified into port-based VLANs. Levels 1-n (UNI, NNI ports): any frame type is transparently switched and is pre-pended with 4 additional .1Q-like bytes. <p>UNI and NNI ports are expecting only frames of the same level. Otherwise traffic is encapsulated into the next level.</p>

8600 modules with multiple physical ports (8648TX, 8616SX, 8632TX, and 8616TX modules) share a common OctaPID. All ports on the same OctaPID must be configured either as normal ports or as UNI/NNI ports. For example, if port 1 on an 8648TX module is configured as a UNI port, then the remaining ports on that OctaPID (ports 2 to 8) must be configured either as UNI ports or NNI ports—they cannot be configured as normal tagged ports.

Switch levels

Stacked VLANs are designed to provide a very scalable hierarchical solution with up to 8 levels. The first layer of the hierarchy is considered to be the user access layer. User traffic can include tagged or untagged traffic. In the case of tagged traffic, the user packets will contain the normal 802.1p/Q tag with the standard Ether-type value of 8100. The subsequent levels within the sVLAN hierarchy are configured to use a different Ether-type than the standard value of 8100. The 8600 Series switch is designed with default Ether-type values for each sVLAN level.

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

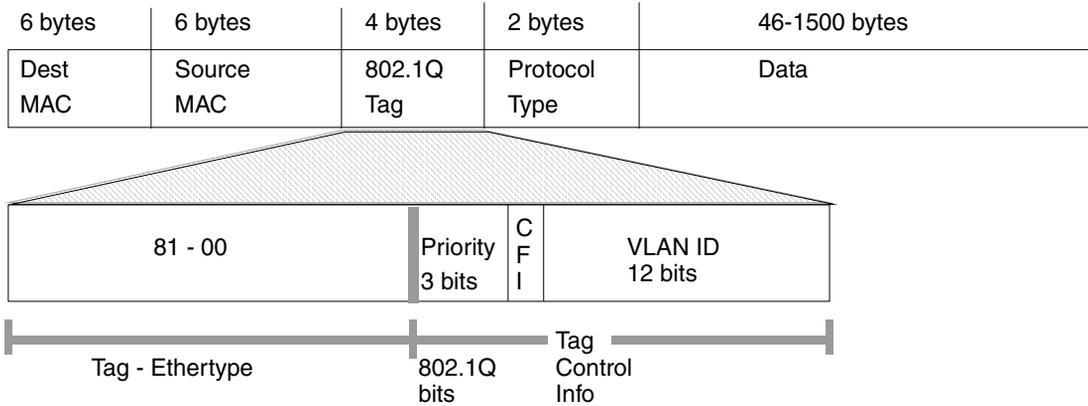
Since each 8600 Series switch can support up to 1980 VLANs, each layer in the sVLAN hierarchy can support up to 1980 VLANs. Given the 8 level hierarchy, the sVLAN network can support thousands of VLANs. This eliminates the issue of VLAN scalability. However, you need to consider certain restrictions when building such networks.

IEEE 802.1Q tag

Provider .1Q-like tags ([Figure 34](#)) have altered Ethertypes. The Ethertype is defined in the sVLAN switch level configuration.

The EtherType 8100 defines a frame as 802.1Q tagged. The 3 priority bits are defined in IEEE 802.1Q as the quality of service bits. The 12 bits for VLAN ID allow for 4096 individual VLAN addressing.

Figure 34 IEEE 802.1Q tag



UNI port behavior

A UNI port is always an untagged, port-based sVLAN. All traffic, untagged or tagged, is classified as a member of the per port configured customer VLAN (sVLAN).

NNI port behavior

An NNI port switches ingressing traffic, based on regular destination MAC lookup, on a per-sVLAN basis. The Ethertype of the 802.1Q tag-like frame has to be equivalent on both sides of the sVLAN NNI link in order for it to correctly switch traffic. Therefore the switch levels of both switches connecting through NNI links with each other must be the same.

sVLAN and SMLT

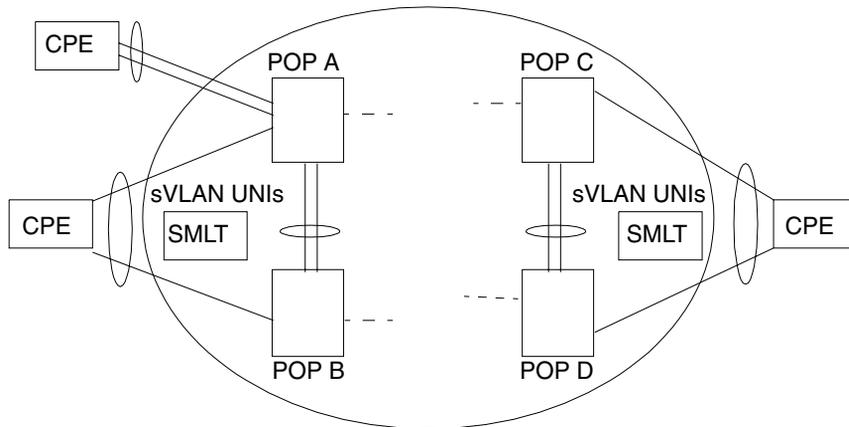
Instead of using Spanning Tree in the provider core, SMLT can be used to provide a redundant architecture.

UNI ports and SMLT

Figure 35 shows dual homing of CPEs to sVLAN UNI ports. The CPE devices are transparent to Q tags. The SMLT IST pairs are:

- POP A and POP B
- POP C and POP D

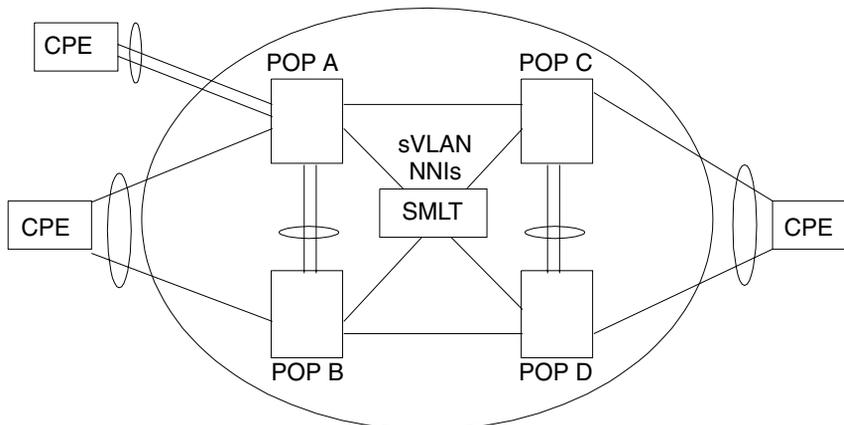
Figure 35 Dual-homing of CPE to sVLAN UNI ports



NNI ports and SMLT

Figure 36 shows an SMLT full mesh core for the sVLAN provider network.

Figure 36 SMLT full mesh core for sVLAN provider network



For more information about designing with SMLT, see [“SMLT” on page 92](#).

Network loop detection and prevention

Customer traffic loops through a provider core can pose a serious threat to network stability. Loops can occur when customers:

- loop traffic back to a redundant connection to one service provider
- loop traffic between two service providers used for redundancy

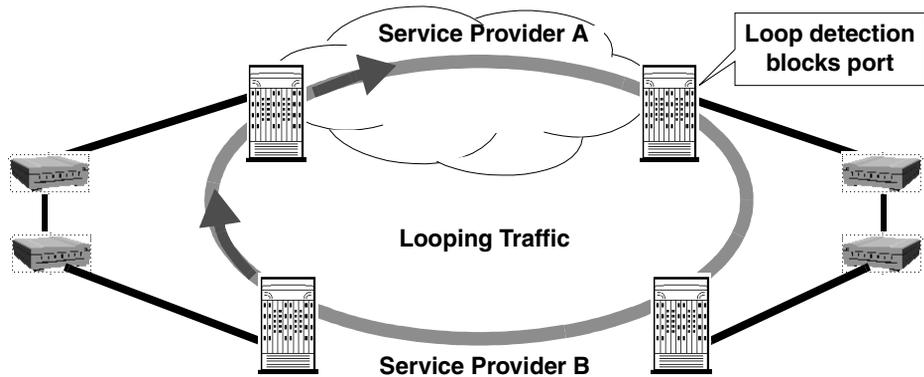
Customer loops result in the following:

- Looping packets saturate the pipes.
- The same MAC addresses will be learned on sVLAN UNI and NNI ports in a rapid sequence.

In either case, the customer’s service is completely shut down. Loops could lead to high control plane utilization because the core switch has to relearn the MAC addresses during its non-stop flapping from the UNI port to the NNI port.

Figure 37 shows how the port loop detection feature discovers loops and disables VLAN on the port.

Figure 37 Customer traffic loops through a service provider core



Note: Loop-detection should be enabled on all UNI customer ports and on SMLT links. Loop-detection should NOT be enabled on IST links. To enable loop detection from Device Manager, select Edit > Port > VLAN > LoopDetect.

Loop detection is triggered when a MAC flaps between two or more ports x number of times within a timer interval of y . A Trap is sent to the management stations and a log entry indicates that a loop occurred.

A VLAN that was disabled when a loop was detected can be enabled using the following command:

```
config ethernet <slot/port> action clearLoopDetectAlarm
```

sVLAN multi-level onion architecture

It is possible to design multi-level sVLANs to increase VLAN scalability. MAC bridging limitations still apply, including MAC learning.

Figure 38 shows the structure of a one-level design. The customer-facing ports on Level 1 devices are sVLAN UNI ports. The core connections are sVLAN NNI ports.

Figure 38 One-level sVLAN design

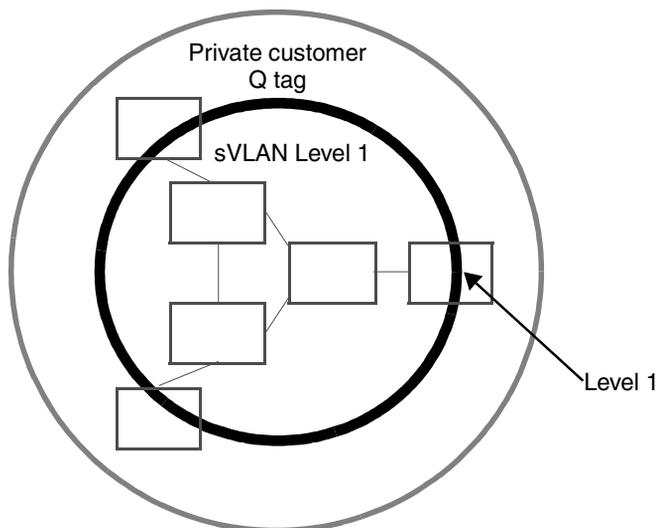


Figure 39 shows the structure of a two-level design. The level 2 facing ports on the level 1 devices are sVLAN NNI ports. The level 1 facing ports on the level 2 devices are sVLAN UNI ports. The ports within the level 2 domain are sVLAN NNI ports.

Figure 39 Two-level sVLAN design

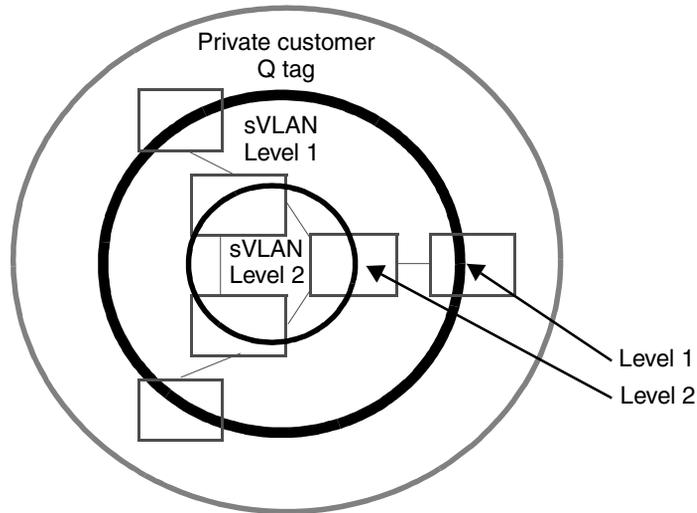
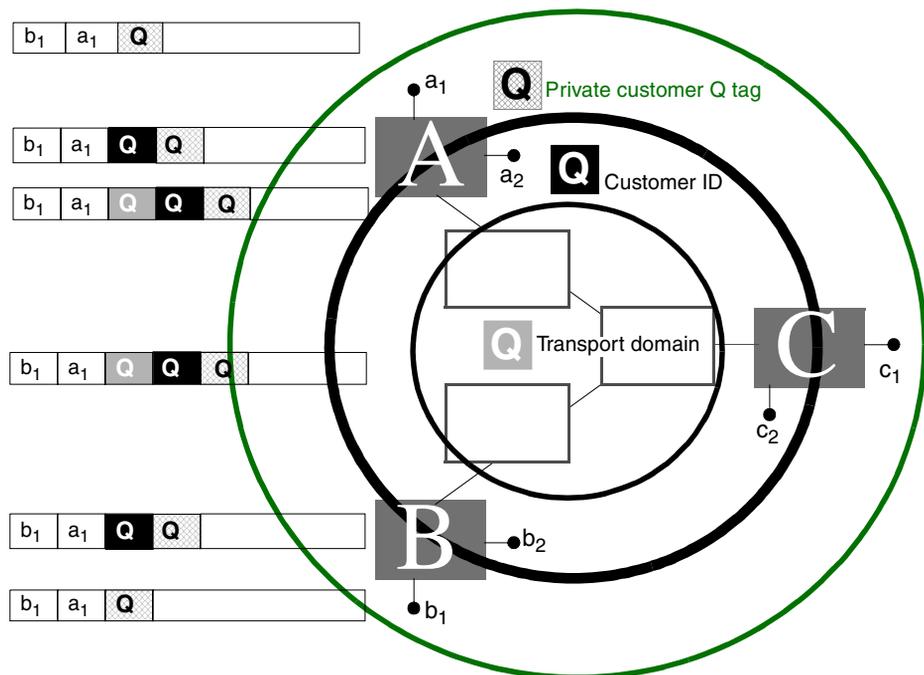


Figure 40 shows the MAC addresses and the Q and Q-like headers in a multi-level onion design sVLAN.

Figure 40 Multi-level onion design sVLAN with Q tags



Network level requirements

Since sVLAN is based on regular VLAN bridging, all MAC addresses of an sVLAN are seen by all provider switches having this sVLAN provisioned. In this architecture, for a regular 8600 E-type module, 24k total MAC addresses are supported. The M-type modules scale up to over 100k MAC addresses.

Independent VLAN learning limitation

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

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

When multiple sVLAN levels are used, sVLANs are aggregated into another level, which could introduce duplicate MAC addresses, learned on different ports. The result is a flapping MAC address from the provider NNI port to another provider NNI port, or a customer UNI port.

Duplicate MAC addresses can be very common for control traffic such as VRRP where VRRP SRC MAC addresses are defined by the IETF RFC and are therefore used by many customers.

To overcome such issues, it is recommended that you connect routers to UNI ports, limiting the amount of MAC addresses and the potential for duplicate MAC addresses.

sVLAN and network or device management

Normal VLANs are currently not supported on sVLAN NNI links. In order to transport regular VLANs in an sVLAN network, it is recommended that you use separate links between the core devices.

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

sVLAN restrictions

The following are sVLAN restrictions.

- For 8648 and 8632 modules, the eight 10/100 ports that share an OctaPID must run in the same mode—either normal or sVLAN UNI/NNI.
- For 8616 modules, the two GIG ports that share an OctaPID must run in the same mode—either normal or sVLAN UNI/NNI

- 8672 and 8684 modules do not support sVLAN.
- SVAN NNI ports do not support normal VLANs (non-sVLANs)
- Routing is not supported on sVLANs.
- IP filters are not supported on sVLAN.
- QoS can be applied through sVLAN QoS only (no filter support).
- sVLAN switches cannot be managed In-band—an out of band network is recommended for management. Connect the Management Ethernet Port to a separate Management sVLAN, and bridge it to the NMS segment.

For information about configuring sVLAN using Device Manager or the CLI, see the publication, *Configuring Layer 2 Operations: VLANs, Spanning Tree, and Multilink trunking*.

Chapter 4

Designing Layer 3 switched networks

This chapter describes some general design considerations you need to be aware of when designing Layer 3 switched networks. Design factors for the following protocols are presented here:

Topic	Page number
VRRP	next
ICMP redirect messages	152
Subnet-based VLANs	155
PPPoE protocol-based VLAN design	157
BGP	163
OSPF	172
IPX	180
IP routed interface scaling considerations	182

VRRP

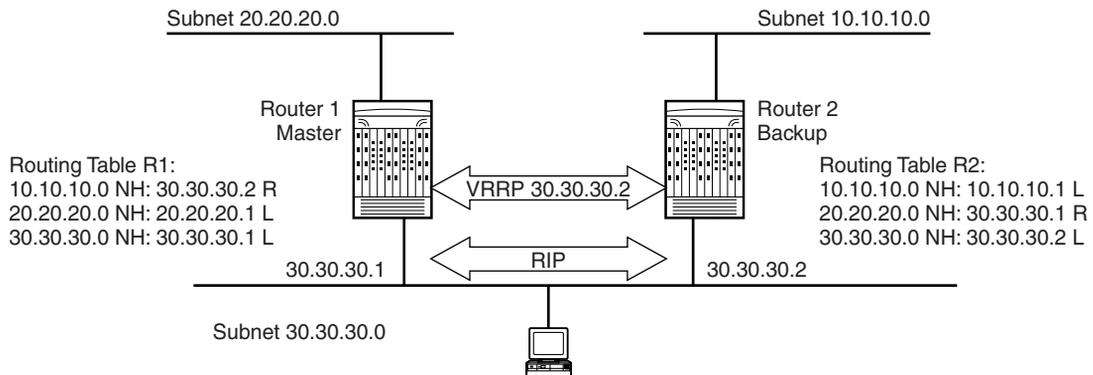
The following design guidelines apply to VRRP.

VRRP and other routing protocols

If, on an IP interface, VRRP and another IP protocol such as OSPF, RIP, or DVMRP are configured, Nortel Networks recommends that you do not use a physical IP address as the virtual IP address.¹ Instead, use a third IP address. Using the physical IP address as the virtual IP address can lead to malfunctioning of the routing protocol in certain circumstances (Figure 41).

When backup master is enabled, it is recommended that with SMLT, you ensure that the virtual IP address and VLAN IP address are not the same.

Figure 41 Sharing the same IP address



10626EA

When VRRP and routing protocols are on the interface, an issue occurs when sharing the same IP address as shown in Figure 41.

¹ When backup-master is enabled on the switch, the VRRP virtual IP address and VLAN IP address cannot be the same.

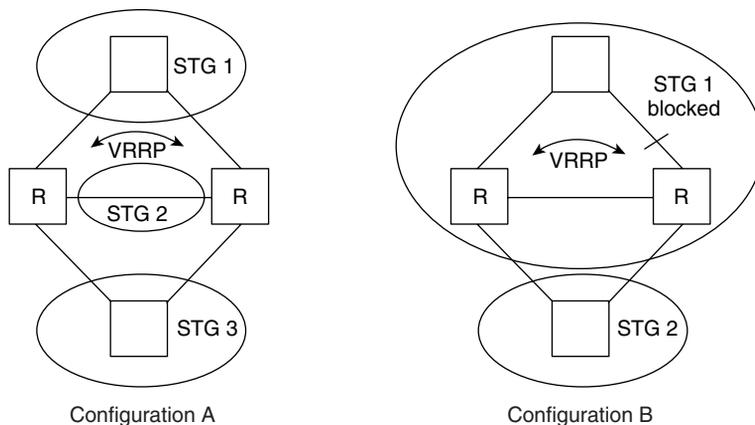
In this example, confusion arises because R1's routing table shows 30.30.30.2 as reaching network 10.10.10.0. Address 30.30.30.2 is local (VRRP Master) to R1, so it does not send traffic to R2. As a result, the traffic is dropped locally. To address this problem, you should use a different IP address for VRRP, other than the local address if a routing protocol is enabled on the VRRP interfaces.

VRRP and STG

Figure 42 shows two possible configurations of VRRP and STG. VRRP protects clients and servers from link or aggregation switch failures and your network configuration should limit the amount of time a link is down during VRRP convergence.

In Figure 42, configuration A is optimal because VRRP convergence occurs within 2-3 seconds. In configuration A, three STGs are configured with VRRP running on the link between the two routers (R). STG2 is configured on the link between the two routers, thus separating the link between the two routers from the STGs found on the other devices. All uplinks are active.

Figure 42 VRRP and STG configurations

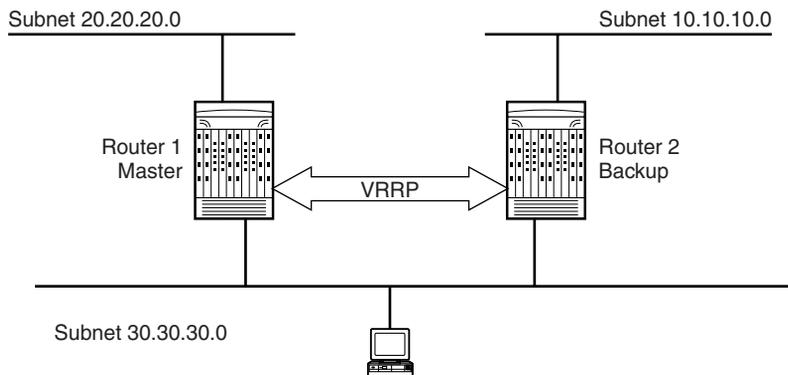


In configuration B, VRRP convergence takes between 30 and 45 seconds because it depends on spanning tree convergence. After initial convergence, spanning tree blocks one link, an uplink, and so only one uplink is used. If an error occurs on the uplink, spanning tree reconverges, which can take up to 45 seconds. After reconvergence, VRRP can take a few more seconds to failover. For VRRP and SMLT information, refer to [“Layer 3 traffic load sharing”](#) on page 103.

ICMP redirect messages

Traffic from the client on subnet 30.30.30.0 destined for the 10.10.10.0 subnet is sent to routing switch 1 (VRRRP Master) in [Figure 43](#). It is then forwarded *on the same subnet* to routing switch 2 where it is routed to the destination. Routing switch 1 sends an ICMP redirect message for each packet received to the client to inform him of a shorter path to the destination through routing switch 2.

Figure 43 ICMP redirect messages diagram

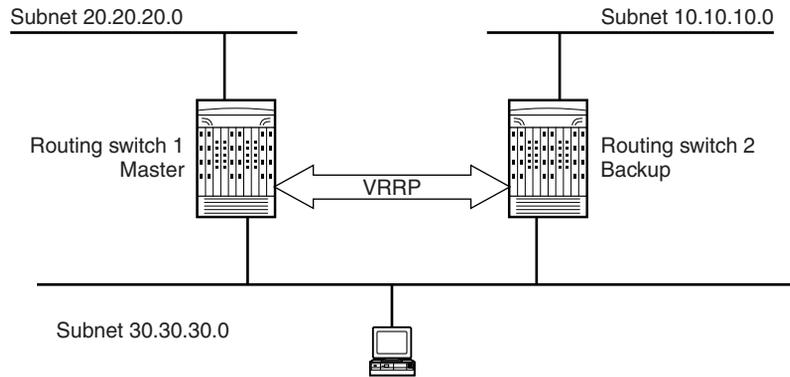


10627EA

Avoiding excessive ICMP redirect messages

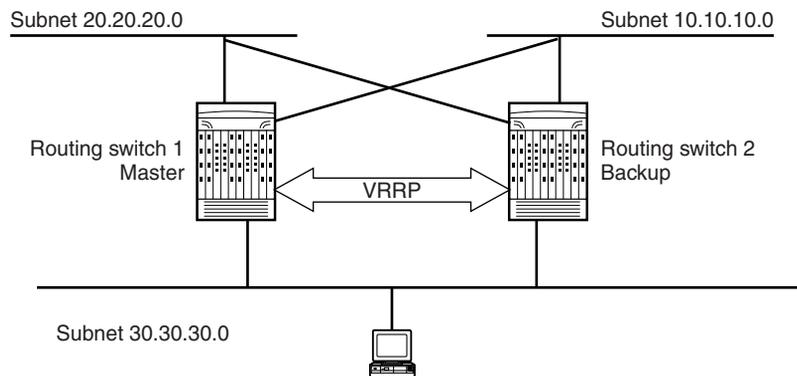
If network clients do not recognize ICMP redirect messages, there are three different network designs you can use to avoid excessive ICMP redirect messages. Option 3 is the one that Nortel Networks recommends.

Option 1 is shown in [Figure 44](#). Here, you enable ICMP redirect generation on the routing switches to let the client learn the new shorter path to the destination. The clients then populate a route entry in their routing table that uses a direct path to the destination through routing switch 2.

Figure 44 Avoiding excessive ICMP redirect messages- option 1

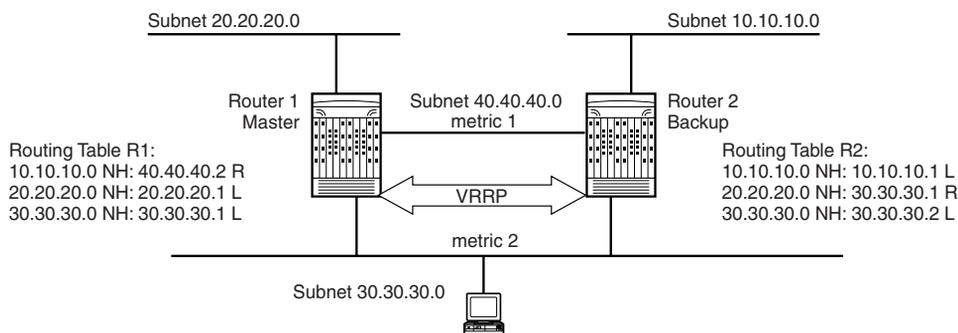
10628EA

Option 2 is shown in [Figure 45](#). Here, you 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. You do this by building symmetrical networks based upon the network design examples presented in [Chapter 2, “Designing redundant networks,”](#) on [page 53](#).

Figure 45 Avoiding excessive ICMP redirect messages- option 2

10629EA

Option 3, the recommended option, is shown graphically in [Figure 46](#). It includes a routed link 40.40.40.0 between routing switch 1 and routing switch 2 with the lowest metric (1). If you increase the metric to 2 or greater on access subnet 30.30.30.3, routing switch 1 uses the inter-switch link to send traffic to routing switch 2 to reach network 10.10.10.0 and no longer issues a redirect message.

Figure 46 Avoiding excessive ICMP redirect messages- option 3

10626BEA

Subnet-based VLANs

You can use subnet-based VLANs to classify end users in a VLAN based on their source IP addresses. For each packet, the switch performs a look-up and based on the source IP address and mask, determines which VLAN the traffic is classified in. You can also use subnet-based VLANs for security reasons to allow only users on the appropriate IP subnet to access to the network. Note that you cannot classify non-IP traffic in a subnet-based VLAN.

Subnet-based VLAN and IP routing

You can enable routing in each subnet-based VLAN. You do so 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.

Subnet based VLAN and VRRP

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

Subnet-based VLAN and multinetting

You can use subnet-based VLANs to achieve a multinetting functionality. The important difference here is that multiple subnet-based VLANs on a port can only classify traffic based on the sender's IP source address. Thus, you cannot multinet by using multiple subnet-based VLANs between routers (L3 devices). Multinetting is supported, however, on all "enduser-facing" ports.

Subnet-based VLAN and DHCP

You cannot classify Dynamic Host Configuration Protocol (DHCP) traffic into subnet-based VLANs because DHCP requests do not carry a specific source IP address, but an all broadcast address. To support DHCP to classify subnet-based VLAN members, you must 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 enduser traffic is classified appropriately into the subnet-based VLAN.

Subnet-based VLAN scalability

The switch supports a maximum number of 300 subnet-based VLANs.

Subnet-based VLAN and wireless terminals

Subnet-based VLANs are incompatible with some wireless terminals. This is especially true in those configurations where you use the Passport 8600 as a classification device (i.e., an IP subnet-based VLAN and a port-based VLAN configured on the same port).

During the roaming phase, wireless terminals may lose the session with their application servers. This is because of the absence of the IP header in the frames that these terminals can send during this roaming phase. Thus, the frames are sent in the port-based VLAN, and not in the IP subnet-based VLAN. Previously, the IP subnet-based VLAN was used to isolate these terminals. When designing your network, it is recommended that you ensure that your wireless access devices are operating correctly.

PPPoE protocol-based VLAN design

Point-to-Point Protocol over Ethernet (PPPoE) allows you to connect multiple computers on Ethernet to a remote site through a device such as a modem. You can use PPPoE to allow multiple users (for example, an office environment, or a building with many users) to share a common line connection to the Internet.

PPPoE combines the Point-to-Point (PPP) protocol, commonly used in dial-up connections, with the Ethernet protocol, which supports multiple users in a local area network. The PPP protocol information is encapsulated within an Ethernet frame (see RFC 2516: Point-to-Point Protocol over Ethernet).

The example in this section shows how to use PPPoE protocol-based VLANs, a feature introduced in release 3.5, to redirect PPPoE Internet traffic to a service provider network, while the IP traffic goes to a routed network. The example uses two features introduced in the 3.5 release:

- PPPoE protocol-based VLANs
- Disabling IP routing per port in a routed VLAN

This example can be used in a service provider application to redirect subscriber Internet traffic to a separate network from the IP routed network. It can also apply to enterprise networks that need to isolate PPPoE traffic from the routed IP traffic, even when this traffic is received on the same VLAN.

Implementing bridged PPPoE and IP traffic isolation

This example shows a configuration with bridged PPPoE and IP traffic isolation to achieve the following goals:

- Enable users to generate IP and PPPoE traffic where IP traffic needs to be routed and PPPoE traffic needs to be bridged to the ISP network. If any other type of traffic is generated, it is dropped by the Layer 2 switch or the 8600 Series switch (when users are attached directly to the 8600).
- Each user is assigned a different VLAN from other users (that is, every subscriber is assigned a VLAN).
- Each user has two VLANs when directly connected to the 8600—one for IP traffic and the other for PPPoE traffic.
- PPPoE bridged traffic must preserve user VLANs.

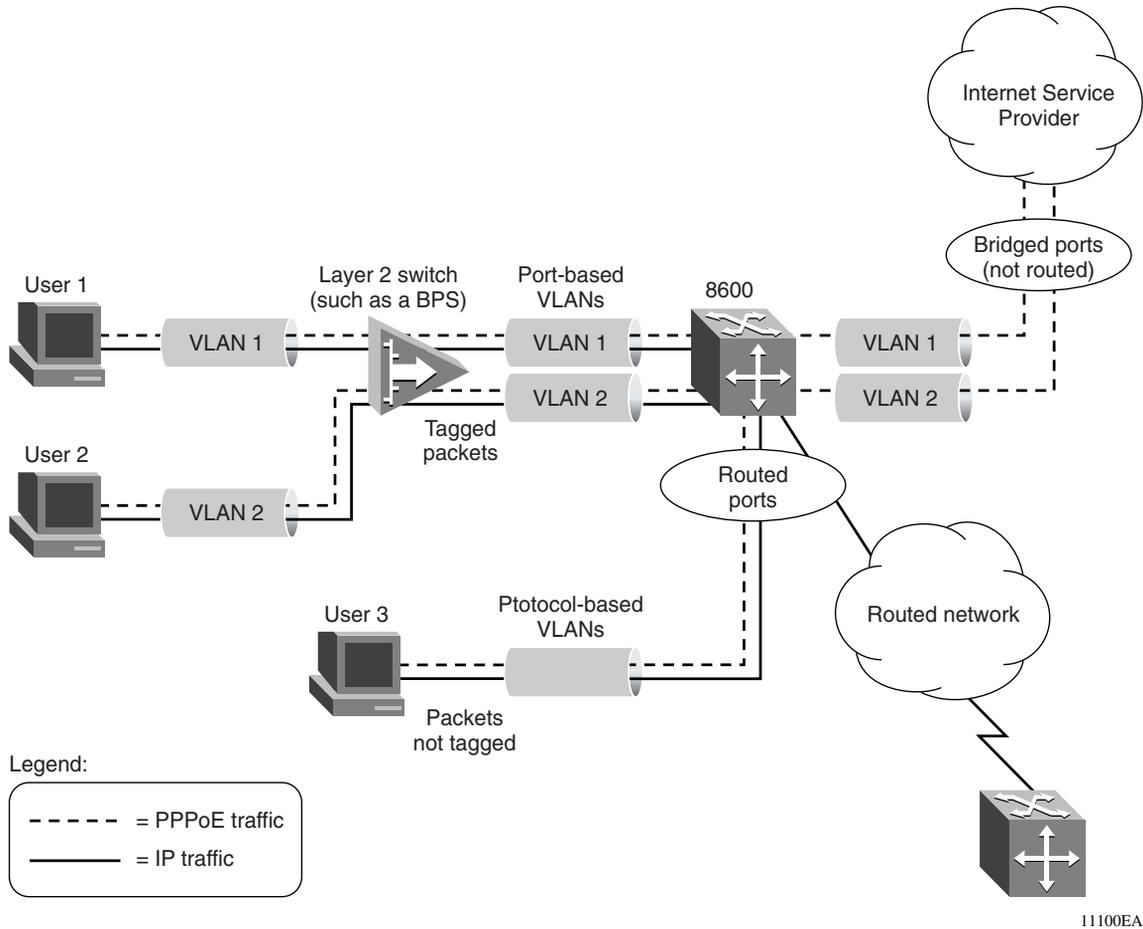
In this example, consider the following two aspects of the configuration:

- [Indirect connections](#) where users are attached to a Layer 2 switch
- [Direct connections](#) where users are attached directly to the 8600 Series switch.

In [Figure 47 on page 159](#), both PPPoE and IP traffic are flowing through the network. Below are some assumptions and configuration requirements:

- PPPoE packets between the users and the ISP are bridged.
- Packets received from the Layer 2 switch are tagged, while packets received from the directly connected user (User 3) are not tagged.
- IP packets between the user and the 8600 are bridged, while packets between the 8600 and the routed network are routed.
- VLANs between the Layer 2 switch and the 8600 are port-based.
- VLANS from the directly connected user (User 3) are protocol-based.
- The connection between the 8600 and the ISP is a single port connection.
- The connection between the Layer 2 switch and the 8600 can be a single port connection or a MultiLink Trunk (MLT) connection.
- 8600 ports connected to the user side (Users 1, 2, and 3) and the routed network, are routed ports.
- 8600 ports connected to the ISP side are bridged (not routed) ports.

Figure 47 PPPoE and IP traffic separation



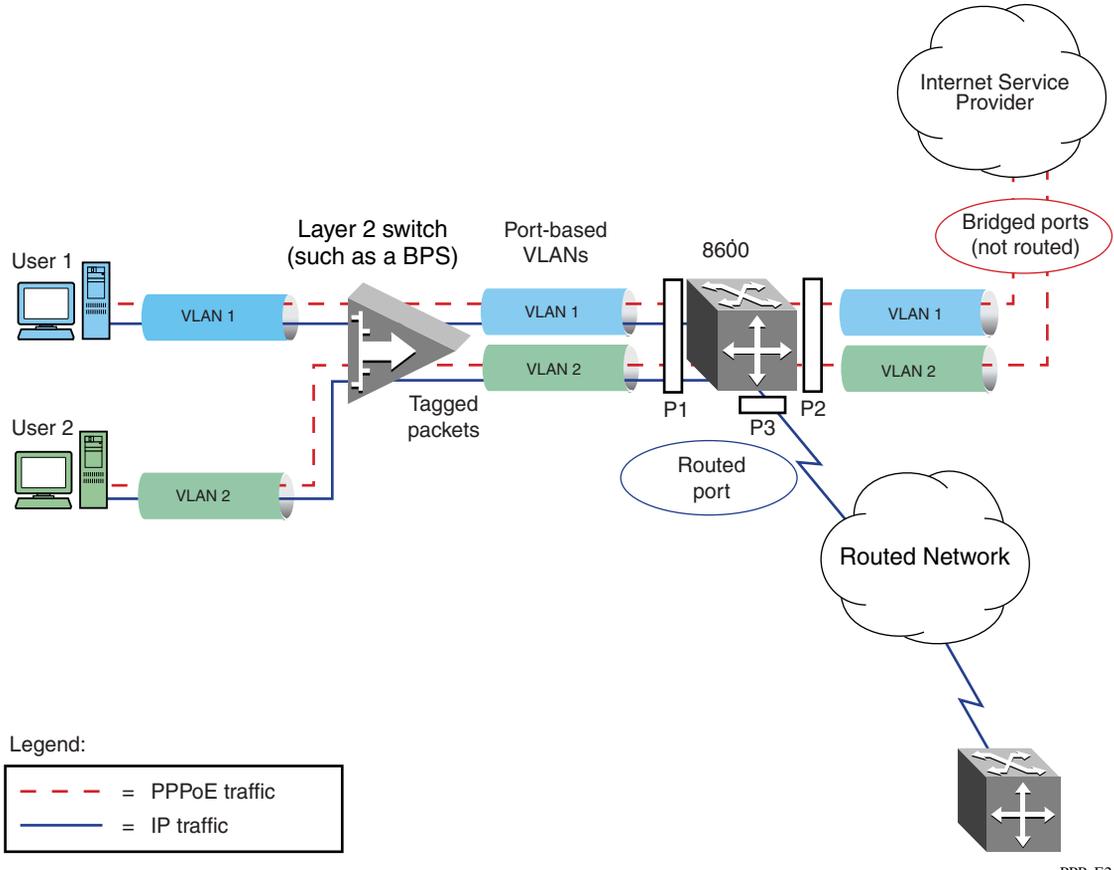
Indirect connections

Figure 48 on page 161 shows that the 8600 Series switch uses 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 non-tagging (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 this 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 48 Indirect PPPoE and IP configuration

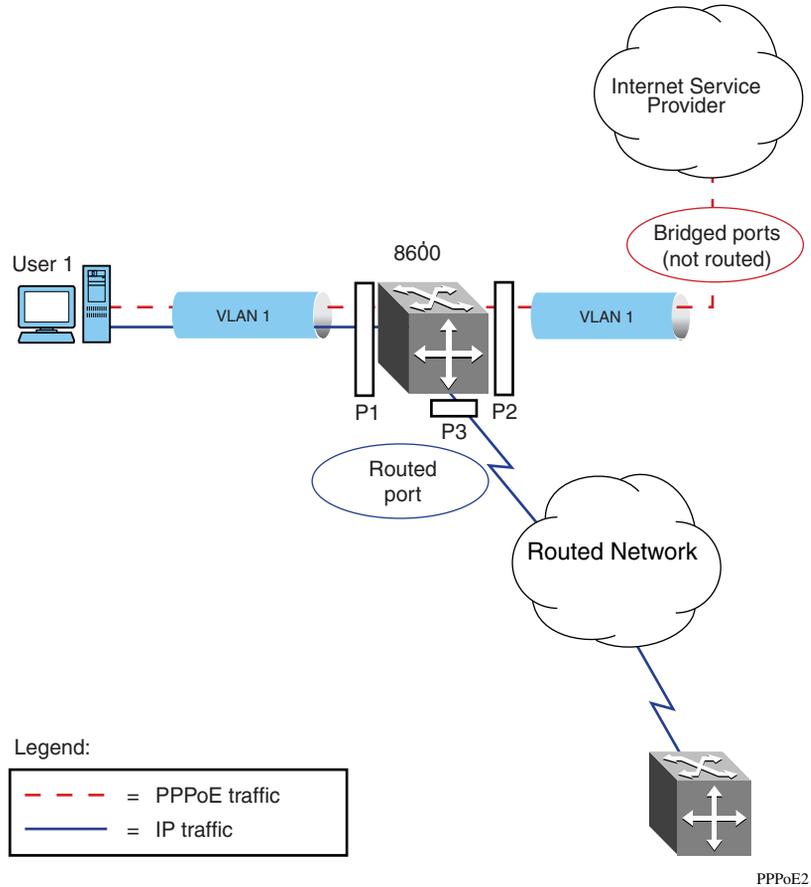


Direct connections

Figure 49 on page 163 shows that, to directly connect to the passport 8600 switch, a user must create two protocol-based VLANs on the port—one for PPPoE traffic and one for IP traffic. 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 CDN network.
P3 can be configured for either tagging or non-tagging (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. Disabling routing per port is not required given that the routed IP VLANs are not configured on port 2 as they are for indirect connections.

Figure 49 Direct PPPoE and IP configuration

BGP

This section provides a general overview, hardware and software dependencies, scaling information, convergence performance, design scenarios, and OSPF interactions for Border Gateway Protocol (BGP).

Overview

Since release 3.3 of the Passport 8000 Series software, the Passport 8600 includes BGP4 functionality. BGP is an exterior gateway protocol designed to exchange network reachability information with other BGP systems in other autonomous systems, or within the same autonomous system (AS). This network reachability information includes information on the AS list that reachability information traverses. This information is sufficient to construct a graph of AS connectivity from which you may prune routing loops and enforce some policy decisions at the AS level.

BGP4 provides you with a new set of mechanisms for supporting classless inter-domain routing. These mechanisms include support for advertising an IP prefix and eliminate the concept of network *class* within BGP. BGP4 also introduces mechanisms, which allow you to aggregate routes, including aggregating AS paths. Note that BGP aggregation does not occur when routes have different multi- exit discs or next hops.

Hardware and software dependencies

The table that follows describes the software and hardware necessary to run BGP.

Software	Hardware
Passport 8000 Series software version 3.3 or above	BGP supported on: <ul style="list-style-type: none">• all I/O modules• on both switch fabric 8690 and 8691 Note: For large BGP environments, Nortel Networks recommends you use the 8691SF.

Scaling considerations

Scaling considerations include:

- BGP peering
- route management
- ECMP support

Each of these are explained in the subsections that follow.

BGP peering

BGP allows you to create routing between two sets of routers operating in different administrative systems. A group of routers that operates in two distinct systems is an AS. An AS can use two kinds of BGP methods:

- Interior BGP (IBGP) - refers to routers that use BGP within an autonomous system. BGP information is redistributed to Interior Gateway Protocols (IGPs) running in the autonomous path.
- Exterior BGP (EBGP) - refers to routers that use BGP across two different autonomous paths.

The Passport 8600 supports a maximum of 10 peers both internal and external. Note that there is no software restriction that prevents you from configuring more than 10 peers. It is recommended that you contact your Nortel Networks sales representative for the evolution of the BGP scaling numbers.

Route management

The number of supported routes include the maximum number of forwarding routes on the I/O modules for:

- 32K modules (including normal and E-modules) = 20,000
- M-modules (128K) = 119,000

Refer to the *Release Notes for the Passport 8000 Series Switch Software Release 3.5* for the latest scalability numbers for route forwarding.

ECMP support

BGP equal-cost multipath (ECMP) support 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, with a maximum of 4 next hop entries per equal cost path.

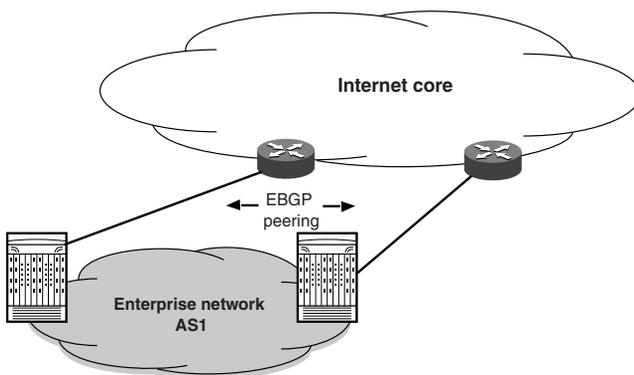
Design scenarios

In situations with a maximum of 10 peers and 100K routes, the Passport 8600 operates as an ideal BGP edge device. Note that the Passport 8600 is currently not positioned as a core Internet BGP router. The following design scenarios describe more typical Passport 8600 BGP applications.

Internet peering

With BGP functionality on the Passport 8600 platform, you can perform Internet peering directly between the Passport 8600 and another edge router. In such a scenario, you use each Passport 8600 for aggregation and peer it with a Layer 3 edge router (Figure 50).

Figure 50 Internet peering

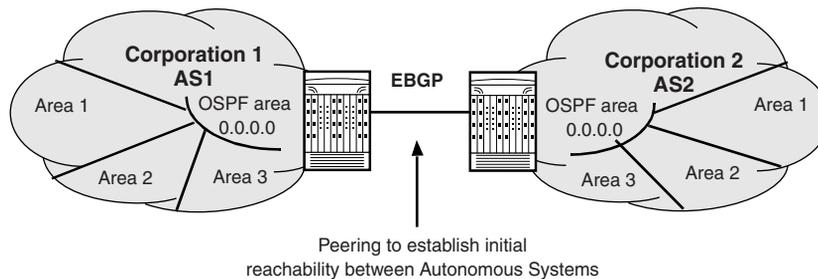


In cases where the Internet connection is single-homed, it is recommended that you advertise Internet routes as a default route to the IGP in order to reduce the size of the routing table.

BGP applications to connect to an AS

You can implement BGP with the Passport 8600, so that autonomous routing domains, such as OSPF routing domains, are connected. This strategy effectively allows the two different networks to begin communicating quickly over a common infrastructure, thus allowing network designers additional time to plan the IGP merger. Such a scenario is particularly effective when network administrators wish to merge two OSPF area 0.0.0.0's (Figure 51).

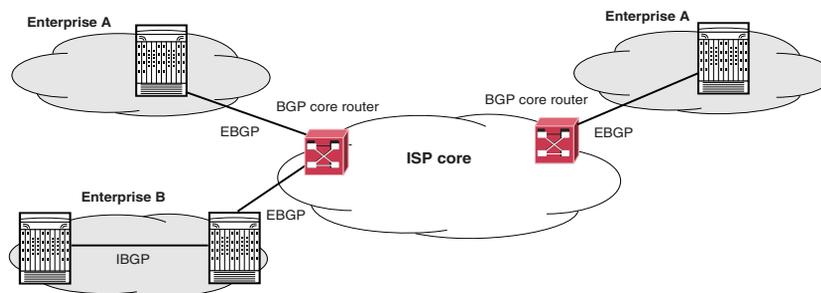
Figure 51 BGP's role to connect to an AS



Edge aggregation

You can use the Passport 8600 to perform edge aggregation with multiple/PoP edge concentrations. The Passport 8600 provides GE or 10/100 EBGP peering services to the enterprise. Should you wish to inter-work with Multiprotocol Label Switching (MPLS)/Virtual Private Network (VPN) (RFC 2547) services at the edge, this particular scenario is ideal. You use BGP here to inject dynamic routes, instead of using static routes or RIP (Figure 52).

Figure 52 Edge aggregation



ISP segmentation

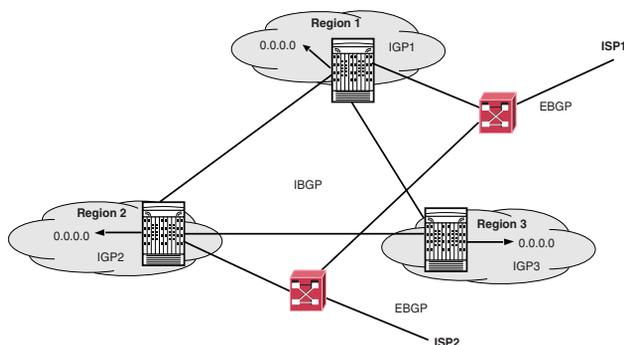
You can also use the Passport 8600 as a peering point between different regions or autonomous systems that belong to the same ISP. In such cases, you may define a region as an OSPF area, AS, or a part of an AS.

Multiple regions separated by IBGP

You can divide the AS into multiple regions, each running different IGPs. You interconnect regions logically via a full IBGP mesh. Each region then injects its IGP routes into IBGP and injects a default route inside the region. Thus, each region defaults to the BGP border router for destinations that do not belong to the region.

You can then use the *community* attribute to differentiate between regions. You can also use this 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 (Figure 53).

Figure 53 Multiple regions separated by IBGP



In Figure 53, note the following:

- The AS is divided into 3 regions, each running different and independent IGPs
- Regions are logically interconnected via a full mesh IBGP, which also provides Internet connectivity

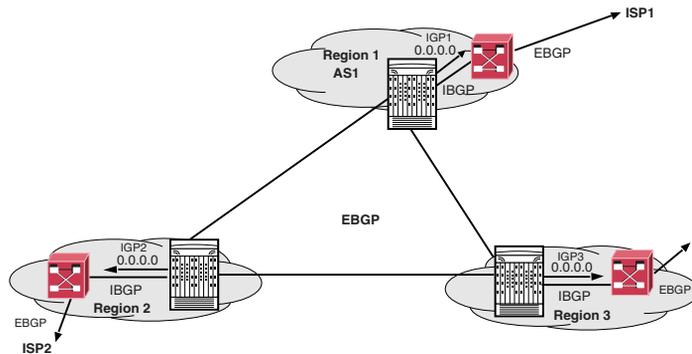
- Internal non-BGP routers in each region default to the BGP border, which contains all routes
- If the destination belongs to any other region, the traffic is directed to that region; otherwise, the traffic is sent to the Internet connections according to BGP policies

Multiple regions separated by EBGP

If you need to set multiple policies between regions, you can represent each region as a separate AS. You then implement EBGP between ASs, while IBGP is implemented 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.

You can obtain AS numbers from the Inter-Network Information Center (NIC), or by using private AS numbers. When using the latter, be sure to design your Internet connectivity very carefully. For example, you may wish to 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 then strips the private AS numbers to the Internet in order to prevent them from leaking to the providers.

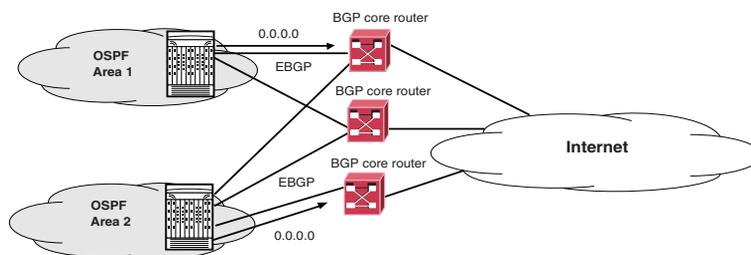
Figure 54 Multiple regions separated by EBGP



Multiple OSPF regions peering with Internet

Figure 55 illustrates a design scenario in which you use multiple OSPF regions to peer with the Internet.

Figure 55 Multiple OSPF regions peering with the Internet



Multi-homed to non-transit AS/single provider

To control route propagation and filtering, it is recommended in RFCs 1772 and 2270 (and often by the providers themselves) that multi-homed, non-transit Autonomous Systems *not* run BGP4. To address the load sharing and reliability requirements of a multi-homed customer, you should instead use BGP between them.

Considerations

When configuring BGP, be aware of the following design considerations:

- A default parameter (max-prefix) limits the number of imported routes from a peer. (The default value is set to 12000). The purpose of this parameter is to prevent non-M mode configurations from accepting more routes than it can forward to.

It is recommended that you use a setting of 0 to accept an unlimited number of prefixes. For instructions on modifying this parameter, see *Configuring BGP Services* in the Passport 8000 Series documentation set.

- BGP will not operate with an IP router in non-forwarding (host-only) mode. Thus, you should ensure that the routers you want BGP to operate with are in forwarding mode.

- If you are using BGP for a multi-homed AS (one that contains more than a single exit point), Nortel Networks recommends that you use OSPF for your IGP and BGP for your sole exterior gateway protocol. Otherwise, you should use intra-AS IBGP routing.
- If OSPF is the IGP, use the default OSPF tag construction. Using EGP or modifying 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 (i.e., the typical case for IBGP speakers), consider using the address of the router's circuitless (virtual) IP interface as the local peer address. In this way, you ensure that BGP is reachable as long as there is an active circuit on the router.
- By default, BGP speakers do not advertise or inject routes into its 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. You should also make consistent routing policy decisions on external BGP connections.
- No current option is available to allow you to enable/disable the Multi-Exit Discriminator selection process.
- You cannot disable the aggregation when routes have different MEDs (MULTI_EXIT_DISC) or NEXT_HOP.

For a complete list of other release considerations, see the *Release Notes for the Passport 8000 Series Switch Software Release 3.5*.

Interoperability

BGP interoperability has been successfully demonstrated between the Passport 8000 Series software release 3.3, Cisco 6500 software release IOS 11.3, and Juniper M20 software release 5.3R2.4. Refer to *Configuring BGP Services* for more information and the list of CLI commands corresponding to the Nortel Networks BGP implementation in the Passport 8600.

OSPF

This section describes some general design considerations and presents a number of design scenarios for OSPF.

Scalability guidelines

You should follow these OSPF scalability guidelines:

- Maximum number of supported OSPF areas per switch: 5
- Maximum number of total OSPF adjacencies per switch: 80
- Maximum number of total routes per switch: 15k

To determine OSPF link state advertisement (LSA) limits:

- Use the CLI command `show ip ospf area` to determine the LSA_CNT and to obtain the number of LSAs for a given area.
- Use the following formula to determine the number of areas:

$$\sum \text{Adj}_N * \text{LSA_CNT}_N < 40k$$

N: from 1 to number of areas per switch

Adj_N = number of Adjacencies per Area N

LSA_CNT_N = Number of LSAs per Area N

For example, assume that a switch has a configuration of 3 areas with a total of 18 adjacencies and 1k 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)

You can then calculate the scalability formula as follows:

$$3*500+10*1000+5*200=12.5k < 40k$$

This ensures that the switch is operating within accepted scalability limits.

Design guidelines

Nortel Networks recommends that you stay within the previously-mentioned boundaries when designing OSPF networks. Follow these 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

Typically, you should configure wiring closet subnets as OSPF passive interfaces unless they form a legitimate routing path for other routes.

- Limit the number of OSPF areas per switch to as few as possible to avoid excessive shortest path calculations

Be aware that the Passport switch has to execute the Dijkstra algorithm for each area separately.



Note: The limits mentioned here are not hard limits, but a result of scalability testing with switches under load with other protocols running in the network. (The other protocols are not scaled to the limits). Depending upon your network design, these number may vary.

- Ensure that the OSPF dead interval is at least 4 times the OSPF hello interval

OSPF route summarization and black hole routes

When you create an OSPF area route summary on an area boundary router (ABR), be aware that the summary route can attract traffic to the ABR that it does not have a specific destination route for. If you have enabled ICMP unreachable message generation on the switch, this may result in a high CPU utilization rate.

To avoid such a scenario, Nortel Networks 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 in the routing table is dropped by the hardware.

OSPF network design scenarios

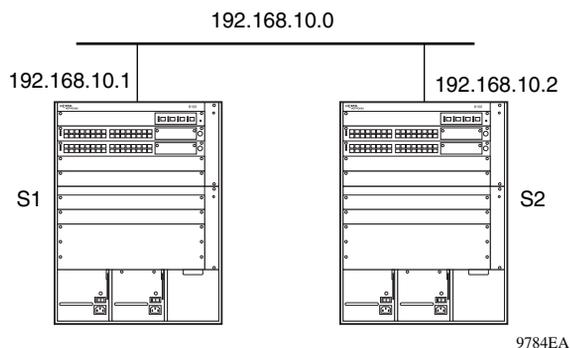
These OSPF network design scenarios are presented in the sections that follow:

- OSPF on one subnet in one area
- OSPF on two subnets in one area
- OSPF on two subnets in two areas

Scenario 1: OSPF on one subnet in one area

Scenario 1 is for 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 ([Figure 56](#)).

Figure 56 Enabling OSPF on one subnet in one area



The routers in scenario 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

In scenario 1, to configure S1 for OSPF, perform the following tasks:

- 1 Enable OSPF globally for the [Product Name (long)] in the IP Routing > OSPF > General window in the JDM or by entering the **config ip ospf admin-state enable** command in the CLI.



Note: OSPF must be globally enabled before any of the following configuration procedures can take effect.

- 2 Verify that IP forwarding is enabled for the switch in the IP Routing > IP > IP window in the JDM or by entering the **config ip forwarding enable** command in the CLI.
- 3 Enter an IP address, subnet mask and VLAN ID for the port in the Edit > Port > IP address insert window in the JDM or by entering the **config ethernet <port> ip create <ipaddr/mask> <vid>** command in the CLI.
- 4 If RIP is not required on the port disable it in the Edit > Port > RIP window in the JDM or by entering the **config ethernet <port> ip rip disable** command in the CLI.
- 5 Enable OSPF for the port in the Edit > Port > OSPF window in the JDM or by entering the **config ip ospf interface 192.168.10.1 admin-status enable** command in the CLI.

When you have completed these tasks, carry out the same sequence of tasks to configure S2 for OSPF, substituting the IP address for S2 in place of the IP address shown in step 5. After you have configured S2, the two switches elect a designated router (DR) and a backup designated router (BDR) and exchange hello packets to synchronize their link state databases.

You can review the relationships between the switches in the JDM or in the CLI by performing the following tasks.

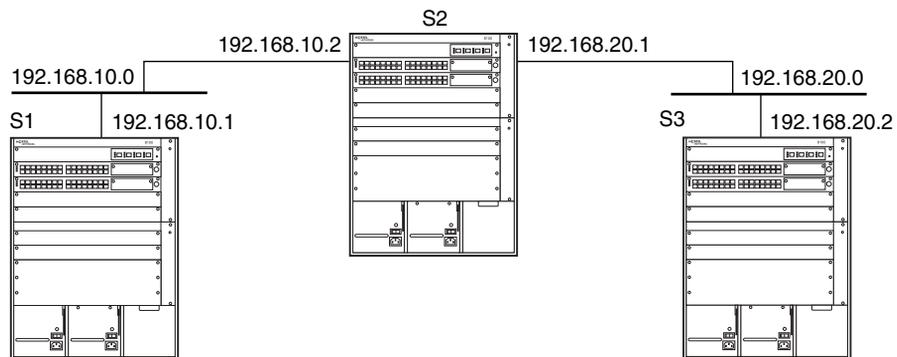
- View which router has been elected as DR and which router has been given the role of BDR either in the IP Routing > OSPF > Interface window in the JDM or by entering the **show ip ospf interface** command in the CLI.
- View the LSAs that were created when the switches synchronized their databases in the IP Routing > OSPF > Link State Database window in the JDM, or by entering the **show ip ospf lsdb** command in the CLI.

- View IP information about neighbors in the IP Routing > OSPF > Neighbors window in the JDM of by entering the `show ip ospf neighbors` command in the CLI.

Scenario 2: OSPF on two subnets in one area

Figure 57 shows a configuration for scenario 2 which enables OSPF on three switches, switch 1 (S1) and switch 2 (S2) and switch 3 (S3), that operate on two subnets in one OSPF area.

Figure 57 Configuring OSPF on two subnets in one area



9786EA

The routers in scenario 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

In scenario 2, to configure OSPF for the three routers perform the following tasks:

- Enable OSPF globally for each router.
- Insert IP addresses, subnet masks, and VLAN IDs for the OSPF ports on S1 and S3 and for the two OSPF ports on S2. Configuring two ports on S2 enables routing and establishes IP addresses related to two networks and two connecting ports.

- Enable OSPF for each of the four OSPF ports that you have allocated IP addresses

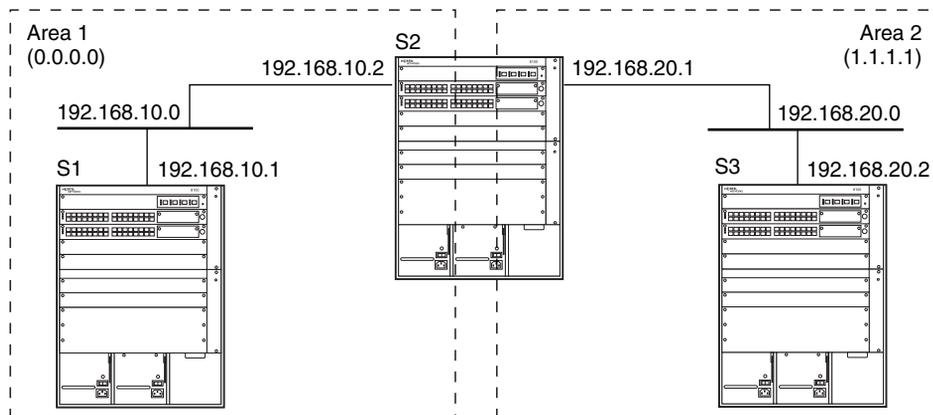
When all three switches are configured for OSPF they will elect a DR and BDR for each subnet and exchange hello packets to synchronize their link state databases.

To review the relationships among the three switches in the OSPF configuration, follow the review procedures described in scenario 1 on [page 174](#). In this scenario S1 is directly connected to S2 and S3 is directly connected to S2, but any traffic between S1 and S3 is indirect, passing through S2.

Scenario 3: OSPF on two subnets in two areas

[Figure 58](#) shows a configuration for scenario 3 which enables OSPF on three switches, S1, S2, and S3, that operate on two subnets in two OSPF areas. S2 becomes the ABR for both networks.

Figure 58 Configuring OSPF on two subnets in two areas



9787EA

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, and 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, and is in OSPF area 2.

To configure OSPF for scenario 3, perform the following tasks in sequence:

- 1 Enable OSPF globally for all three switches
- 2 Configure OSPF on one network.
 - On S1, insert the IP address, subnet mask, and VLAN ID for the OSPF port, and 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



Note: Both routable ports belong to the same network. Therefore, by default, both ports are in the same area.

- 3 Configure 3 OSPF areas for the network in the IP Routing > OSPF > Area > Insert Area window in the JDM or by entering the **config ip ospf area <ipaddr> create** command in the CLI, where *ipaddr* is a dotted decimal notation for the OSPF area.
- 4 Configure OSPF on two additional ports in a second subnet. OSPF is already enabled for the S2 and S3 but you must 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.

All three switches should now be configured for OSPF and should be exchanging hello packets.

When you review the relationships among the three switches in the OSPF configuration note the following:

S2 is confirmed as the ABR because “true” appears in the `AreaBdrRtrStatus` field. In the CLI enter **show ip ospf interface** info.

- View router status either in the IP Routing > OSPF > Interface window in the JDM or by entering the **show ip ospf interface** command in the CLI.
 - S1 is the BDR for area 1
 - S2 is the DR for area 1 and is also the BDR for area 2
 - S3 is the DR for area 2
 - S2 is the ABR for areas 1 and 2
- View neighbor status either in the IP Routing > OSPF > Neighbors window in the JDM or by entering the **show ip ospf neighbors** command in the CLI.
 - S1 has S2 as its only neighbor
 - S2 has both S1 and S3 as neighbors
 - S3 has S2 as its only neighbor
- View the link state advertisements (LSAs) that were created when the switches synchronized their databases in the IP Routing > OSPF > Link State Database window in the JDM, or by entering the **show ip ospf lsdb** command in the CLI.
- View IP routing information either in the IP Routing > IP > IP Route window in the JDM, or by entering the **show ip route info** command in the CLI.



Note: In an environment with a mix of Cisco and Nortel switches/routers, you have to manually modify the OSPF parameter `RtrDeadInterval` to 40 seconds.

IPX



Note: With release 3.3, the Passport 8600 now supports the concept of tick and hop routing. This parameter is a global parameter.

You should be aware of the following IPX design considerations: get nearest server (GNS) and logical link control (LLC) encapsulation and translation. Both of these are explained in the upcoming sections.

GNS

IPX clients use the GNS request to find a server for login. If there is a server available on the same network segment, this server answers the GNS request with a GNS response. If there is no server present, the routing device provides the GNS response.

With release 3.1 and above, Passport chooses the closest Netware server services based on the following algorithm:

- The Passport 8600 switch checks the route cost
- If there are multiple services with the same RIP route cost, the switch uses the lowest SAP hop count
- If multiple services with the same SAP cost are available, the switch responds with the services in alphabetical order, providing a means of load balancing user network logins over multiple servers.

If you encounter connection problems because the [Product Name (long)] is responding with a Netware service that might not be the most optimal, increase hop counts to that Netware server using following the CLI command:

```
config ipx static-route create <IPX-network-number> <nexthop>  
<hop-count> <tick-count>
```

where:

- *IPX-network-number* is the destination IPX network number for the route.
- *nexthop* is the IPX address of the next router.

- *hop-count* is the number of passes through a router.
- *tick-count* is the number of ticks (1/18th of a second).

LLC encapsulation and translation



Note: The Passport 8616SXE module and all other enhanced Gigabit modules (E-modules) support LLC translation to and from Gigabit Ethernet (GE) ports.

LLC translation to and from GE ports is not supported on other modules. To avoid network connection problems, avoid setups that require LLC translation. You can do so by using one encapsulation type throughout your network.

If you have client switches with LLC encapsulation and another encapsulation, do not use LLC encapsulation over the Gigabit Ethernet connection.

IPX RIP/SAP policies

With IPX RIP policies introduced in release 3.3, you can shield the view of networks from users on different network segments by configuring route filters. Route filters give you greater control over the routing of IPX packets from one area of an IPX internetwork to another.

Using route filters helps maximize the use of the available bandwidth throughout the IPX internetwork, and helps improve network security by restricting a user's view of other networks. You can configure inbound and outbound route filters on a per-interface basis, instructing the interface to advertise/accept or drop filtered RIP packets. The action parameter that you define for the filter determines whether the router advertises, accepts, or drops RIP packets from routers that match the filter criteria. The same concept applies to SAP (Service Advertisement Protocol).

See *Configuring IPX Routing Operations* in the Passport 8000 Series documentation for information on configuring IPX RIP/SAP policies.

IP routed interface scaling considerations

Release 3.5 and above allow the support for up to 1980 IP routed interfaces. However, to configure more than 512 IP routed interfaces, you will need the MAC upgrade kit (Part # DS1404015).

There are several considerations that you need to take into account when configuring a large number of IP routed interfaces. Follow the guidelines below:

- Use passive interfaces on most of the configured interfaces. You can only make very few interfaces active. (See below.)
- For DVMRP, you can have up to a maximum of 80 DVMRP active interfaces and 1900 passive interfaces. This assumes that no other protocols are running. If you need to run other routing protocols, you can enable IP forwarding and use routing policies and default route policies to perform IP routing. If you need to use a dynamic routing protocol, you need to have very few interfaces with OSPF or RIP enabled, while one or two will allow the connection of the switch to other switches to exchange dynamic routes.
- With PIM, you should have a maximum of 10 PIM active interfaces and all the rest passive when using 1980 interfaces. Also, in this case it is recommended to use IP routing policies with one or two IP unicast active interfaces.

Chapter 5

Enabling Layer 4-7 application services

This chapter describes the Nortel Networks Alteon Web Switching Module (WSM) and provides some general information you need to be aware of when utilizing the Passport 8600 and WSM. Specific topics included here are:

Topic	Page number
Introduction	next
Layer 4-7 switching	184
WSM architecture	187
Applications and services	192
Network architectures	202
Architectural details and limitations	206

Introduction

As each company strives to increase market share, deliver better service, and provide higher returns for shareholders, its network infrastructure assumes an increasingly significant role. Mission-critical applications mandate extreme levels of performance, availability and scalability and thus, make obvious the need for Layer 4-7 switching.

With the advent of the Internet and intranet, networks that connect server, employees, customers, and suppliers have become critical. Network downtime is unacceptable and poor performance of Web-based applications and online services can virtually shut down a business. The mass proliferation of servers, network devices and security solutions has created the requirement for enterprises and service providers to create high performance data center environments.

The complexity in scaling, managing and guaranteeing the availability of applications and services is one of the critical factors that makes Layer 4-7 a major requirement in today's networks. Many applications require multiple servers because one server does not provide enough power or capacity, and a single server cannot ensure the level of reliability and availability for business critical communication.

Layer 4-7 switching

Layer 4-7 switching means that switching is based on higher level protocol header information in the packet. By facilitating deep-packet inspection on TCP and UDP headers, Layer 4-7 switching allows intelligent routing for common applications including Hypertext Transfer Protocol (HTTP), File Transfer Protocol (FTP), domain name server (DNS), secure socket layer (SSL), Real-Time Streaming Protocol (RTSP), and Lightweight Directory Access Protocol (LDAP).

Layer 4-7 switching deals with the intelligent distribution of network traffic and requests across multiple servers or network devices. It permits applications and services to scale, while simultaneously eliminating single points of failure on the network. Layer 4-7 switching brings availability, scalability and fault tolerance to high performance networks. In addition, this type of intelligent traffic management allows you to segregate content across multiple servers and devices, accelerate it, and then prioritize it for delivery across available network resources.

Layer 4-7 switching enables at least four major applications for high performance networks, including:

- Server load balancing
- Global server load balancing
- Firewall and VPN load balancing
- Transparent cache redirection

For additional information, see [“Applications and services” on page 192](#).

Layer 4-7 switching in the Passport 8600 environment

The WSM speeds application performance and facilitates the availability and scalability of critical network services by migrating high-level networking functions from software to hardware. By using the WSM in a Passport 8600, you can perform wire-speed, deep-packet inspection, TCP session analysis, and Intelligent Traffic management.

The WSM provides all the necessary Layer 4-7 services including:

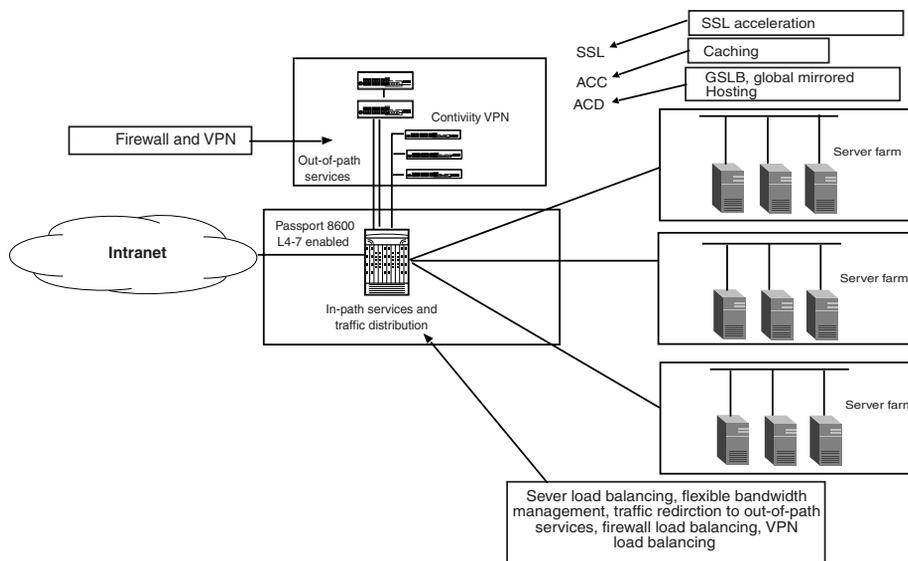
- local/global server load balancing
- web cache redirection
- firewall load balancing,
- VPN load balancing,
- streaming media load balancing
- Intrusion Detection System (IDS) load balancing
- bandwidth management
- DoS attack protection
- session persistence
- direct server return
- network failure recovery

For additional information, see [“Applications and services”](#) on page 192.

WSM location

The WSM resides inside the Passport 8600 as an intelligent module and transforms the 8600 into a complete Layer 2-7 intelligent routing solution. Enterprises, service providers, hosters, content providers and E-businesses can now obtain Alteon WebOS traffic management services in a cost- effective, easily-customizable I/O module.

At the same time, they can aggregate large numbers of 10/100/1000 Ethernet connections to servers, routers, firewalls, caches, and other essential networking devices. The WSM meets the demands of high performance networks by handling entire network sessions and real-time device and load conditions to direct requests and sessions to the most appropriate networking resource ([Figure 59](#)).

Figure 59 WSM's role as an intelligent module

WSM components

The Alteon WSM has four front-facing ports that you can configure to support either dual-media 10/100 or 1000BASE-SX connections to network devices, such as a upstream routers or pools of Intrusion Detection Servers. The remaining ports (four gigabit connections) are rear-facing and connect to the backplane of the Passport switch chassis. In this way, the WSM can enable all of the Layer 2 and Layer 3 fan-out modules and ports with Alteon WebOS traffic management and intelligent Internet services.

Figure 60 WSM ports

Up to eight Alteon WSM modules are supported in the Passport 8010 chassis. You can connect and configure all the Alteon WSM and Web OS applications and services via the Passport CLI, JDM, and Optivity Network Management Systems.

WSM architecture

The WSM was designed to take advantage of the density and robustness of the Passport 8600's Layer 2-3 capabilities. It provides high performance intelligent routing based on Layer 4-7 information to all ports.

The WSM also allows you to:

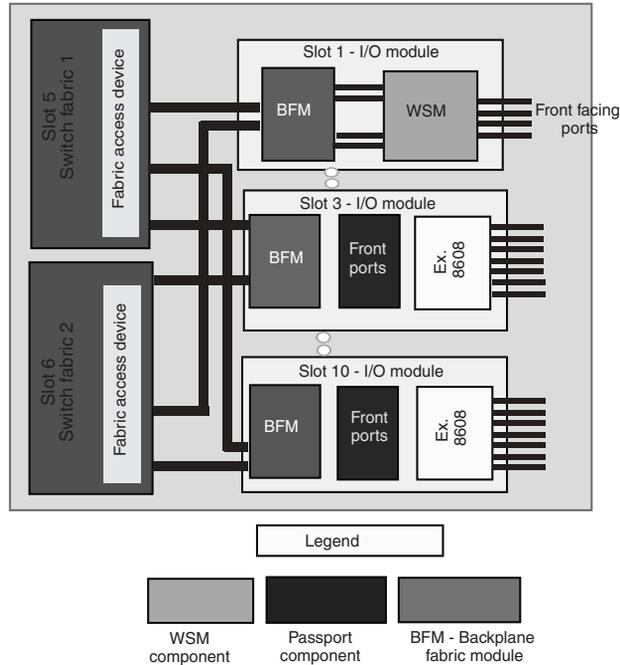
- Represent groups of real servers or network devices with a single instance (Virtual IP),
- Balance the traffic to this cluster of network devices (server load balancing)
- Limit traffic to individual devices or servers (persistent connections) and clusters via specific Layer 4-7 policies

Client and server connections through the WSM can use either Layer 2 or Layer 3 communication with the Passport 8600. Clients connect to the client-side VLAN and servers connect to a unique server-side VLAN. This ensures that there is no looped traffic.

Servers and client can exist on different subnets. Along with the unique two VLAN approach to processing client and server traffic, the overall configuration process has been simplified. Via the WSM default configuration, elements have also been automated to enable easy integration into the Passport 8600 environment.

The simplified data path architecture ([Figure 61](#)) shows that traffic from a Passport 8600 I/O module traverses the Passport switch fabric to the backplane fabric module (BFM) of the WSM. It then connects to the WSM using two dynamically created MLTs, tagged as 802.1q. Each MLT consists of two Gigabit links. These MLTs are set up automatically by the Passport 8600 when the WSM is initialized.

If you are connecting servers and clients to the Passport 8600 I/O module, it is recommended that you create two separate VLANs, one for the clients and one for the servers. Then, assign one dynamically-created MLT to each VLAN.

Figure 61 WSM data path architecture

The WSM has 4 front-facing ports (1, 2, 3, and 4). You can configure each of these at 10/100Mbps via an RJ-45 port or 1000 Mbps via an SX port, but not both. It also has 4 rear-facing, Gigabit ports which are used for connectivity to the Passport 8600 through the backplane. The WSM has two pre-configured trunks, each of which contains 2 rear-facing ports.

Passport default parameters and settings

At WSM initialization, the Passport 8600 dynamically creates two MLTs to establish communication between the Passport backplane and WSM ports. The higher MLT ID (32) goes to STG 64/VLAN 4093 and STG 1/VLAN 1, while the other MLT is user-configurable and by default, is not assigned to any VLAN and spanning tree group. You can configure the two dynamically-created MLTs and assign them to any VLAN and spanning tree group.

Table 16 provides more detail on the Passport default parameters and their settings.

Table 16 Passport default parameters and settings

Parameter	Setting
Passport MLT 31 (server MLT)	Upon initialization, BFM ports 1 and 2 are combined to create MLT group 31 when factory defaults are used.
Passport MLT 32 (client MLT)	Upon initialization, BFM ports 3 and 4 are combined to create MLT group 32 when factory defaults are used.
Passport VLAN 1 (client processing)	When using the factory settings on the Passport 8600, BFM ports 3 and 4 are added to VLAN1 by default.
Passport VLAN 4093	This is reserved for in-band management of the WSM and is automatically created when the WSM is initialized with the Passport 8600 chassis.
Passport STG 1	STG 1 is the default spanning tree for the Passport 8600. When an 8600 is started with the factory default configuration, all ports are automatically added to VLAN 1. VLAN 1 is assigned to spanning tree group 1. When you insert a WSM in the Passport 8600 chassis, BFM ports 3 and 4 are added to VLAN 1.
Passport STG 64	Refer to the “ VLAN 4093 and STG 64 ” section that follows and <i>Installing the Web Switch Module for the 8000 Series Switch</i> for more information.

VLAN 4093 and STG 64

VLAN 4093 is configured as follows during WSM initialization:

- Rear-facing ports 7 and 8
- IP address assigned in the range 172.31.255.246/28 - 172.31.255.253/28
- IP interface 256
- WSM trunk group 3

VLAN 4093 is configured as follows during Passport 8600 initialization:

- BFM ports 3 and 4
- IP address 172.31.255.245/28
- Mask 255.255.255.240
- Default IP management subnet

The Passport 8600 uses STG 64 for internal operation, while inserting the WSM. BFM ports 3 and 4 are added to STG 64.

For a more detailed description of STG 64 and VLAN 4093, see *Installing the Web Switch Module for the 8000 Series Switch*.

WSM default parameters

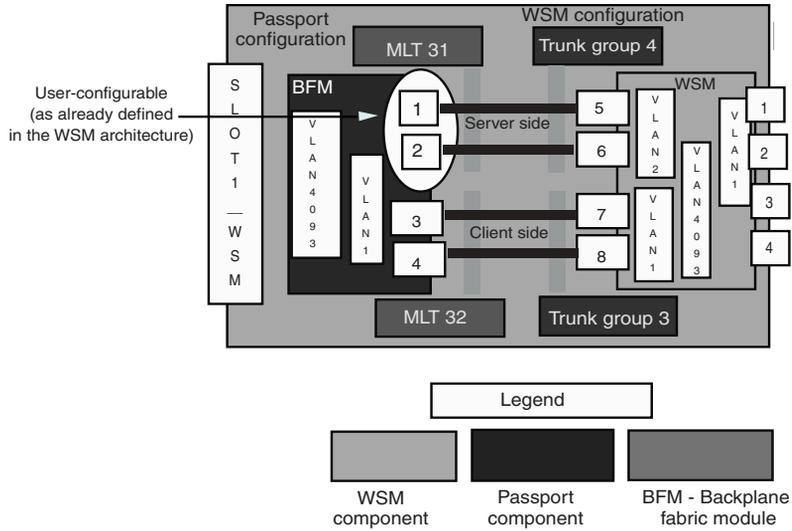
[Table 17](#) provides more detail on the WSM default parameters.

Table 17 WSM default parameters

Parameter	Setting
WSM VLAN 1 (client processing)	On the WSM, rear-facing ports 7-8, as well as front-facing ports 1-4 are added to VLAN 1. For more information, see <i>Installing the Web Switch Module for the 8000 Series Switch</i> .
WSM VLAN 2 (server processing)	On the WSM, rear-facing ports 5-6 are added to VLAN 2. For more information, see <i>Installing the Web Switch Module for the 8000 Series Switch</i> .
Trunk group 3 (client MLT)	WSM rear-facing ports 7 and 8 are combined to create trunk group 3.
Trunk group 4 (server MLT)	WSM rear-facing ports 5 and 6 are combined to make trunk group 4.
WSM STP1 (i.e., STG groups are referred to as STPs on the WSM)	VLAN 1 is added to STP 1.

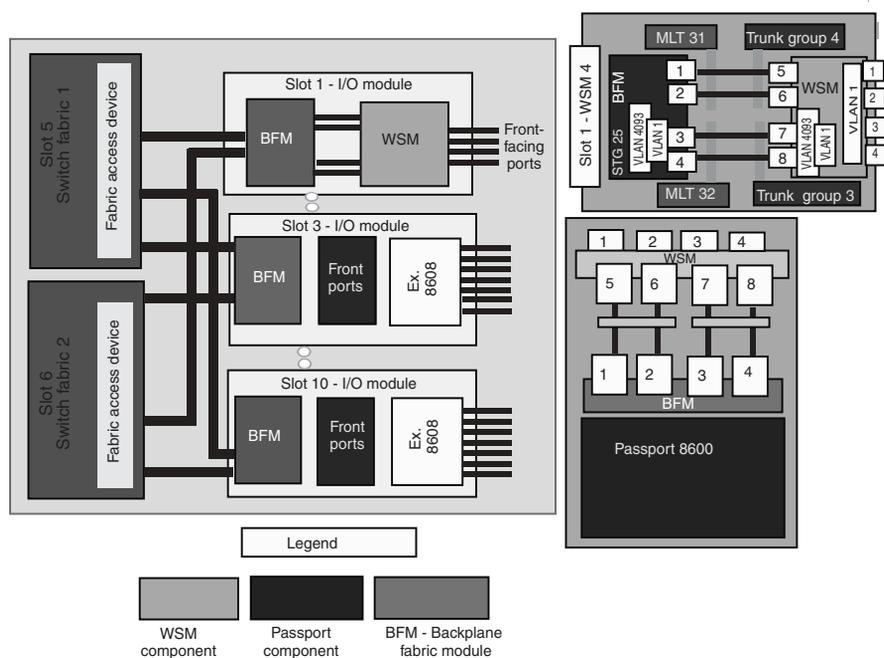
[Figure 62](#) shows the detailed WSM data path architecture.

Figure 62 Detailed WSM data path architecture



By making each WSM trunk a member of a different VLAN and by running STP (Spanning Tree Protocol), this architecture ensures connectivity with the WSM without introducing bridging loops.

[Figure 63](#) shows a single WSM default architecture.

Figure 63 Single WSM default architecture

Applications and services

This section summarizes some leading Layer 4-7 applications and services. The object here is to help you understand how the WSM can help by improving the performance, scalability and availability of critical applications and devices in your network. For additional information, refer to the *Alteon Web OS Switch Software 10.0 Application Guide*.

Local server load balancing

The proliferation of servers and network devices to perform critical tasks for enterprises, E-commerce businesses and service providers has led to numerous scalability and manageability challenges. Load balancing offers you a cost-effective way to resolve such issues.

Server load balancing (SLB) allows you to configure the WSM to balance user-session traffic among a pool of available servers or devices that provide shared services. SLB benefits your network by providing:

- Increased efficiency for server utilization and network bandwidth

With SLB, your Passport 8600 is aware of the shared services provided by your server pool and can then balance user session traffic among the available and appropriate resource. Important session traffic gets through more easily, thus reducing user competition for connections on overutilized devices. For greater control, traffic is distributed according to a variety of user-selectable rules.

- Increased reliability and availability of services to users

If any device in a server pool fails, the remaining servers continue to provide access to vital applications and data. You can bring the failed device back without interrupting access to services.

- Increased scalability of services

As users are added and server capabilities become saturated, you can add new servers seamlessly to the existing network

The WSM acts as the front-end to servers and network devices, interpreting user sessions requests and distributing them among the available and appropriate resources. Load balancing via the WSM is performed in the following ways:

- Virtual server-based load balancing

This is the traditional load balancing method. You configure the WSM to act as a virtual server and it is given a virtual server IP address (or range of addresses) for each collection of services it distributes. You can have as many as 255 virtual servers on the Passport 8600, each distributing up to eight different services (up to a total of 2048 services).

Each virtual server is assigned a list of IP addresses of the real servers in the pool where its services reside. When you request a connection to a service, you communicate with a virtual server on the WSM.

When the WSM receives your request, it binds the session to the IP address of the best available resource and remaps the fields in each frame from virtual addresses to real addresses. IP, FTP, RTSP, and static session WAP are examples of some of the services that use virtual servers for load balancing

- Filtered-based load balancing

A filter allows you to control the types of traffic permitted through the WSM. You configure filters to allow, deny, or redirect traffic according to IP address, protocol, or Layer 4 port criteria. In filtered-based load balancing, you use a filter to redirect traffic to a real server group.

If you configure the group with more than one real server entry, redirected traffic is load balanced among the available real servers in the group. Firewall load balancing, WAP with RADIUS snooping, IDS and WAN links use redirection filters to load balance traffic

- Content-based load balancing

Content-based load balancing uses Layer 7 application data such as URLs, cookies, and host headers to make intelligent load balancing and routing decisions. URL-based load balancing, browser-smart load balancing, and cookie based preferential load balancing are a few examples of content load balancing

Another key element of SLB is determining the health and availability of each real server or device. By default, the WSM checks each service on each real server every two seconds. If a service does not respond to four consecutive health checks, the WSM declares the service unavailable.

Health checking metrics

Metrics select the most appropriate real server to receive and service the client connection. [Figure 64](#) illustrates this process graphically.

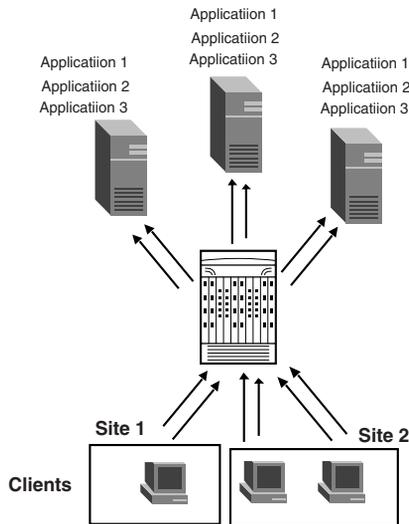
Figure 64 Metric selection process

Table 18 provides information on several of the available metrics. For more detailed information on these and the other available metrics, see the *Alteon Web OS Switch Software 10.0 Application Guide*.

Table 18 Health checking metrics

Metric	Description
Minmisses	Optimized for application redirection. This metric uses the IP address information in the client request to select a server. Based on its calculated score, the server that is most available is assigned the connection. This metric attempts to minimize the disruption of persistency when servers are removed from service. Only use this metric when persistence is a must.
Hash	Uses the destination IP address for application redirection, the source IP for SLB, and both for firewall load balancing. It ensures that requests are sent to the same server to: <ul style="list-style-type: none"> maximize successful cache hit ensure that client information is retained between sessions ensure that unidirectional flows of a given session are redirected to the same firewall
Least connections	Uses the number of connections currently open on each real server in real time to determine which one receives the request. The server with the fewest connections is considered the best choice.

Table 18 Health checking metrics (continued)

Metric	Description
Round robin	Issues new connections to each server in turn. When all the real servers in a group have received at least one connection, the issuing process starts over.
Response time	Uses real server response time to assign sessions to servers. The WSM monitors and records the amount of time it takes for each server to reply to the health check and adjusts the real server weights. In such a scenario, a server with half the response time as another server will receive a weight twice as high and receive more requests.
Bandwidth	Uses the octet counts to assign sessions. The servers that process more octets are considered to have less available bandwidth. The higher the bandwidth used, the smaller the weight assigned to the server. The next request then goes to the real server with the highest amount of free bandwidth. This bandwidth metric requires identical servers with identical connections.

GSLB

You enable global server load balancing (GSLB) via a license on the WSM. GSLB allows you to overcome many scalability, availability and performance issues that are inherent in distributing content across multiple geographic locations. By serving content from several different points, GSLB helps alleviate the impact.

GSLB allows you to balance server traffic load across multiple physical sites. Specifically, the WSM's GSLB implementation takes into account an individual site's health, response time, and geographic location. It then integrates the resources of the dispersed server sites for complete global performance.

GSLB also enables enterprises to meet the demand for higher content availability by distributing content and decision making. In this way, it ensures that the best-performing site receives the majority of traffic, thus enabling network administrators to build and control content by user, location, target application and more.

On the WSM, GSLB is based on the domain name server (DNS) and proximity by source IP address. Each WSM is capable of responding to clients' resolution requests with a list of addresses of distributed sites, prioritized by performance, geography and other criteria.

Application redirection

Application redirection improves network bandwidth utilization and provides unique network solutions. You can create filters to redirect traffic to cache and application servers improving speed of access to repeated client access to common Web or application content, which in turn frees up valuable network bandwidth.

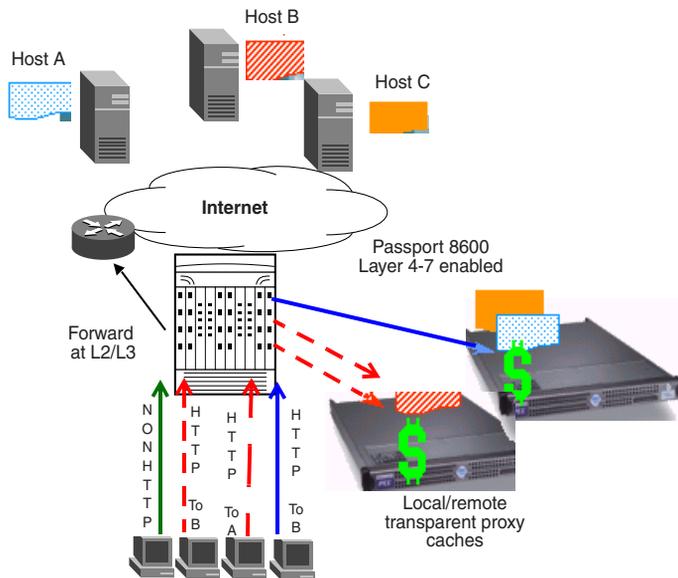
Application redirection helps to reduce traffic congestion by intercepting outbound client requests and redirecting them to a group of application or cache servers on a local networks. If the WSM recognizes the request as one that can be handled by a local network device, it routes it locally instead of sending the request across the Internet.

In addition to increasing the efficiency of a network, the WSM with application redirection allows clients to access information much faster and lowers WAN access costs.

The WSM also supports content intelligent application redirection, which allows a network administrator to redirect requests based on different HTTP header information. [Table 19](#) lists the available types of application redirection.

Table 19 Application redirection types

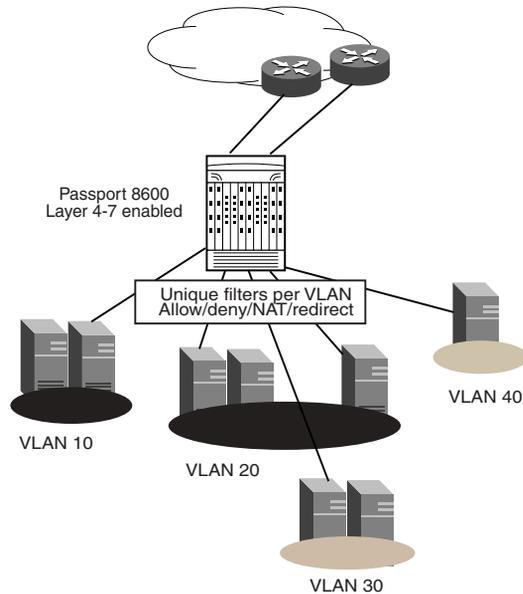
Application redirection type	Description
URL-based	Separates static and dynamic content requests and provides you with the ability to send requests for specific URLs or URL strings to designated cache devices. The WSM off loads the overhead processing from the cache server and only sends appropriate requests to the cache server farm.
HTTP header-based	Allows you to define host names and string IDs that will be redirected to cache server farms. For example if you want all domain names that end with .net or .uk not to go to a cache server, you can do so in a by creating a simple configuration.
Browser-based	Allows you to configure the user-agent to determine if client request will be redirected to a cache or server farm. Thus, you can send different browser types to the appropriate sites locally and on the internet (Figure 65).

Figure 65 Browser-based application redirection

VLAN filtering

On the WSM, you can apply filters per switch, per port or per VLAN. The advantage here is that VLAN-based filtering allows a single WSM to provide differentiated services for multiple groups, customers, users, or departments.

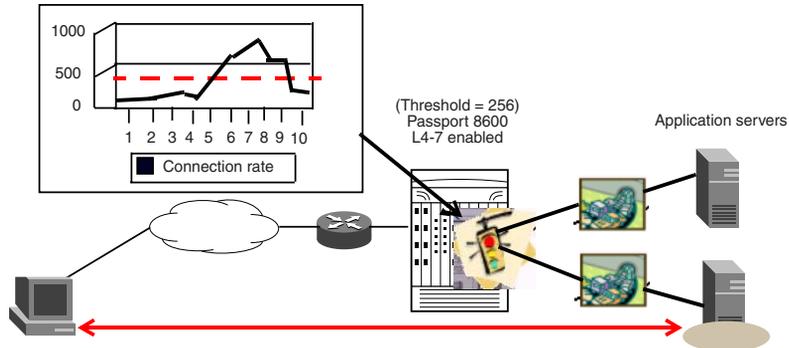
For example, you can define separate filters for the Finance department and Marketing department on the same WSM on two different VLANs. [Figure 66](#) shows how you can assign different filters to unique VLANs that allow, deny or redirect client requests, thus enabling differentiated service per group.

Figure 66 VLAN filtering

Application abuse protection

The WSM allows you to prevent a client or group of clients from claiming all the TCP or application resources on the servers. Thus, you automatically protect your applications from unnecessary abuse or usage via the WSM. You do so by monitoring the rate of incoming requests for connections to a virtual IP address and limiting the client request with a known set of IP addresses.

You ensure application abuse protection by defining the maximum number of TCP connection requests that will be allowed within a configured time window. The WSM then monitors the number of new TCP connections and when it exceeds the configured limit, any new TCP connections are blocked or *held down*. Specifically, the client is held down for a specified period of time after which new connections are permitted. [Figure 67](#) shows the application abuse protection process graphically.

Figure 67 Application abuse protection

Layer 7 deny filters

The WSM can secure your network from virus attacks by allowing you to configure the WSM with a list of potential offending string patterns (HTTP URL request). The WSM then examines the HTTP content of the incoming client request for the matching pattern.

If the matching virus pattern is found, the packet is dropped and a reset frame is sent to the offending client. SYSLOG messages and an SNMP trap are generated to warn you of a possible attack, while back-end devices and servers are automatically protected because the request is denied at the WSM ingress port.

Network problems addressed by the WSM

Table 20 describes a number of common network problems and explains how the WSM helps address them.

Table 20 Network problems addressed by the WSM

Problem	Description	Resolution
Network requests are inefficiently directed	Lower performing servers receive excessive requests while other are underutilized	WSM load balancing algorithms direct traffic and requests to the server or network device that is in the best position to handle it. The benefit here is increased efficiency and better utilization of network resources.
Network device failure leading to costly downtime	Server or network device is unavailable due to a hardware or OS failure	The WSM routes traffic to healthy and available resources only. The benefit here is that by proactively monitoring network element health and status, the WSM keeps your network downtime at a minimum, and network failures transparent. Once a failed element responds properly to health checks, it is automatically added to the online operations, thus easing network administration.
Critical application failure	Individual applications can hang or stop responding even though other applications on the same server are healthy	The WSM monitors individual application health and when necessary, redirects requests to other servers where the service is running properly. The advantage here is that failures are transparent, and critical applications remain available and active.
Traffic exceeds network limits	As traffic increases, servers are unable to respond to requests promptly	The WSM enables you to set thresholds for acceptable performance parameters, and it automatically redirects requests if a server is not responding. You can also set the maximum number connections per server to eliminate server overloading. The advantage to this is you always experience the level of service you anticipate and receive the content you are looking for. Furthermore, you can easily scale your solution by adding more servers to logical application groups.

Network architectures

This section describes various network architectures available for you to use when configuring the Passport 8600 and WSM for L2- L7 processing.

These architectures are not exhaustive. However, they do reflect the most common configurations. In most cases, you can mix and match the methods described here to accommodate specific requirements. The purpose of this section is to provide you with a framework of the various methods available to build upon.

Be aware that the following architectures are based on an SLB example and for simplicity sake, use VLAN 1 (client processing) and VLAN 2 (server processing).



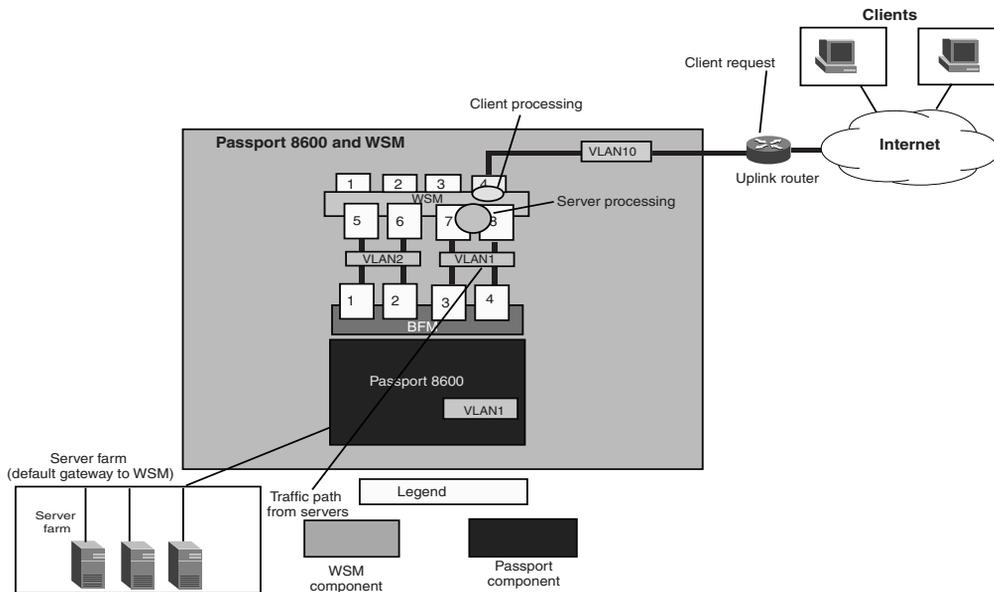
Note: You have the flexibility here to define appropriate VLANs if VLAN 1 and VLAN 2 are not available. However, you must ensure they are configured on both the Passport 8600 and WSM.

Using the Passport 8600 as a Layer 2 switch

Most architectures use the Passport 8600 as a Layer 3 switch to route traffic from the client and server to the WSM. Occasionally, you may need to implement Layer 4-7 services and applications using the Passport 8600 as a Layer 2 switch, however. Such occasions arise if you are aggregating optical Ethernet connections via the Passport 8600 I/O modules.

The sample architecture in [Figure 68](#) shows traffic entering through a Passport 8600 I/O module and traversing the backplane at Layer 2 to the WSM. In this example, client requests are coming from the Internet using an uplink router connected to the WSM front-facing port server farm. In turn, this server farm is connected to the Passport 8600 I/O module.

VLAN 1 is created in the Passport 8600 and BFM ports 3 and 4 (WSM dynamic MLT) are assigned to VLAN 1. An IP address is then assigned to VLAN 1 in the WSM consisting of Ports 7 and 8. The servers can point to the IP interface in the WSM as their default gateway. The Passport 8600 is providing a Layer 2 switching path here for the servers connected to the I/O module.

Figure 68 The Passport 8600 as a Layer 2 switch

Leveraging Layer 3 routing in the Passport 8600

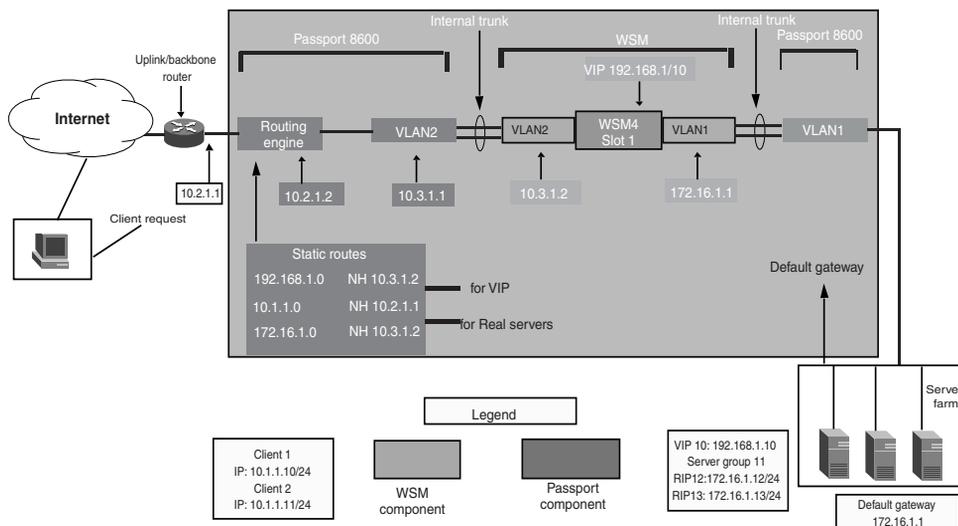
This architecture uses the Passport Layer 3 routing engine to direct traffic to the WSM. In this configuration, your client traffic is aggregated elsewhere and is routed or switched to the Passport 8600 and WSM. The routing engine of the Passport 8600 appropriately routes traffic to the WSM.

In [Figure 69](#), the client initiates a request to access a VIP that first traverses the uplink router connected to the Internet cloud. This request is forwarded to the Passport 8600 and enters the switch on one of the Passport 8600 I/O modules. The Passport 8600 routing engine makes a decision on the next-hop based on static-route entries. A static route is created in the Passport 8600, so that all traffic destined for the VIP is forwarded to the WSM.

In this example, the routing engine forwards the packet to a WSM interface in VLAN 2 where Layer 4-7 processing occurs. The WSM selects a real server and routes the request out the VLAN that houses the server. On egress, the traffic is sent out VLAN 1 across the backplane to the appropriate server connected to the Passport 8600 I/O module in VLAN 1.

In this type of design, you can utilize both the Layer 2 switching and Layer 3 routing engine in the Passport 8600, as well as the Layer 4-7 switching and server load balancing capabilities in the WSM.

Figure 69 Layer 3 routing in the Passport 8600

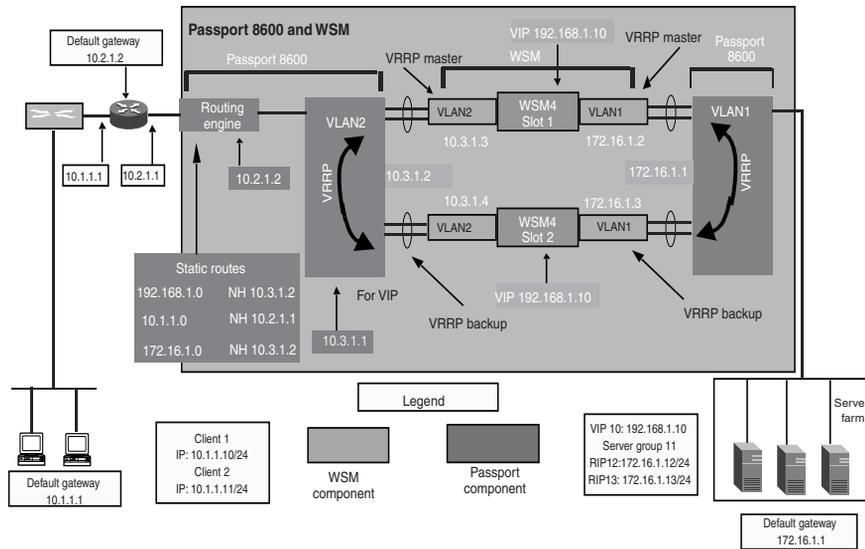


Implementing L4-7 services with a single Passport 8600

The following architecture provides you with a high availability scenario using a single Passport 8600 with multiple WSMs operating in active/standby redundancy mode. From a price standpoint, it is very common for architectures like this to use redundant modules and fabrics instead of an entire switch. There are also times when you may find a module failover preferable to an entire network path switch failover.

This architecture (Figure 70) is running two instances of VRRP (one for client access and one for server access) on the WSMs. The goal here is to offer high-availability. VRRP on the WSMs can communicate over the Passport backplane, which is the preferred method, since the Passport 8600 re-configures dynamic MLT connections to every WSM installed in the chassis.

You should ensure that VRRP communications occur over an available data path. Do this in the event that a WSM serving as the VRRP Master fails. If it does, the standby WSM can then re-fashion itself as the master.

Figure 70 Multiple WSMs using a single Passport 8600

Implementing L4-7 services with dual Passport 8600s

The following architecture utilizes a pair of Passport 8600s with multiple WSMs installed to offer a full-nodal redundancy, high-availability solution.

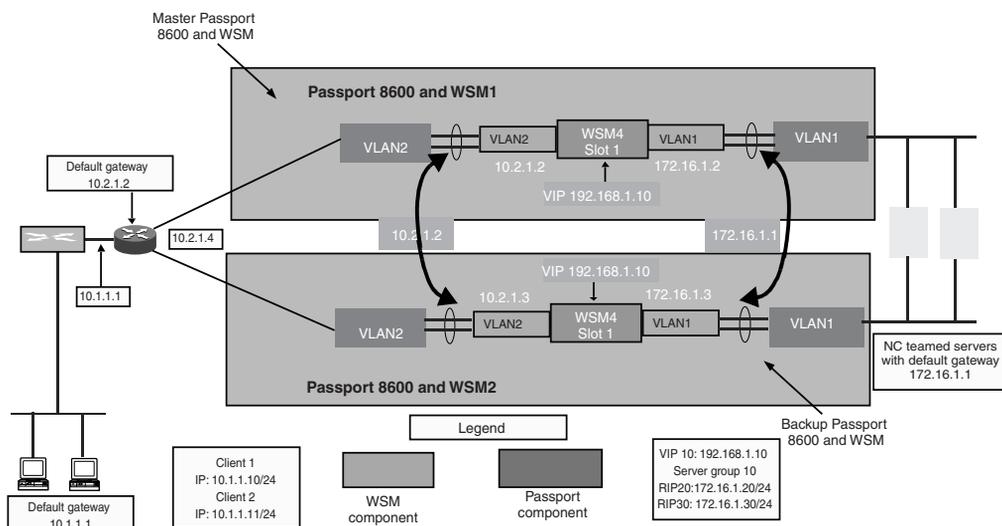
This architecture allows you to use both the clients and servers to create a single network route that provides hot-standby access to the Passport 8600 for L4-7 services. The client-side router and server-side each communicate with a VRRP instance that is running between the WSMs. These instances determine which Passport 8600 and WSM is the master (the one that accepts the traffic request) and which is the backup.

In [Figure 71](#), VRRP is implemented along the data path on the front-end and out of the data path on the back-end. This ensures that a failure on any component along the data path triggers a failover. This implementation avoids the situation when the inter-switch link on VLAN 1 fails causing a failover when it is not required.

The simplest method for you to configure servers in a high-availability mode is to employ NIC teaming. With NIC teaming, two NICs share the same IP address, permitting switchover to a live element should the interfacing switch, line, or NIC fail.

In this implementation, you configure a single IP address that corresponds to a single virtual MAC address. Since the IP address and MAC address never change, upstream and downstream network devices do not need to perform updates. Since VRRP is running on the WSM, a failure of the master WSM still allows traffic to traverse the VLAN 1 link as long as the top Passport 8600 is running.

Figure 71 Dual chassis high availability



Architectural details and limitations

WSM architectural details and limitations include the following:

- User and password management
- Passport unknown MAC discard
- Syslog
- Image management
- SNMP and MIB management
- Console and management support

Each of these topics is explained in the subsections that follow.

User and password management

The Passport 8600 password management and access levels determine the WSM access levels. All user access levels on the WSM are enabled by default. It is important to note here that login IDs and passwords are case-sensitive.

During the boot process, the Passport 8600 and WSM login passwords are synchronized. You can change passwords for the various access levels from the Passport 8600 config CLI password menu. These accounts/passwords are then mapped to the access levels on the WSM.

Note that password changes only impact the local Passport 8600. As a result, if you insert the same WSM into another Passport 8600 chassis, the password and access levels implemented on that Passport 8600 chassis apply after WSM boot process.

There are a total of 11 access levels on the Passport 8600, including 6 native levels of access: RWA, RW, L3, L2, L1, and RO. With Release 3.2.2, the Passport 8600 with WSM has added 5 more access levels for L4-7 configuration which are mapped to the corresponding access levels on the Passport 8600. These include L4Admin, SLBAdmin, Oper, L4Oper, and SLBOper.

You can change the login name and password pending user requirements. You cannot add additional access levels, nor can you delete them.

Table 21 shows the password mapping for the Passport 8600 login and WSM access levels.

Table 21 Passport 8600 and WSM user access levels

Login ID	Passport 8600 access	WSM access	Description and tasks performed
Rwa	rwa	admin	<p>Passport 8600- Read/write/all access. You have all the privileges of read-write access and the ability to change the security settings. Security settings include access passwords and the Web-based management user names and passwords.</p> <p>WSM- The SuperUser administrator has complete access to all menus, information, and configuration commands on the WSM, including the ability to change both the user and administrator passwords.</p>
rw	rw	admin	<p>Passport 8600- Read/write access. You can view and edit most device settings. You cannot change the security and password settings.</p> <p>WSM- Same as <i>admin</i> WSM access level.</p>
l3	l3	user	<p>Passport 8600- Layer 3 read/write access. You can view and edit device settings related to Layer 2 (bridging) and Layer 3 (routing) functionality. You cannot change the security and password settings.</p> <p>WSM- As a user, you have no direct responsibility for switch management. You can view all switch status information and statistics, but cannot make any configuration changes to the switch.</p>
l2	l2	user	<p>Passport 8600- Layer 2 read/write access. You can 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.</p> <p>WSM- Same as <i>user</i> WSM access level.</p>
l1	l1	user	<p>Passport 8600- Layer 1 read/write access. You can view most switch configuration and status information and change physical port parameters.</p> <p>WSM- Same as <i>user</i> WSM access level.</p>
ro	ro	user	<p>Passport 8600- Read-only access. You can view the device settings, but you cannot change any of the settings.</p> <p>WSM- Same as previous <i>user</i> WSM access level</p>

Table 21 Passport 8600 and WSM user access levels (continued)

Login ID	Passport 8600 access	WSM access	Description and tasks performed
l4admin	ro	l4admin	Passport 8600- Read-only access. You can view the device settings, but you cannot change any of them. WSM- The Layer 4 administrator configures and manages traffic on the lines leading to the shared Internet services. In addition to SLB administrator functions, the Layer 4 administrator can configure all parameters on the Server Load Balancing menus, including filters and bandwidth management.
slbadmin	ro	slbadmin	Passport 8600- Same as <i>ro</i> Passport 8600 access level. WSM - The SLB administrator configures and manages Web servers and other Internet services and their loads. In addition to SLB operator functions, the SLB administrator can configure parameters on the Server Load Balancing menus, with the exception of not being able to configure filters or bandwidth management.
oper	ro	oper	Passport 8600- Same as <i>ro</i> Passport 8600 access level. WSM- The Operator manages all functions of the switch. In addition to SLB operator functions, the Operator can reset ports or the entire switch
l4oper	ro	l4oper	Passport 8600- Same as <i>ro</i> Passport 8600 access level. WSM- The Layer 4 Operator manages traffic on the lines leading to the shared Internet services. This user currently has the same access level as the SLB Operator.
slboper	ro	slboper	Passport 8600- Same as <i>ro</i> Passport 8600 access level. WSM- The SLB Operator manages Web servers and other Internet services and their loads. In addition to being able to view all switch information and statistics, the SLB Operator can enable/disable servers using the Server Load Balancing operation menu.

Passport unknown MAC discard

As a key security component, you can enable the unknown MAC discard feature on the Passport 8600. It discards and prevents any unknown MAC addresses from accessing specific ports.

If you enable unknown MAC discard on BFM ports 3 and/or 4, connectivity to the WSM is lost. This results in warning messages similar to the following:

```
[09/13/02 16:56:49] WARNING Task=tRcIpTask An intrusion MAC
address:00:60:cf:50:52:60 at port 2/3
[09/13/02 16:57:37] WARNING Task=tCppRxTask An intrusion MAC
address:00:50:8b:d3:4e:fd at port 2/4
```

This action prevents you from connecting to the WSM. To restore the connection, you must disable the feature on both BFM ports x/3 and x/4, or you must configure the switch to allow specific MAC addresses. By configuring the known MAC addresses of the WSM in the *add-allow-mac* attribute, you can manually enable WSMs with the unknown MAC discard feature. This prevents unwanted network devices (such as sniffers) from accessing the network.

Syslog

In order for the WSM to generate SYSLOG messages to the SYSLOG host, you must configure the SYSLOG facility on the WSM to match that of the Passport 8600. The facility range provided on both components goes from 0 to 7.

The Passport 8600 has 4 severity levels (Info, Warning, Error, and Fatal) that can be generated as SYSLOG messages. You can map each Passport 8600 severity level accordingly to the eight severity levels of the standard UNIX SYSLOG daemon.

The WSM has 5 severity levels (Notice, Warning, Error, Critical, and Alert) that are generated directly as SYSLOG messages to the SYSLOG host. You cannot modify the severity levels. You need only configure the local facility on the WSM to match the facility of the Passport 8600 in order to generate SYSLOG messages. For more detail on the SYSLOG messages generated on the WSM, refer to the *Alteon Web OS Switch Software 10.0 Command Reference*.

Image management

You manage both the boot and switch images from the Passport 8600 WSM command level by using the 8600's **copy** command. You are required to enter the TFTP server address and boot/switch file name as part of the download process.

Copying to the WSM from a TFTP server, or from the WSM to a TFTP server requires that you create a temp file in the /flash directory. If there is not enough space available in /flash, the copy operation will fail.

You can also select which switch image to boot after a WSM reset. You do so at the Passport 8600 WSM level using **setboot** [*<slotId>*] [*<image-choice>*], or in the WSM via **/boot/image**. The *active*, *backup*, or *factory* configuration you load after a WSM reset is still set in the WSM via **/boot/conf**).

You can still copy and paste a configuration file/script to the WSM. You must connect to the WSM from the Passport 8600 level, apply, and save to update the configuration.

The WebOS still provides you with revert function (forgets un-applied changes) and revert apply (reverts back to previously saved config, without rebooting the WSM) commands.

SNMP and MIB management

Since the WSM is a switch within a switch architecture, it still retains its own SNMP agent. The SNMP interface is via the Passport 8600 CPU proxy. You utilize a special SNMP community string to select a WSM agent. The SNMP agent on the WSM communicates to the management station via VLAN 4093 on the Passport 8600.

When you reset the WSM with the factory configuration, the read and write community strings for SNMP are set to *wsm_xx*, where *xx* indicates the slot in which you inserted the WSM at the time of the reset. Any changes to the read and write community strings require a *wsmreset*.

You are unable to set the read and write community to the default Passport 8600 read and write community strings (public and private respectively). Since the WSM requires a reboot to effect changes in the community string, the strings are then set back to the default (i.e. *wsm_xx*).

You can find the detailed SNMP MIBs and trap definitions of the WSM SNMP agent in the following Alteon WebSystems enterprise MIB documents:

- [Altroot.mi](#)- Alteon product registrations, which are returned by sysObjectID.

- Altswitch.mib- Alteon enterprise MIB definitions.
- Alttrap.mib- Alteon enterprise trap definitions.

The MIB definitions reside on the JDM for the Passport 8600 which allow for the SNMP functions (Get, Set, Traps) for the WSM. The WSM also supports standard MIBs including RFC 1213, 1573, 1643, 1493, and 1757.

Due to SNMP incompatibility between in the Passport 8600 and the WSM, you cannot configure SNMP V2 on the Passport 8600 (using the **config sys set snmp trap-recv xx.xx.xx.xx v2c public** command). An error message displays should you try to do so.

Console and management support

Console port access to the WSM is supported through the front-panel maintenance port. The maintenance port uses a DIN-8 interface, so a DIN-8 to DB9-Female cable is required to connect to a standard PC COM port (DCE). This cable ships with the WSM.

It is recommended you only use the maintenance port for serial download of a software image to the WSM, when you cannot log in to the WSM CLI via the Passport 8600 CLI, or when logging of the boot process and errors are required.

During the boot process while the WSM is initializing, you can login to the console using the admin password. This functionality is available in order to allow direct connectivity to the WSM for maintenance purposes. However, once the card has registered, you cannot log in using the local admin password. At that point, accessing the console requires a valid Passport 8600 password.



Note: Only JDM version 5.5.x and above supports the Passport 8600 and WSM. If you use any version prior to 5.5, you can adversely affect automatic configuration of the WSM.

WAN link load balancing

WAN link load balancing is only supported through the front-facing ports of the WSM. This is because WAN link load balancing requires a proxy IP address (PIP). You cannot apply a PIP to a trunk group, or MLT 31 or 32 of the BFM ports.

VRRP hot standby

Hot standby mode is not supported on MLT 4 or rear-facing ports 7 and 8 of the WSM because it causes the switch to lose connectivity. In order to alleviate the high cost of spanning tree convergence times, Alteon has enabled some extensions to VRRP.

The Alteon Web Switch allows you to define a port as *hotstan*. By enabling hot standby on a port, you allow the hot standby algorithm to control the forwarding state. Essentially, this algorithm puts the master VRRP switch in forwarding mode and blocks the backup switch.

If you configure hot standby mode on backplane ports 7 or 8, this causes the backup switch to lose connectivity. This is because the hot standby algorithm has disabled the backup switch management ports.

Chapter 6

Designing multicast networks

This chapter provides information on designing networks supporting IP multicast on the Passport 8600 switch. The following features are described here:

Topic	Page number
Multicast handling in the Passport 8600	next
Multicast and MLT	216
IP multicast scaling	221
General IP multicast rules and considerations	226
IGMP and routing protocol interactions	237
DVMRP general design rules	240
General design considerations with PIM-SM	248
Multicast and SMLT	266
Reliable multicast specifics	275
TV delivery and multimedia applications	277
IGAP	281
PIM-SSM and IGMPv3	284

Multicast handling in the Passport 8600

The Passport 8600 provides a unique architecture that handles IP multicast in an efficient and optimized manner where a packet is duplicated only when needed. At the ingress side, hardware IP multicast (IPMC) records are used to determine the destination ports of the packet. A packet that matches a hardware record is forwarded to the switch fabric based on a pointer that points to the information on

the destination modules in the chassis and the destination ports on these modules. The switch fabric uses this information to determine how many copies are required and sends one copy per board that has receivers attached to it. A board that does not have receivers will not get a copy of a multicast packet.

At the board level, a multicast packet that is received will be duplicated to the receiver ports at the forwarding engine level using an egress forwarding pointer to forward to destination ports.

All IP multicast records that have the same group and sources in the same subnet will share the same egress-forwarding pointer. With DVMRP, all IP multicast records that have the same destination group and ingress VLAN also share the same egress forwarding pointer for IP multicast bridged traffic. This provides higher scalability for the system. The total number of available records in a Passport 8600 is 32K. For the M-modules introduced in Release 3.3, it is 128K. Refer to [“DVMRP scalability” on page 221](#) and [“PIM-SM and PIM-SSM scalability” on page 222](#) for specific scaling numbers per protocol.

Multicast and MLT

Release 3.5 introduces a feature that allows distribution of IP multicast streams over links of an MLT. With releases prior to release 3.5 or with non-E or M modules with any release, a multicast stream uses the link where the IGMP query, PIM hello, or DVMRP probe was received. Hence, without the new feature, multiple streams are not distributed between the available MLT links. If the link used by multicast traffic becomes unavailable, the multicast streams switch to another active link in the MLT group.

If you need to use several links to share the load of several multicast streams between two switches, use one of the following methods:

- [“DVMRP or PIM route tuning to load share streams,” next](#)
- [“Multicast flow distribution over MLT” on page 219](#)

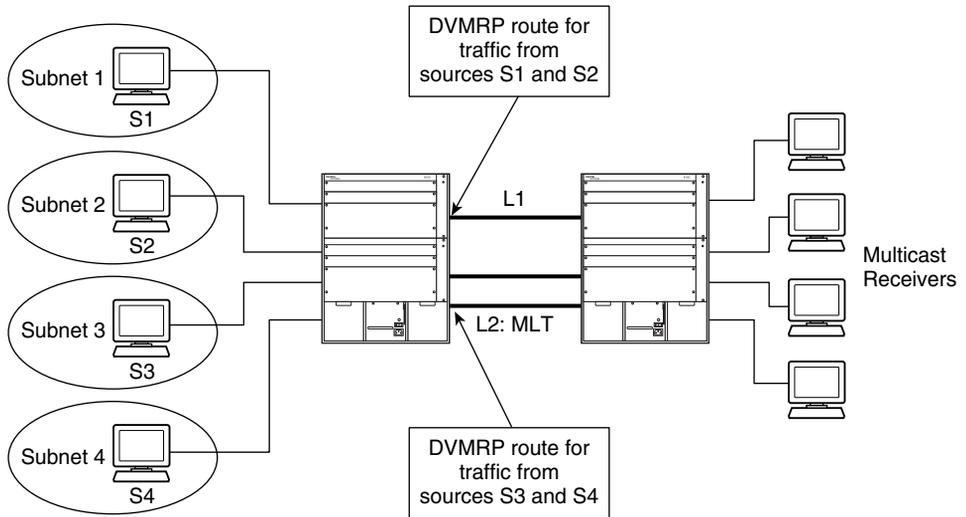
DVMRP or PIM route tuning to load share streams

You can use DVMRP or PIM routing to distribute multicast traffic. With this method, you must distribute sources of multicast traffic on different IP subnets and design routing metrics so that traffic from different sources flows on different paths to the destination groups.

[Figure 72](#) illustrates a way to distribute multicast traffic sourced on different subnets and forwarded on different paths. In [Figure 72](#), multicast sources S1 to S4 are on different subnets and you use different links for every set of sources to send their multicast data: S1 and S2 send their traffic on a common link (L1) and S3 and S4 on another common link (L2). These links can be MLT links, such as the L2 link. Unicast traffic is shared on these MLT links, while multicast uses only one of the MLT links. Receivers can be anywhere on the network. This design can be worked in parallel with unicast designs and does not impact unicast routing in the case of DVMRP.

Note that in this example, sources have to be on the same VLAN interconnecting the two switches together. 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. Use the CLI command `config ip dvmrp interface <IP Interface> metric <metric-value>` to change the metric value for an interface in order to provide different paths to different sources.

Figure 72 Traffic distribution for multicast data



9894EA



Note: When multicast is used in MLT configurations, Nortel Networks recommends using E- or M-modules if the MLT on the Passport 8600 is connected to a non-Passport 8600 device.

Multicast flow distribution over MLT

MultiLink Trunking (MLT) provides a mechanism for distributing multicast streams over an MLT. It does so based on source-subnet and group addresses and in the process provides you with the ability to choose the address and the bytes in the address for the distribution algorithm. As a result, you can now distribute the load on different ports of the MLT and aim (whenever possible) to achieve an even distribution of the streams. In applications like TV distribution, multicast traffic distribution is particularly important since the bandwidth requirements can be substantial when a large number of TV streams are employed.



Note: The multicast flow distribution over MLT feature is supported only on 8000 Series E- or M-modules. As a result, all the cards that have ports in an MLT must be 8000 Series E- or M-cards in order to enable multicast flow distribution over MLT.

Multicast flow distribution over MLT is based on source-subnet and group addresses. To determine the 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.

For example, consider:

Group address G[0].G[1].G[2].G[3], Group Mask
GM[0].GM[1].GM[2].GM[3], Source Subnet address S[0].S[1].S[2].S[3],
Source Mask SM[0].SM[1].SM[2].SM[3]

Then, the Port =:

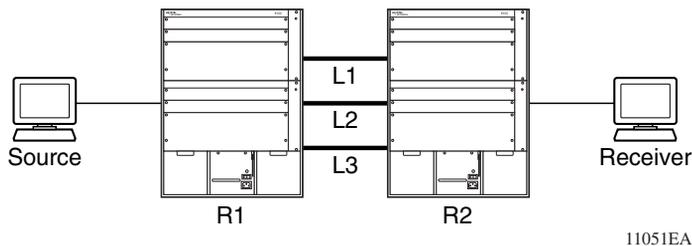
$$\begin{aligned} &(((((((G[0] \text{ AND } GM[0]) \text{ xor } (S[0] \text{ AND } SM[0])) \text{ xor } (G[1] \text{ AND } GM[0] \\ &) \text{ xor } (S[1] \text{ AND } SM[1]))) \text{ xor } (G[2] \text{ AND } GM[2]) \text{ xor } (S[2] \text{ AND } SM[2] \\ &))) \text{ xor } ((G[3] \text{ AND } GM[3]) \text{ xor } (S[3] \text{ AND } SM[3]))) \text{ MOD (active ports} \\ & \text{ of the MLT)} \end{aligned}$$

Stream failover consideration

The traffic interruption issue described below happens only in a PIM domain that has the “multicast MLT flow redistribution” feature enabled. (For information on this feature, see *Configuring IP Multicast Routing Protocols*.)

Figure 73 illustrates a normal scenario where streams are flowing from R1 to R2 through an MLT. The streams are distributed on links L1, L2 and L3.

Figure 73 Multicast flow distribution over MLT



If link L1 goes down, the affected streams get distributed on links L2 and L3. However, with redistribution enabled, the unaffected streams (which were flowing on L2 and L3) will also start distributing. Since the Passport 8600 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 Passport 8600 prunes these streams.

To avoid the above issue, make sure the `activity-chk-interval` command is set to its default setting of 210 seconds. If the activity check fails when the (S,G) entry timer expires (210 seconds), the Passport 8600 deletes the (S,G) entry and the corresponding hardware. The (S,G) entry and hardware will get recreated when packets corresponding to the (S,G) reach the switch again. The potential issue is that there might be a short window of traffic interruption during this deletion-creation period.

IP multicast scaling

IP multicast scaling depends on several factors. There are some limitations that are related to the system itself and other limitations that are related to how the network is designed.

The following sections provide the scaling number for DVMRP and PIM in a Passport 8600 network. These numbers are based on testing 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 here are recommended for general network designs.

DVMRP scalability

See the following sections for information on DVMRP scalability:

- [“Interface scaling,”](#) next
- [“Route scaling” on page 222](#)
- [“Stream scaling” on page 222](#)

Interface scaling

In the Passport 8000 Series software, there are no restrictions on what VLAN IDs can be configured with DVMRP. You can configure up to 500 VLANs for DVMRP. In earlier releases of the software, these numbers were more restrictive: the 3.0.x releases allow for 64 interfaces, while 3.1 allows for 200 interfaces. When configuring more than 300 DVMRP interfaces, you need to use the 8691SF that has 128MB of RAM.

Release 3.5 allows a maximum of 1980 DVMRP interfaces. Because of this, you should configure most interfaces as passive DVMRP interfaces (80 active interfaces maximum). This is particularly appropriate when the number of DVMRP interfaces approaches the limit. When this happens, it is recommended that you configure only a few interfaces as active DVMRP interfaces (the rest are passive). In general, when the number of interfaces is higher than 300, Nortel Networks recommends that you always use the 8691SF.

Route scaling

In the Passport 8600 switch, the number of DVMRP multicast routes can scale up to 2500 routes when deployed with other protocols such as OSPF, RIP and IPX/RIP. Note that with the proper use of DVMRP routing policies, your network will scale very high. For information on using the default route or announce and accept policies, refer to [“DVMRP policies” on page 242](#).

Stream scaling

In the Passport 8000 Series software, the recommended number of active multicast source/group pairs (S,G) is 2000. A source/group pair contains both a unicast IP source address and a destination multicast group address.

Nortel Networks recommends that the number of source subnets times the number of receiver groups not exceed 500. If more than 500 active streams are needed, you should group senders into the same subnets in order to achieve higher scalability. You should also give careful consideration to traffic distribution to ensure that the load is shared efficiently between interconnected switches (see [“Multicast and MLT” on page 216](#) for more information).



Note: The limits mentioned here are not hard limits, but a result of scalability testing with switches under load with other protocols running in the network. Depending upon your network design, these numbers may vary.

PIM-SM and PIM-SSM scalability

See the following sections for information on PIM-SM scalability:

- [“Interface scaling,”](#) next
- [“Route scaling” on page 223](#)
- [“Stream scaling” on page 224](#)
- [“Improving multicast scalability” on page 224](#)

Interface scaling

In the Passport 8000 Series software, you can configure up to 1980 VLANs for PIM. When configuring more than 300 PIM interfaces, you need to use the 8691SF that has 128MB of RAM. Note that interfaces running PIM have to run a unicast routing protocol which puts stringent requirements on the system. As a result, the 1980 interface number may not be supported in some scenarios, especially if the number of routes and neighbors is high. With a high number of interfaces, you should take special care to reduce the load on the system.

Use a very low number of IP routed active interfaces and better, use IP forwarding without any routing protocol enabled on the interfaces with only one or two with a routing protocol. You can perform proper routing by using the IP routing policies to announce and accept routes on the switch. Also, it is essential that you use the PIM passive interface introduced in the 3.5 release on the majority of the interfaces for proper operation. Nortel Networks recommends a maximum of 10 active PIM interfaces on a switch when the number of interfaces exceeds 300.



Note: Nortel Networks does not support more than 80 **active** interfaces and recommends the use of not more than 10 PIM active interfaces in a large scaled configuration with more than 500 VLANs. If you configure any more interfaces, they must be **passive**. For information on configuring PIM interfaces, see *Configuring IP Multicast Routing Protocols* in the Passport 8000 Series documentation set.

Route scaling

When using PIM-SM, the number of routes can scale up to the unicast route scaling since PIM uses the unicast routing table for its forwarding decisions. Thus, for higher route scaling, Nortel Networks recommends that you use OSPF.

As a general rule, a well designed network should not have too many routes in the routing table. For PIM to work properly, however, you should ensure that all subnets configured with PIM are “reachable” using the information in the unicast routing table. For the RPF check, PIM requires the knowledge of the unicast route to reach the source of any multicast traffic. For more detailed information, see [“PIM network with non-PIM interfaces” on page 265](#).

Stream scaling

In the Passport 8000 Series software, Nortel Networks recommends that with PIM-SM you limit the maximum number of active multicast S,G pairs to 2,000. A source, group pair contains both a unicast IP source address and a destination multicast group address.

You should also ensure that the number of source subnets times the number of receiver groups does not exceed 500.



Note: The limits mentioned here are not hard limits, but a result of scalability testing with switches under load with other protocols running in the network. Depending upon your network design, these number may vary.

Improving multicast scalability

To increase multicast scaling, follow these six network design rules:

- **Rule 1:** Whenever possible, use simple network designs that do not have VLANs spanning several switches. Instead, use routed links to connect switches.
- **Rule 2:** Whenever possible, group sources should send to the same group in the same subnet. The Passport 8600 uses a single egress forwarding pointer for all sources in the same subnet sending to the same group. Be aware that these streams will still have separate hardware forwarding records on the ingress side.

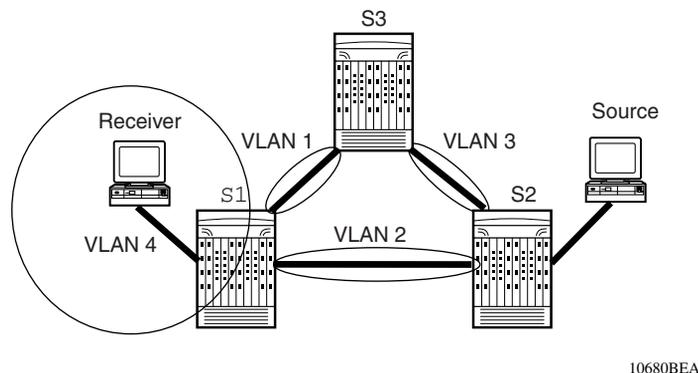
You can use the CLI command `show ip mroute-hw group trace` to obtain information about the ingress and egress port information for IP multicast streams flowing through your switch.

- **Rule 3:** Do not configure multicast routing on edge switch interfaces that will never contain multicast senders or receivers. By following this rule, you:
 - Provide secured control on multicast traffic entering or exiting the interface.
 - Reduce the load on the switch as well as the number of routes, as for example, in the case of DVMRP. This improves overall performance and scalability.

- **Rule 4:** Avoid initializing very high numbers (several hundreds) of multicast streams simultaneously. Initial stream setup is a process heavy task and initializing a large number may slow down the setup time of these streams and could in some cases result in some stream loss.
- **Rule 5:** Whenever possible, do not connect IP multicast sources and receivers on VLANs that interconnect switches. In some cases, such as the design shown in [Figure 74](#), this may result in more hardware records being consumed. By placing the source on the interconnected VLAN, traffic takes two paths to the destination depending on the RPF checks and the shortest path to the source.

For example, if a receiver is placed on VLAN 1 on switch S1 and another receiver is placed on VLAN 2 on this switch, traffic may be received from two different paths to the two receivers. This results in the use of two forwarding records. When the source on switch S2 is placed on a different VLAN than VLAN 3, traffic takes a single path to switch S1 where the receivers are located.

Figure 74 IP multicast sources and receivers on interconnected VLANs



10680BEA

- **Rule 6:** Use the default timer values for PIM and DVMRP. When timers are used for faster convergence, they usually adversely affect scalability since control messages are sent more frequently (e.g. DVMRP route updates). If faster network convergence is required, configure the timers with the same values on all switches in the network. Also, it is necessary for you to perform baseline testing in most cases to achieve optimal values for timers versus required convergence times and scalability. See [“DVMRP timers tuning” on page 241](#) for more detail.

General IP multicast rules and considerations

The following sections provides general rules and considerations to follow when using IP multicast on the Passport 8600 switch. It includes recommendations on proper network design for:

- [“IP multicast address ranges,”](#) next
- [“IP to Ethernet multicast MAC mapping”](#) on page 227
- [“Dynamic configuration changes”](#) on page 229
- [“DMVRP IGMPv2 back-down to IGMPv1”](#) on page 230
- [“TTL in IP multicast packets”](#) on page 230
- [“Multicast MAC filtering”](#) on page 232
- [“Multicast filtering and multicast access control”](#) on page 233
- [“Split-subnet and multicast”](#) on page 236

IP multicast address ranges

IP multicast utilizes 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 when dealing with the IP multicast range of addresses.

Addresses from 224.0.0.0 through 224.0.0.255 are reserved by IANA for link-local network applications. Packets with an address in this range are not forwarded by multicast capable routers by design. For example, OSPF uses both 224.0.0.5 and 224.0.0.6 and VRRP uses 224.0.0.18 to communicate across a local broadcast network segment.

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. RFC1700 contains a complete list of these reserved numbers.

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 applications, such as one-to-many applications. (See draft-holbrook-ssm-00.txt for more details). 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, meaning they are reserved for use in private domains and should not be advertised outside that domain. This multicast range is analogous to the 10.0.0.0/8, 172.16.0.0/20, and 192.168.0.0/16 private address ranges in the unicast IP space.

Technically, 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.

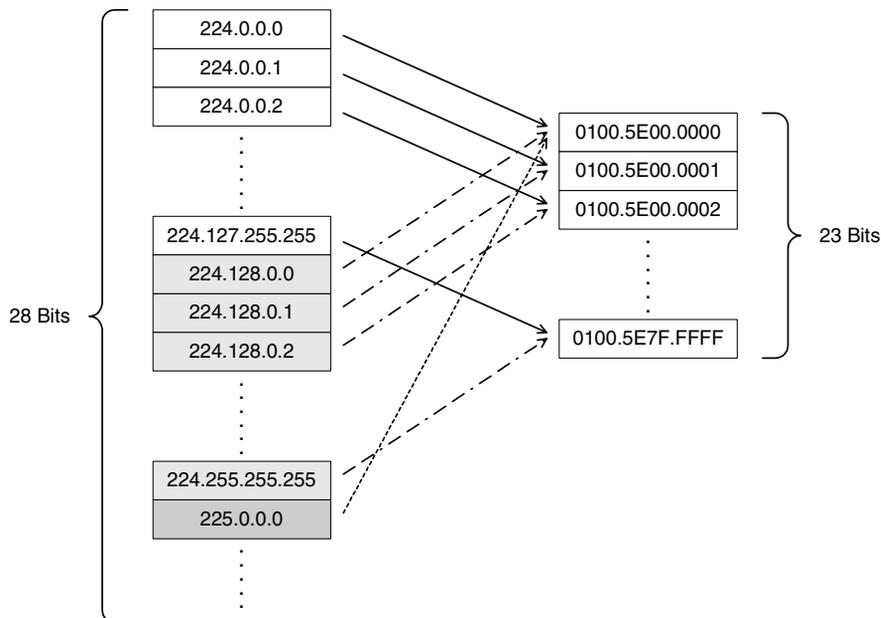
Note that while you are free to use any multicast address you choose on your own private network, even reserved addresses, it is generally not a good network design practice to allocate public addresses to private network entities. This is true with regard to both unicast host and multicast group addresses on private networks. To prevent private network addresses from escaping to a public network, you may wish to use the Passport 8600 announce and accept policies described on [page 242](#).

IP to Ethernet multicast MAC mapping

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, however, Ethernet has only 23 addressing bits assigned to IP multicast. Ethernet's multicast MAC address space is much larger than 23 bits, but only a sub-range of that larger space has been allocated to IP multicast by the IEEE. 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 (Figure 75). 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 75 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 designing the group addresses for multicast applications, you should take care to efficiently distribute streams only to hosts that are receivers. The Passport 8600 switches IP multicast data based on the IP multicast address and 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 or the other stream. This increases the load on both the hosts and the switch. To avoid this extra load, it is recommended that you manage the IP multicast group addresses used on the network.

At the same time, however, it is worth noting that the Passport 8600 does not forward IP multicast packets based on multicast MAC addresses- even when bridging VLANs at Layer 2. Thus, the Passport 8600 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 IP group address, regardless of whether the VLAN is bridged or routed. You should be aware then that while the Passport 8600 does not suffer from the problem described in the previous example, other switches in the network might. This is particularly true of pure L2 switches.

In a network that includes non-Passport 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 addressing needs of even the largest private enterprise.

Dynamic configuration changes

It is not recommended that you perform dynamic configuration changes in IP multicast when multicast streams are flowing in a network. This is particularly true when you change:

- the protocol running on an interface from PIM to DVMRP or vice versa
- the IP address and/or subnet mask for an interface

For such changes, Nortel Networks recommends that you stop all multicast traffic that is flowing in the network. If the changes are necessary and there is no control on the applications sending multicast data, it may be necessary for you to disable the multicast routing protocols before performing the change. For example, you should consider doing so before making interface address changes. Note that in all cases, these changes will result in traffic interruptions in the network since they impact neighborhood state machines and/or stream state machines.

DMVRP IGMPv2 back-down to IGMPv1

The DMVRP standard states that when a router operates in IGMPv2 mode and another router is discovered on the same subnet in IGMPv1 mode, you must take administrative action to back the router down to IGMPv1 mode. When the Passport 86000 switch detects an IGMPv1 only router, it automatically downgrades from IGMPv2 to IGMPv1 mode.

This feature saves network down time and configuration effort. However, it is not possible to dynamically switch back to IGMPv2 mode because multiple routers, including the Passport 8600 switch, now advertise their capabilities as limited to IGMPv1 only. To return to IGMPv2 mode, the Passport 8600 switch must lose its neighbor relationship. Subsequently when the switch reestablishes contact with its neighboring routers, the Passport 8600 switch operates in IGMPv2 mode.

You can view the IGMP configured mode and the operational mode either through the CLI or Device Manager.

TTL in IP multicast packets

The Passport 8600 switch treats multicast data packets with a Time To Live (TTL) of 1 as expired packets and sends them to the CPU before dropping them. You can avoid this situation by ensuring 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.

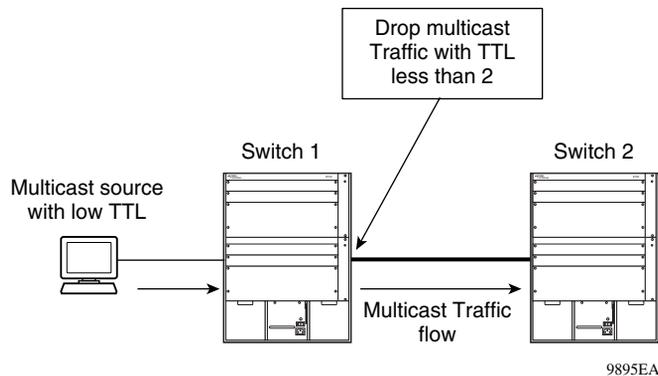


Note: Nortel Networks recommends using a TTL value of 33 or 34 to minimize the effect of looping in an unstable network.

To avoid sending packets with a TTL of 1 to the CPU, the Passport 8600 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 to the same group with sources on the same originating subnet as the stream with a TTL of 1.

To ensure that a switch does not receive multicast streams with a TTL=1, thus pruning other streams originating from the same subnet for the same group, you can configure the upstream Passport 8600 switch (Switch 1) to drop multicast traffic with a TTL < 2 (see Figure 76). In this configuration, all streams that egress the switch (Switch 1) with a TTL of 1 are dropped. In Device Manager, select IP Routing > Multicast > Interface to configure the TTL for every DVMRP interface.

Figure 76 Passport 8600 Switches and IP multicast traffic with low TTL



Changing the accepted egress TTL value does not take effect dynamically on active streams. To change the TTL, disable DVMRP, then enable it again on the interface with a TTL > 2. Use this workaround in a Passport 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 start sending its multicast data with a TTL lower than 2. Otherwise, all of its traffic is dropped and the load on the switch is increased. Note that enhanced modules (E- or M-modules), which provide egress mirroring, do not experience this behavior.

Multicast MAC filtering

Certain network applications, such as Microsoft Network Load Balancing Solution or NFS, require the ability for multiple hosts to share a multicast MAC address. Instead of flooding all ports in the VLAN with this multicast traffic, this feature allows you to forward traffic to a configured subset of the ports in the VLAN. Note that this multicast address is not an IP multicast MAC address, so you should not confuse this feature with IP multicast functionality.

At a minimum, you must map the multicast MAC address to a set of ports within the VLAN. In addition, if traffic is being routed on the local Passport 8600, you must configure an ARP entry to map the shared unicast IP address to the shared multicast MAC address. This is true since the hosts can also share a virtual IP address, and packets addressed to the virtual IP address need to reach them all.

It is recommended that you limit the number of such configured multicast MAC addresses to a maximum of 100. This number is inter-related with the maximum number of possible VLANs you can configure. For example, for every multicast MAC filter you configure, the maximum number of configurable VLANs on the Passport 8600 is reduced by one. Similarly, configuring large numbers of VLANs reduces the maximum number of configurable multicast MAC filters downwards from 100.

Release 3.5 introduced the possibility to configure under this feature the addresses starting with 01.00.5E that are reserved for IP multicast address mapping. When using a configuration with these addresses, you should be very careful of not having IP multicast enabled with streams that match the configured addresses. This will result in a malfunction of the IP multicast forwarding as well as in the Multicast MAC filtering function.

Multicast filtering and multicast access control

This section shows how multicast access policies are implemented in release 3.5 and in releases prior to 3.5.

New release 3.5 multicast access control policies

Release 3.5 introduces a complete set of new multicast access control policies that flexibly and efficiently protect a network from unwanted multicast access as well as multicast spoofing. These policies are:

- `deny-tx` — Prevents a matching source from sending multicast traffic to the matching group on the interface where the `deny-tx` access policy is configured.

The `deny-tx` access policy is the opposite of `allow-only-tx` and conflicts with `allow-only-both`. The `deny-tx` access policy cannot exist with these “allow” access policies for the same prefix-list on the same interface at the same time.

- `deny-rx` — Prevents a matching group from receiving IGMP reports from the matching receiver on the interface where the `deny-rx` access policy is configured.

The `deny-rx` access policy is the opposite of `allow-only-rx` and conflicts with `allow-only-both`. The `deny-rx` access policy cannot exist with these “allow” access policies for the same prefix-list on the same interface at the same time.

- `deny-both` — Prevents a matching IP address from both sending multicast traffic and receiving IGMP reports from a matching receiver on an interface where the `deny-both` access policy is configured.

The `deny-both` access policy is the opposite of `allow-only-both` and conflicts with the other “allow” access policies. The `deny-both` access policy cannot exist with any “allow” access policies for the same prefix-list on the same interface at the same time.

- `allow-only-tx` — Allows only the matching source to send multicast traffic to the matching group on the interface where the `allow-only-tx` access policy is configured. This access policy discards all other multicast data received on this interface.

The `allow-only-tx` access policy is the opposite of `deny-tx` and conflicts with `deny-both`. The `allow-only-tx` access policy cannot exist with these “deny” access policies for the same prefix-list on the same interface at the same time.

- `allow-only-rx` — Allows only the matching group to receive IGMP reports from the matching receiver on the interface where the `allow-only-rx` access policy is configured. This access policy discards all other multicast data received on this interface.

The `allow-only-rx` access policy is the opposite of `deny-rx` and conflicts with `deny-both`. The `allow-only-rx` access policy cannot exist with these “deny” access policies for the same prefix-list on the same interface at the same time.

- `allow-only-both` — Allows only the matching IP address to both send multicast traffic to and receive IGMP reports from the matching receiver on an interface where the `allow-only-both` access policy is configured. This access policy discards all other multicast data and IGMP reports received on this interface.

The `allow-only-both` access policy is the opposite of `deny-both` and conflicts with the other “deny” access policies. The `allow-only-both` access policy cannot exist with any “deny” access policies for the same prefix-list on the same interface at the same time.

Multicast access policies before release 3.5

In the Passport 8000 Series software, a common IGMP code is used for IGMP snooping, PIM, and DVMRP routing. You can deploy multicast access policies on IGMP snooping to control which hosts can send or receive data for a multicast session based on VLAN and multicast group address.

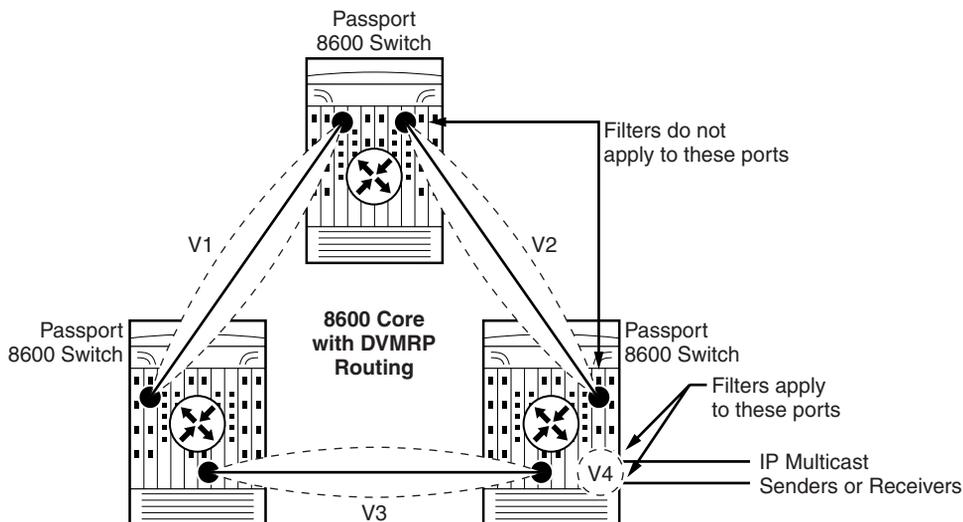
Guidelines for multicast access policies

Use the following guidelines for 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 to the hardware ASICs when the first packet of a multicast stream is received by the switch.

Multicast access policies can be applied on a DVMRP or PIM routed interface if IGMP reports control 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 77](#) 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 77 Applying IP Multicast access policies for DVMRP

10361EA

The following rules and limitations apply to IGMP access policies when used with IGMP versus DVMRP and PIM:

- Static member applies to snooping, DVMRP and PIM on both interconnected links and edge ports.
- Static Not Allowed to Join applies to snooping, DVMRP and PIM on both interconnected links and edge ports.
- For multicast access control, denyRx applies to snooping, DVMRP and PIM. DenyTx and DenyBoth apply only to snooping on the Passport 8600, but not on Passport 8100.

Split-subnet and multicast

The split subnet issue arises when a subnet is divided into two non-connected sections in a network. This results in erroneous routing information on how to reach the hosts on that subnet being produced. This problem applies to any type of traffic. However, it has a larger impact on a network with PIM-SM running.

When a network is running PIM and there is the potential of a split-subnet situation, you should ensure that the RP is not placed on a subnet that can become a split subnet. Also, you should avoid having receivers on this subnet. Since the RP is an entity that has to 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. This is true even for receivers and senders that are not part of the split-subnet.

IGMP and routing protocol interactions

The following cases provide you with design tips for those situations where Layer 2 multicast is used along with Layer 3 multicast. This is typically the case when a Layer 2 edge device is connected to one or several Layer 3 devices. The cases that follow involve IGMP interactions with PIM and DVMRP protocols.



Note: On a Passport 8600 switch, you must configure the IGMP Query Interval with a value higher than 5 to prevent the switch from dropping some multicast traffic.

IGMP and DVMRP

In [Figure 78](#), switches A and B are running DVMRP and switch C is running IGMP Snooping. Switch C connects to A and B through ports P1 and P2 respectively. Ports P1, P2, P3 and P4 are in the same VLAN. A source S is attached to switch A on a different VLAN than the one(s) connecting A to C and a receiver R is attached to switch B on another VLAN.

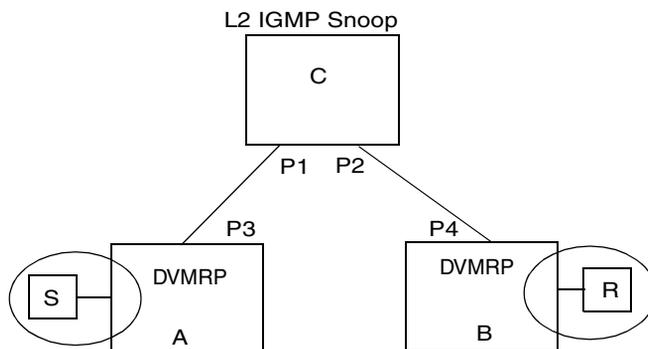
Assume that switch C has not been configured with any mrouter ports. If switch A is the querier, then it becomes the mrouter (multicast router port) port for C. The receiver does not receive data from source S, because C does not forward data on the link between C-B (non-mrouter).

You can surmount this problem in two ways:

- configure ports P1 and P2 as mrouter ports on the IGMP snoop VLAN
or
- 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. If you connect switches A and B together, there is no need for any specific configuration since the issue does not arise.

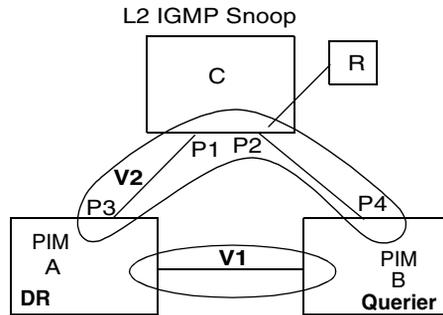
Figure 78 IGMP interaction with DVMRP



IGMP and PIM-SM

In [Figure 79](#), switches A and B are configured with PIM-SM, and switch C is running IGMP Snooping. A and B are interconnected with VLAN 1 and C connects to A and B with VLAN 2.

If a receiver R is placed in VLAN 2 on switch C, it does not receive data. This is because PIM chooses the higher IP address as DR, whereas IGMP chooses the lower IP address as 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.

Figure 79 IGMP interaction with PIM

As in the previous IGMP interaction with DVMRP, you can surmount this problem in two ways:

- Configure ports P1 and P2 as mrouter ports on the IGMP snoop VLAN
or
- 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. Note that this issue does not occur when DVMRP has the querier and forwarder as the same switch, as for example, when IGMPv2 is used.

IGMP and PIM-SSM

The Passport 8000 Series implementation of IGMPv3 for PIM-SSM is not backward compatible with IGMPv1 or IGMPv2. This may result in the switch discarding version 1 and version 2 membership reports.

DVMRP general design rules

The following sections describe DVMRP design rules:

- [“General network design,”](#) next
- [“Sender and receiver placement”](#) on page 241
- [“DVMRP timers tuning”](#) on page 241
- [“DVMRP policies”](#) on page 242
- [“DVMRP passive interface”](#) on page 247

General network design

As a general rule, you should design your network with routed VLANs which 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.



Note: When DVMRP VLANs span more than two switches, temporary multicast delayed record aging on the non-designated forwarder may occur after receivers go away.

DVMRP uses not only the metric, but also the IP addresses to choose the RPF path. Thus, you should take great care when assigning the IP addresses in order to ensure the utilization of the best path.

As with any other distance vector routing protocol, note that DVMRP suffers from count-to-infinity problems when there are loops in the network. This makes the settling time for the routing table higher, so it is something that you should be aware of when designing your network.

Sender and receiver placement

Another useful rule you should follow is to avoid connecting your senders and receivers to the subnets/VLANs which connect core switches. If you need to connect servers generating multicast traffic or acting as multicast receivers to the core, you should connect them to VLANs different from the ones which connect the switches. As shown in [Figure 77](#), 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.

DVMRP timers tuning

The Passport 8000 Series software allows you to 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).

You may need to change the default values of these timers for faster network convergence in the case of failures or route changes. If so, Nortel Networks 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 values when setting DVMRP timers since this can result in a high switch load trying to process frequent messages. This is particularly true for route update timers, especially in the case of a large number of routes in the network. Also, note that 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 dictates of the DVMRP standard in the relationship between correlated timers. For example, the Route Hold-down = 2 x Route Report Interval.

DVMRP policies

DVMRP policies include:

- [“Announce and accept policies,”](#) next
- [“Do not advertise self”](#) on page 245
- [“Default route policies”](#) on page 246

Announce and accept policies

Announce and accept policies for DVMRP allow you to control the propagation of routing information. Under the multicast routing paradigm, routing information governs which subnets can contain sources of multicast traffic, rather than destinations of multicast streams. In a secure environment, this can be an important issue since DVMRP periodically floods and prunes streams across the network, possibly leading to congestion.

You can successfully filter out subnets that only have multicast receivers by using accept or announce policies without impacting the ability to deliver streams to those same networks. You can also use policies to scale very large DVMRP networks by filtering out routes that are not necessary to advertise.

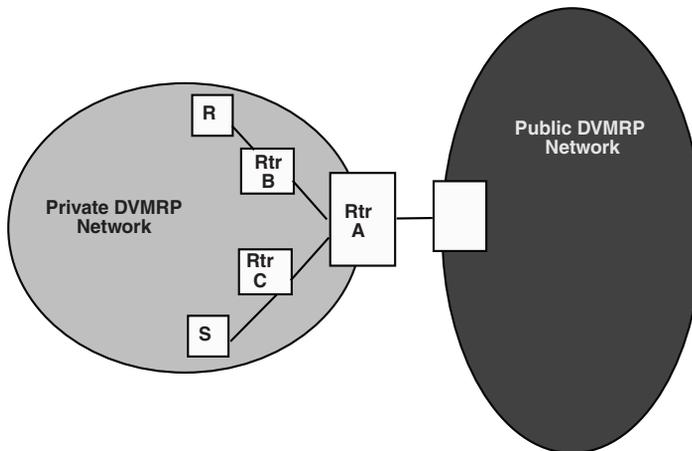
An announce policy affects the routes that are advertised to neighboring DVMRP routers. Thus, while received routes are poison-reversed and added to the local routing table, they are also potentially filtered by an announce policy when advertised. You can use this feature at key points in the network to limit the scope of certain multicast sources. An announce policy effectively allows the local router to receive the stream, while propagating it on a subset of outgoing interfaces. If there are no potential egress interfaces for a particular multicast source (i.e., the local router has no need for the stream), you may find it more appropriate to use an accept policy.

Announce policy on a border router

[Figure 80](#) 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 ultimate goal here 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, you may find that the most appropriate solution here is to use an announce policy on Router A's interface connecting to the public network. 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.

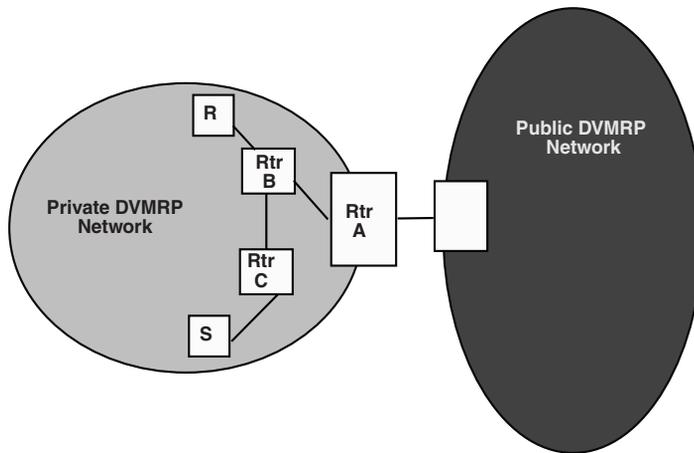
Figure 80 Announce policy on a border router



An accept policy blocks routes upon receipt. When a route is received that is to be filtered by an accept policy, the local router does not poison-reverse the route. Therefore, the remote router does not add the interface to its distribution tree. This effectively prevents any stream from the source from being forwarded over the interface. Like announce policies, you can use accept policies at key points in the network to limit the scope of certain multicast sources.

Accept policy on a border router

[Figure 81](#) illustrates a similar scenario (with the same requirements) as that in described in [Figure 80](#). This time, Router A has only one multicast capable interface connected to the private network. Since one interface precludes the possibility of intra-domain multicast transit traffic, there is no need for private multicast streams to be forwarded to Router A. Thus, you may find it inefficient to use an announce policy on the public interface since private streams are forwarded to Router A just to be dropped (and pruned). Under such circumstances, you will find it more appropriate to use an accept policy on Router A's private interface. Public multicast streams are forwarded into the private network as desired.

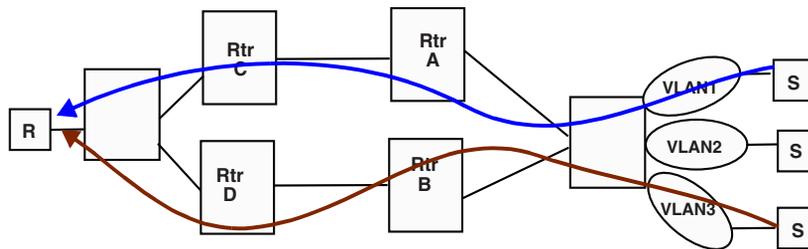
Figure 81 Accept policy on a border router

You may find accept policies useful when you cannot control routing updates on the neighboring router. For example, a service provider cannot directly control the routes advertised by its customer's neighboring router, so the provider may choose to configure an accept policy to only accept certain agreed upon routes.

You can utilize an accept policy in a special way to receive a default route over an interface. If a neighbor is supplying a default route, you may find it desirable to accept only that route while discarding all others, thus reducing the size of the routing table. In this situation, the default route is accepted and poison-reversed, while 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 source subnet. [Figure 82](#) 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. You may find it desirable then to load balance the traffic across the other available links.

In such cases, you can use announce policies on Router A to increase the advertised metric of certain routes to make the path between Routers B and D more preferable. Thus, traffic originating from those subnets takes the alternate route between B and D.

Figure 82 Load balancing with announce policies

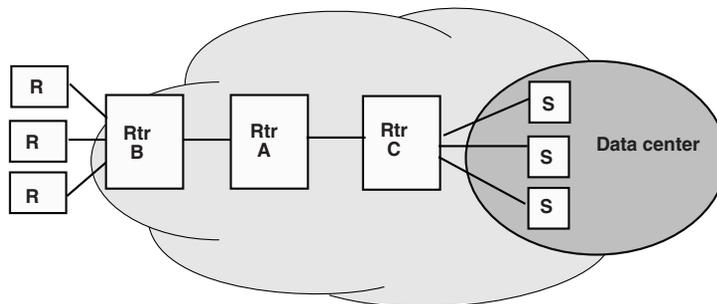
Do not advertise self

The *do not advertise self* feature represents a special case of DVMRP policies. The essential benefit is that it is easier to configure than regular announce policies, while providing a commonly-used policy set. Functionally, DVMRP does not advertise any local interface routes to its neighbors when you enable this feature. However, it will still advertise routes that it receives from neighbors. Because this disables the ability for networks to act as a source of multicast streams, you should not enable it on any routers that are directly connected to senders.

Figure 83 shows a common example of using this feature in DVMRP networks. Router A is a core router that has no senders on any of its connected networks. Therefore, it is unnecessary that its local routes be visible to remote routers, so it is configured to not 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, you configure Router B in a similar fashion to ensure it does not advertise any local routes either.

for the remote router to be visible to its local routes, so it configured to not advertise local routes. This makes it purely a transit router then. In contrast, Router B is an edge router that is connected to potential receivers. None of these hosts are allowed to be a source, so you configure Router B similarly and have it not advertise local routes either.

Since 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 there are certain local routes (that do not contain sources) that should not be advertised, you can selectively enable *do not advertise self* on a per interface basis, or configure announce policies instead.

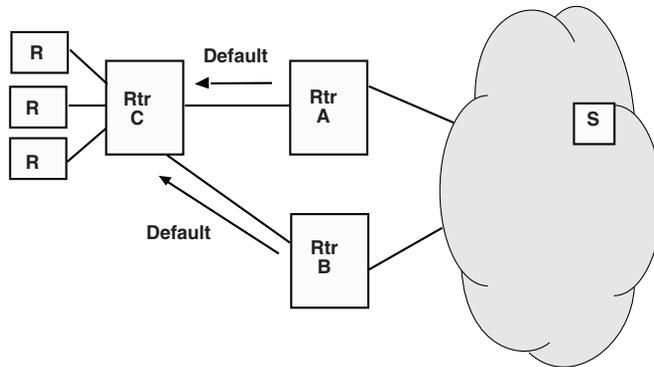
Figure 83 Do not advertise local route policies

Default route policies

DVMRP default route policies are special types of accept and announce policies you apply to the default route. You use the feature primarily 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.

Note that the default route does not appear in the routing table of the supplier. You can configure an interface to not listen for the default route. Once a default route is learned from a neighbor, it is placed in the routing table and potentially advertised to its other neighbors depending on whether or not you configured the outgoing interfaces to advertise the default route. Be aware that 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, while 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. This 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 [Figure 84](#), Router A and B are both advertising the default route to Router C. Because Router A is the preferred path for multicast traffic, you configure it with a lower metric (a value of 1 in this case), than 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 84 Default route

It is also recommended 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 being installed in the routing table.

DVMRP passive interface

The passive interface feature allows you to create a DVMRP interface to act like a IGMP interface only. In other words, no DVMRP neighbors and hence no DVMRP routes are learned on that interface. However, multicast sources and receivers can exist on that interface.

Such a feature is highly useful in cases where you wish to have IGMP snoop and DVMRP on the same switch. Currently, Layer 2 IGMP (IGMP snoop) and L3 IGMP (with DVMRP and PIM) on the same switch operate independently of each other. Thus, if you configure DVMRP on interface 1 and IGMP snoop on interface 2 on Switch A, multicast data with source from interface 1 is not forwarded to the receivers learned on interface 2 and vice versa. To overcome this problem, you can use DVMRP passive interfaces.

A DVMRP passive interface does not send probes or reports, does not listen for probes or reports and does not form neighbor relationships with other DVMRP routers. Instead, it acts exactly like IGMP snoop, except it is on Layer 3. However, the interface routes of the passive interfaces are still advertised on other active DVMRP interfaces. Since the passive interfaces provide less overhead to the

protocol, you will find it highly useful to configure certain interfaces as passive when there are many DVMRP interfaces on the switch. On the Passport 8600, you can change an existing DVMRP interface to a passive interface only if the interface is disabled by management.

You should only configure passive interfaces on those interfaces containing potential sources of multicast traffic. If the interfaces are connected to networks that only have receivers, it is recommended that you use a *do not advertise self* policy on those interfaces.



Note: You should not attempt to disable a DVMRP interface if there are multicast receivers on that interface.

In the event that it is necessary to support more than 512 or so potential sources on separate local interfaces, you should 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.

General design considerations with PIM-SM

The following sections discuss the guidelines you should follow in designing PIM networks:

- [“General requirements,”](#) next
- [“Recommended MBR configuration”](#) on page 251
- [“Redundant MBR configuration”](#) on page 252
- [“MBR and DVMRP path cost considerations”](#) on page 255
- [“PIM passive interface”](#) on page 255
- [“Static RP”](#) on page 256
- [“RP placement”](#) on page 260

General requirements

It is recommended that you design simple PIM networks where VLANs do not span several switches.

PIM relies on the 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) and BSR in the network. In addition, your design should ensure that the path between a sender and receiver contains PIM enabled interfaces. Note that receiver subnets may not always be required in the routing table. However, Nortel Networks recommends that you follow these guidelines in using PIM-SM:

- Ensure that a PIM-SM domain is configured with an 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. Your CLI configuration should be as follows:

```
candrp add 224.0.0.0 mask 240.0.0.0 rp <RP's IP address>
```

- Configure an RP to handle a range of multicast groups 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, be aware that you cannot configure one of the (local) interfaces for the static RP switches as RP. For example:

```
(static rp switch) Sw1 ----- Sw2 (BSR/cand-RP1) -----Sw3
```

you cannot configure one of the interfaces on switch Sw1 as static RP since BSR cannot learn this information and propagate it to Sw2 and Sw3. The PIM RFC requires that you consistently set RP on all the routers of the PIM domain. In other words, you can only add the remote interface candidate-RP1 (cand-RP) to the static RP table on Sw1.

- Static RP cannot be enabled or configured on a switch in a mixed mode of candidate RP and static RP switches, if a switch needs to learn an RP-set and has a unicast route to reach the BSR through this switch.

Example configuration 1

Sw3 (BSR) - Sw4 (candidate RP)

|

|

Sw2 (cannot be configured as a static RP)

|

|

Sw1 (PIM enabled, needs to learn RP-set through Sw2)

Example configuration 2

Sw3 (BSR) - Sw4 (candidate RP)

| /

| Sw5

Sw2 / (cannot be configured as a static RP)

| Sw6

| /

Sw1 (PIM enabled. It does have a route to BSR through Sw5, but shortest path route to BSR is through Sw2).

SPT switchover

When IGMP receivers join a multicast group on a Passport 8600, 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 switchover to the shortest path tree (SPT).

The Passport 8600 does not support a threshold bit rate in relation to SPT switchover in order to guarantee a simple, yet high performance implementation of PIM- Sparse Mode (SM). Note that intermediate routers (i.e. no directly connected IGMP hosts), do not switchover to the SPT until directed to do so by the leaf routers.

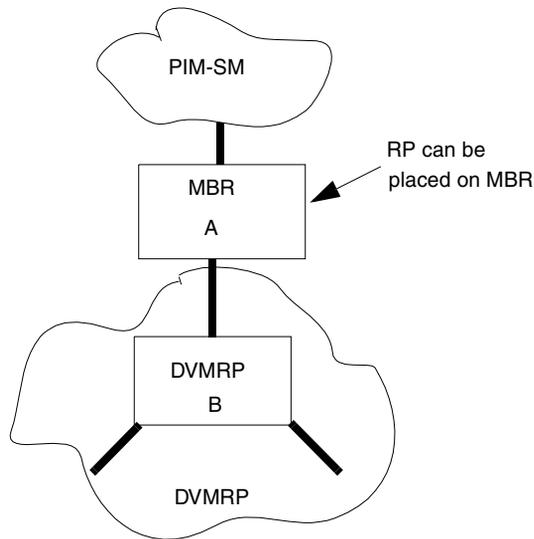
Other vendors may offer a configurable threshold, such as a certain bitrate, at which point SPT switchover occurs. Regardless of the implementation of these features on other equipment in the network, no interoperability issues with the Passport 8600 will result since switching to and from the shared and shortest path tree are independently controlled by each downstream router.

Upstream routers relay the joins or prunes hop by hop upstream, thus building the desired tree as they go. Since any PIM-SM compatible router already supports shared and shortest path trees, you should encounter no compatibility issues stemming from the implementation of configurable switchover thresholds.

Recommended MBR configuration

The MBR functionality provided with the PIM-SM implementation lets you connect a PIM-SM domain to a DVMRP domain. Note that a switch configured as an MBR will have both PIM-SM and DVMRP interfaces.

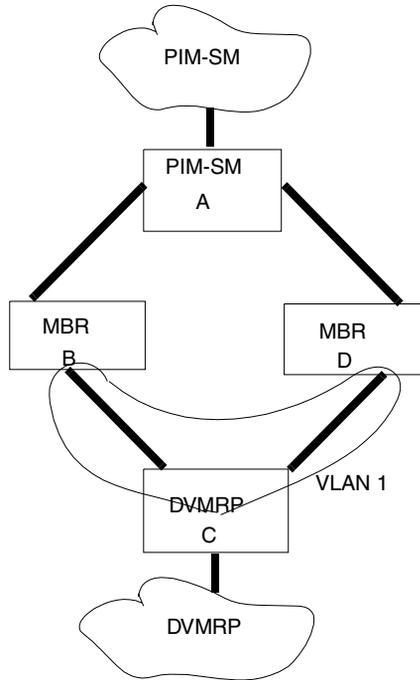
The easiest way to configure an MBR is to have one switch connecting the PIM-SM to the DVMRP domain, even though it is possible to use redundant switches for this purpose. In the first instance, you can use more than one interface on that switch to link the domains together. [Figure 85](#) illustrates this most basic scenario.

Figure 85 MBR configuration

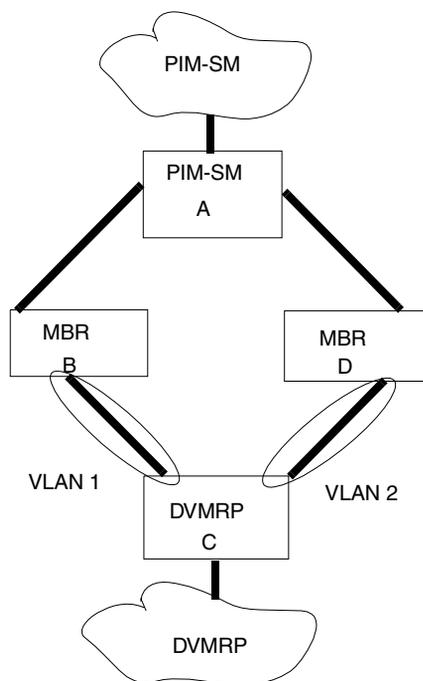
Note that with the Passport 8600 implementation, there is no limitation on the RP placement in the network.

Redundant MBR configuration

Figure 86 shows a redundant MBR configuration where two MBR switches connect a PIM to a DVMRP domain. Be aware that this configuration is not a supported configuration. The reason for this is that MBRs connecting two domains should not extend the same VLAN on the links connected to the same domain.

Figure 86 Redundant MBR configuration

In a proper configuration, you should ensure that the links use two separate VLANs as displayed in [Figure 87](#). For proper operation, ensure that the unicast routes and DVMRP routes always point to the same path in redundant MBR configurations.

Figure 87 Redundant MBR configuration with two separate VLANs

In [Figure 87](#), 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 lets you reach source A through B and that DVMRP shows upstream switch B to reach the source on A. If so, data flows from A to B to C and traffic coming from D is discarded.

If the link between C and B fails, switch C's unicast route indicates 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 will not create an mroute for the stream. Thus, data is discarded and it will never recover. To resolve this issue, you should stop the affected streams until DVMRP ages out the entries. Another alternative is to reinitialize DVMRP (disable and re-enable) and then restart the multicast streams.

If stopping DVMRP or the streams is not possible, you should lower the DVMRP timers for a faster convergence than the unicast routes. This solves the problem since DVMRP learns its routes before PIM learns the new unicast routes and reroutes the stream.

This same problem may happen in dynamic route changes if DVMRP and unicast routes diverge while traffic is flowing. As a result, Nortel Networks recommends that you use the simple design proposed in [“Recommended MBR configuration” on page 251](#) for safe MBR network operation.

MBR and DVMRP path cost considerations

When using the MBR to connect PIM-SM domains to DVMRP domains, ensure that the unicast routes metric is not greater than 32 since issues may occur in the network. This is due to the fact that the DVMRP maximum metric value is 32. On the MBR, DVMRP obtains metric information for the PIM domain routes from unicast protocols. Note that there may be cases where these metrics are higher than 32. If DVMRP finds a route with a metric higher than 32 on the MBR, this route is considered a non-reachable route and the RPF check fails, resulting in data not being 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.

PIM passive interface

Release 3.5 introduced the PIM passive interface feature. The PIM passive interface has the same uses and advantages as the DVMRP passive interface. Refer to [“DVMRP passive interface” on page 247](#) for more details.

Circuitless IP for PIM-SM

The Passport 8600 provides you with a resilient way to configure an RP and BSR for a PIM network using the circuitless IP capability. When configuring an RP or BSR on a regular interface, if this interface becomes non-operational (e.g., because no ports on this interface are active), the RP and BSR become non-operational too. This results in the election of other redundant RPs and BSRs,

if any, and may disrupt IP multicast traffic flowing in the network. As a sound practice for multicast networks design, you should always configure the RP and BSR on a circuitless IP interface to prevent a single interface failure from causing these entities to fail.

It is also recommended that you configure redundant RPs and BSRs on different switches and that these entities be on circuitless interfaces. For the successful setup of multicast streams, you must ensure that there is a unicast route to all these circuitless interfaces from all locations in the network. This is mandatory because every switch in the network needs to reach the RP and BSR for proper RP learning and stream setup on the shared RP tree. When used for PIM-SM, be aware that circuitless IP interfaces can only be utilized for RP and BSR configurations and are not intended for other purposes.

Static RP

You can use static RP very effectively to provide security, interoperability, and/or redundancy for PIM-SM multicast networks. In some networks, the administrative ease derived from dynamic RP assignment may not be worth the security risks involved. For example, if an unauthorized user connects a PIM-SM router to the network, which advertises itself as a candidate RP (C-RP or cand-RP), it may possibly hijack new multicast streams that would otherwise be distributed through an authorized RP. If you are designing networks where security is important, you may find such risks unacceptable. In such cases, you will find a static RP assignment preferable.

The static RP feature also provides you with interoperability capabilities in legacy PIM-SMv1 and mixed vendor PIM-SMv2 networks. Since PIM-SMv1 did not support dynamic RP assignment, auto-RP was developed as a proprietary solution until the standards-based BSR was created as part of the PIM-SMv2 standard. The only RP assignment options you have available for legacy PIM-SMv1 networks are static configuration and auto-RP. For legacy networks that use static RP configuration, you can easily integrate the Passport 8600 into the network.

Auto-RP protocol

Other legacy PIM-SMv1 networks may use the auto-RP protocol. Auto-RP is a Cisco proprietary protocol that uses two proprietary defined protocols to provide equivalent functionality to the standard PIM-SM RP and BSR implemented on the Passport 8600. You can use the static RP feature to interoperate in such

environments. For example, in a mixed vendor network, you can use auto-RP among routers that support the protocol, while other routers such as the Passport 8600 have RP information statically configured. In such a network, you must ensure that the static RP configuration mimics the information that is dynamically distributed in order to guaranteed that multicast traffic is delivered to all parts of the multicast network.

In a mixed auto-RP and static RP network, you also need to ensure that the Passport 8600 does not serve as an RP since the 8600 does not support the Cisco proprietary auto-RP protocol. Therefore, in this type of network, the choice of RPs is limited only to the routers that support the auto-RP protocol.

RP redundancy

The Passport 8600 can also provide RP redundancy when you configure static RPs. You need to implement the same static RP configuration on all PIM-SM routers in the network to ensure consistency of RP selection. Furthermore, you must ensure in a mixed vendor network that the same RP selection criteria is used among all routers.

For example, the Passport 8600 uses the hash algorithm defined in the PIM-SMv2 standard to select the active RP for each group address. (See the [“BSR hash algorithm” on page 260](#) for more information). If a router from another vendor selects the active RP based on lowest IP address, then the inconsistency causes the stream to not be delivered to certain routers in the network.

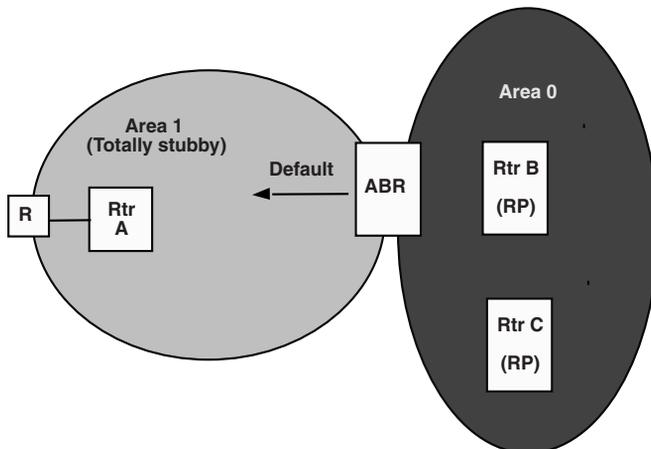


Note: When there is a discrepancy concerning the group address to RP assignment among PIM-SM routers, network outages occur. Routers that are unaware of the true active RP are unable to join the shared tree and receive the multicast stream.

Failure detection of the active RP is determined by the routes available in the unicast routing table. As long as the RP is considered reachable from a unicast routing perspective, the local router assumes the RP is fully functional and attempts to join that RPs shared tree. Thus, you should take extra care when using static and default routes in the network.

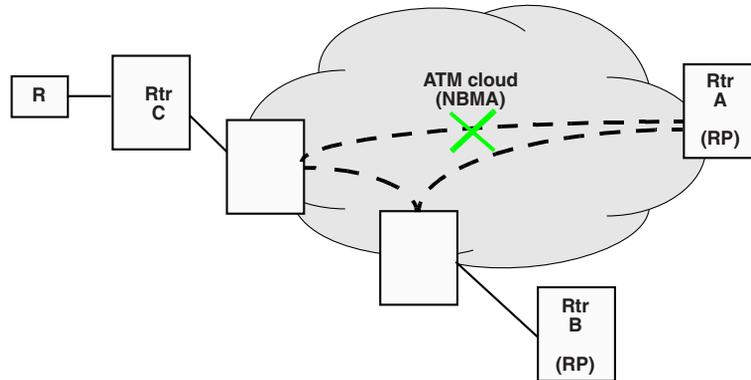
Figure 88 shows a hierarchical OSPF network where a receiver is located in a totally stubby area. If RP B fails, PIM-SM router A does not switch over to RP C because the injected default route in the unicast routing table indicates that RP B is still reachable.

Figure 88 RP failover with default unicast routes



In Figure 89, the IP address you selected for the RP is the interface connected to an ATM non-broadcast multiaccess (NBMA) network. If RP A fails or if the virtual circuits connecting RP A to the ATM network fail, the desired behavior is to have the other multicast routers switch over to RP B. However, because the network or subnet corresponding to the rest of the ATM NBMA network is still reachable, it is still advertised by the unicast routing protocol. Therefore, multicast routers still attempt to use RP A as the active RP.

These figures illustrate that since failover is determined by unicast routing behavior, you need to give careful consideration to the unicast routing design, as well as the IP address you select for the RP.

Figure 89 Interface address selection on the RP

The performance aspect of static RP failover is dependent on the convergence time of the unicast routing protocol. Thus, it is recommended that you use a link state protocol, such as OSPF, in order to take advantage of the quick convergence. For example, if you are using RIP as the routing protocol, an RP failure may take minutes to detect. Depending upon the application, you may find this totally unacceptable. Note, that this problem does not affect routers that have already switched over to the SPT, however. It just affects newly-joining routers.

Non-supported static RP configuration

Be aware that with static RP, dynamic RP learning is not operational. The following example shows a non-supported configuration for static RP. In this example, with static RP and dynamic RP interoperation, there is no RP at switch2. However, (S,G) create/delete occurs every 210 seconds at switch 16.

```
sw 10 ----- sw 2
```

```
|           |
```

```
|           |
```

```
sw 15 ----- sw 16
```

Sw 10, 15, and 16 are under StaticRP, while Sw 2 is under dynamic RP. The src is at Sw 10 with the Rx Sw 15, and 16. RP is then at Sw 15 locally. The Rx on Sw 16 is unable to receive packets because its SPT is going thru Sw 2.

Sw 2 is in a dynamic RP domain. Thus, there is no way to learn about RP on Sw 15. However, you will see an (S, G) record being created/deleted on Sw16 every 210 seconds.

RP placement

The following sections describe guidelines for RP placement:

- [“BSR hash algorithm,”](#) next
- [“RP and extended VLANs”](#) on page 264
- [“Receivers on interconnected VLANs”](#) on page 264
- [“PIM network with non-PIM interfaces”](#) on page 265

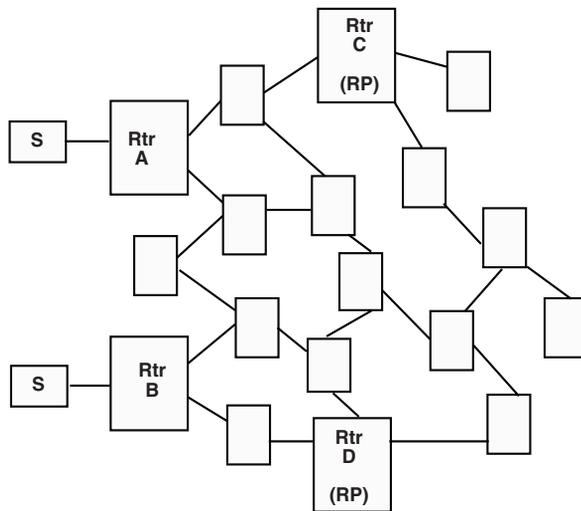
BSR hash algorithm

If your network is using a bootstrap router (BSR) to dynamically advertise candidate RPs and if you require redundancy, you may wish to consider the multicast group address to RP assignment function in order to optimize the flow of traffic down the shared trees. A hash function is used to assign small blocks of multicast group addresses to each RP, which is a candidate for the same range.

The hash mask used to compute the RP assignment is also distributed by the BSR. 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 4 consecutive addresses and assigned to one or the other candidate RP.

Figure 90 illustrates a sub-optimal design where Router A is sending traffic to a group address assigned to RP D. Router B is then sending traffic assigned to RP C. Note that RP C and RP D serve as backups for each other for those group addresses. In such cases, you may find it desirable instead to have traffic from Router A use RP C and traffic from Router B use RP D.

Figure 90 Inefficient group-RP mapping



While providing redundancy in the case of the failure of a particular RP, you can ensure the optimal shared tree is used in two ways.

- 1 Use the hash algorithm to proactively plan and understand the group address to RP assignment.

You use this information to select the multicast group address for each multicast sender on the network and thus, ensure optimal traffic flows. You may also find this method helpful for modeling more complex redundancy and failure scenarios where each group address has three or more C-RPs.

- 2 Allow the hash algorithm to assign the blocks of addresses on the network, and then simply view the results on the Passport 8600 with the **show ip pim active-rp <group>** command.

You can then use the results 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. In the event that there is more than one backup for a group address, the secondary RP is not obvious. In this case, if you use the hash algorithm, it reveals which of the remaining C-RPs will take over for a particular group address in the event of primary RP failure.

Hash algorithm operation

The hash algorithm works as follows:

- 1 For each C-RP with matching group address ranges, a hash value is calculated according to the formula:

$$\text{Value (G,M,C(i))} = (1103515245 * ((1103515245 * (G\&M) + 12345) \text{ XOR } C(i) + 12345) \text{ mod } 2^{31})$$

As you can see, the hash value is a function of the group address (G), the hash mask (M), and the IP address of the C-RP (C(i)). The expression (G&M) guarantees that blocks of group addresses hash to the same value for each C-RP, and 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 C-RP. Thus, the block of 8 addresses are assigned to the same RP selected by the criteria as listed in step 2.

- 2 The results are compared, and the C-RP with the highest resulting hash value is chosen as the RP for the group. In the event of a tie, the C-RP 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.

Cand-RP priority

You can also use the candidate-RP priority parameter to determine an active RP for a group. Here, the hash value for different RPs are compared for only RPs with the highest priority (lowest cand-RP priority value). A cand-RP with highest RP IP address is then chosen as the active RP among RPs with the highest priority value and same hash value.

The current Passport 8600 PIM-SM implementation does not allow you to configure the cand-RP priority. Thus, each RP has a default cand-RP priority value of 0 and the algorithm uses the RP if the group address maps to the grp-prefix you configure for that RP. If a different router in the network is configured with cand-RP with a priority value > 0, the Passport 8600 uses this part of the algorithm in the RP election, even though it cannot be configured with a priority other than 0.

Currently, you cannot configure the hash mask used in the hash algorithm. As a result, the default hash mask of 255.255.255.252 is used, unless you configure a different PIM BSR in the network with a non-default hash mask value. Note that static RP configurations do not use the BSR hash-mask, so they are limited to the default hash mask value when configuring redundant RPs.

For example:

RP1 = 128.10.0.54 and RP2 = 128.10.0.56

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:

grp range 238.1.1.40 - 238.1.1.51 (12 consecutive groups) mapped to 128.10.0.56

grp range 238.1.1.52 - 238.1.1.55 (4 consecutive groups) mapped to 128.10.0.54

grp range 238.1.1.56 - 238.1.1.63 (8 consecutive groups) mapped to 128.10.0.56

RP and extended VLANs

With Release 3.2.2 of the Passport 8000 Series software, you are no longer restricted on where you can place an RP when VLANs extend over several switches. Indeed, you can now place your RP on any switch in the network. When using PIM-SM, however, it is always recommended that you not span VLANs on more than 2 switches.

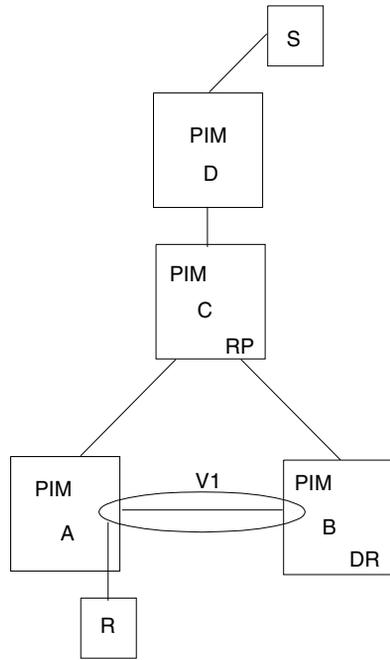
Receivers on interconnected VLANs

There are designs where IP multicast traffic can flow unnecessarily on some links in a PIM-SM domain. Traffic in these cases does not get duplicated to the receivers. However, be aware that it can use extra bandwidth on links in the network.

[Figure 91](#) displays such a situation. Switch B is the Designated Router (DR) between A and B. Switch C is the RP. A receiver R is placed on the VLAN1 interconnecting switches A and B. A source is sending multicast data to receiver R.

The IGMP reports sent by R are then forwarded to the DR and both A and B create (*,G) records. Switch A receives duplicate data, via the path from C to A and via the second path from C to B to A. It discards the data on the second path assuming the upstream source is A to C.

To avoid this situation, Nortel Networks recommends that you not place receivers on V1. This guarantees that no traffic will need to flow between B and A for receivers attached to A. Only through PIM join messages to the RP (for (*,G)) and the source through SPT joins will the existence of the receivers be learned.

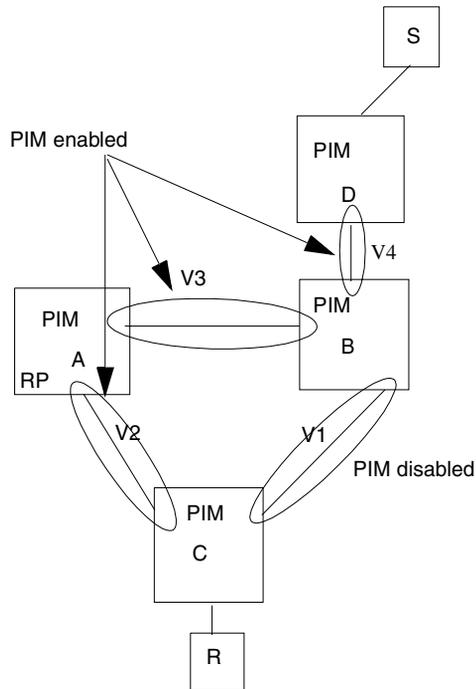
Figure 91 Receivers on interconnected VLANs

PIM network with non-PIM interfaces

As a general rule, in a PIM-SM domain, you need to enable PIM-SM on all interfaces in the network for multicast traffic to flow properly. This is true even if paths exist between all PIM interfaces. The reason for this is that PIM-SM relies on the unicast routing table to determine the path to the RP, BSR and to multicast sources. Thus, you should ensure that all routers on these paths have PIM-SM enabled interfaces.

[Figure 92](#) provides an example of this situation. Here, if A is the RP, then initially, receiver R gets data via the shared tree path (i.e., through switch A).

If the shortest path from C to the source is through switch B, while the interface between C and B does not have PIM-SM enabled, then C will not switch to the SPT. It then starts discarding data coming through the shared tree path (i.e., through A). The simple workaround for this issue is for you to enable PIM on VLAN1 between C and B.

Figure 92 PIM network with non-PIM interfaces

Multicast and SMLT

This section discusses guidelines that you should follow for multicast and SMLT configurations:

- [“Triangle designs,”](#) next
- [“Square designs”](#) on page 269
- [“Design that avoids duplicate traffic”](#) on page 270
- [“DVMRP versus PIM”](#) on page 272

Triangle designs

A triangle design is an SMLT configuration where you connect edge switches, or SMLT clients to two aggregation switches. You connect the aggregation switches together with an IST that carries all the SMLTs configured on the switches.

With a triangle design, the following configurations are supported:

- a configuration with all Layer 2 IP multicast with IGMP snooping
- a configuration with Layer 2 snooping at the client switches and Layer 3 routing with DVMRP or PIM-SM at the aggregation switches

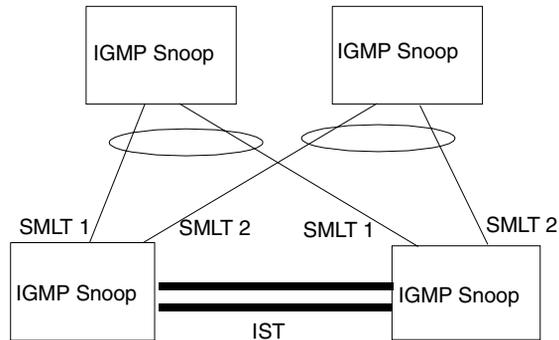
If the client switches are Passport 8600 switches, you need to use E- or M-modules for proper operation. Using non-E-modules, may result in loops developing on the SMLT links. You can position switches, other than the Passport 8600, such as the Baystack 450 or the BPS 2000, as clients in a triangle configuration.

The sub-sections that follow describe the supported SMLT triangle design configurations in more detail.

All Layer 2 IGMP snooping

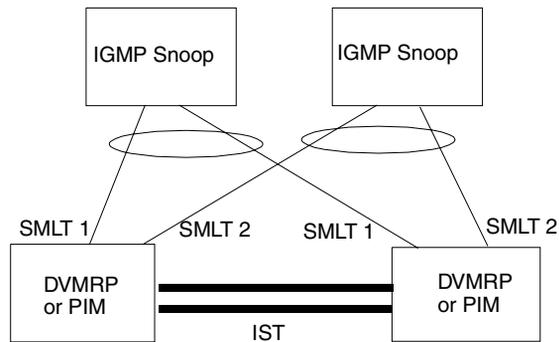
The design shown in [Figure 93](#) is supported. In this design, you need to attach a querier to every VLAN that is configured with IGMP snooping. You can attach this querier to any of the switches. If you require redundancy for the querier, then you can configure several queriers. However, they must be on the same switch. For different VLANs, you can locate the querier(s) on different switches. You can use the queriers for IP multicast routing, if necessary.

If the client switches are Passport 8600 switches, you must use E- or M-modules in this design. Note that you do not need to include E- or M-modules in the aggregation switches, however.

Figure 93 Layer 2 IGMP snooping

Layer 2 and Layer 3 multicast

To avoid using an external querier to allow correct handling and routing of multicast traffic to the rest of the network, Nortel Networks recommends that you use the triangle design with IGMP snooping at the client switches. You should then use multicast routing using DVMRP or PIM at the aggregation switches as shown in [Figure 94](#).

Figure 94 Multicast routing using DVMRP or PIM

Client switches run IGMP snooping and the aggregation switches run PIM or DVMRP. This design is simple and lets you perform IP multicast routing by means of DVMRP or PIM for the rest of the network. The aggregation switches are the queriers for IGMP, thus, there is no need for you to have any external querier to activate IGMP membership. These switches also act as redundant switches for IP multicast.

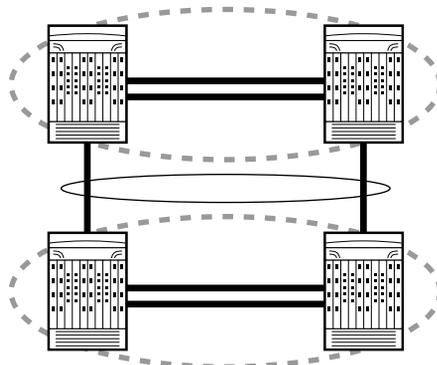
Multicast data flows through the IST link when receivers are learned on the client switch and senders are located on the aggregation switches, or when sourced data comes through the aggregation switches. This data is destined for potential receivers attached to the other side of the IST. It does not reach the client switches through the two aggregation switches since only the originating switch forwards the data to the client switch receivers.

Note that you should always place any multicast receivers and senders on the core switches on different VLANs than those that span the IST.

Square designs

A square design is a design where you connect a pair of aggregation switches to another pair of aggregation switches (Figure 95). If you connect the aggregation switches in a full mesh, then it is a full mesh design. Note that the full mesh design does not support SMLT and IP multicast.

Figure 95 Square design- full mesh configuration



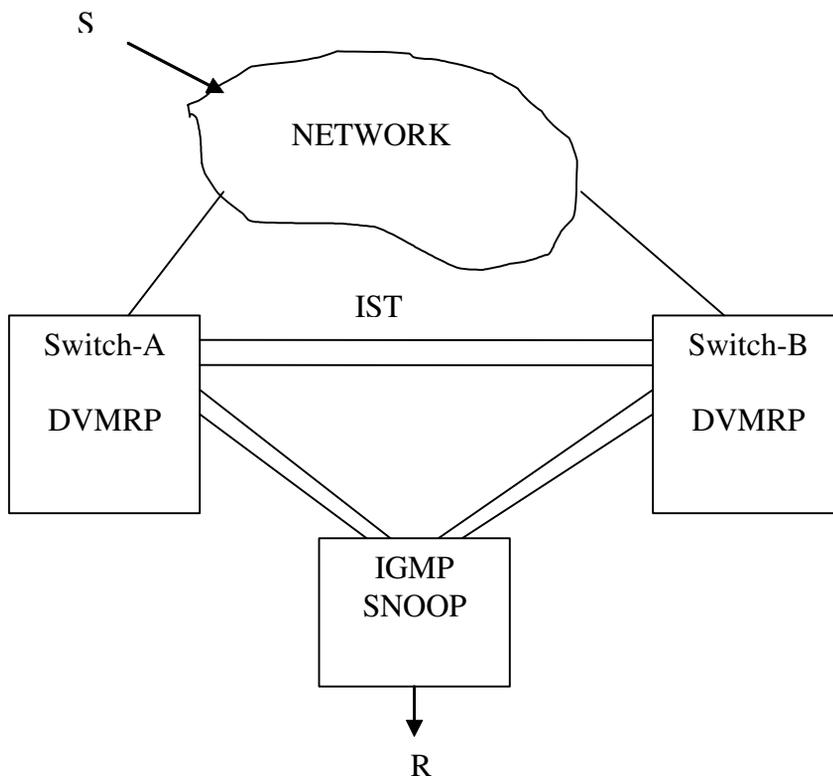
10674BEA

In a square design, you must configure all switches with IGMP snooping. No IP multicast routing is allowed with DVMRP or PIM. You also have to configure a querier to connect to every VLAN with IGMP snooping. If you need redundancy for the querier, you can configure several queriers. However, they have to reside on the same switch. For different VLANs, you can configure the querier(s) on different switches. If necessary, you can also use the queriers for IP multicast routing.

Design that avoids duplicate traffic

This section describes a potential “duplicate traffic” issue that can occur with multicast and SMLT networks. When the path to a source network on the aggregation switches is the same for both switches, it can result in duplicate traffic. This section provides a solution to avoid this issue. [Figure 96](#) illustrates the issue and the solution.

Figure 96 Multicast and SMLT design that avoids duplicate traffic



Switch-A and Switch-B learn the DVMRP route for sender (S) network with the same metric.

- Assume that Switch-A is the querier for the interface connected to the Layer 2 switch with IGMP snooping.
- When receiver R sends an IGMP report, A learns the receiver on the SMLT port and forwards the IST-IGMP message to B.
- On receiving the message from A, B learns the receiver on its SMLT port connected to the Layer 2 switch. So, both A and B have local receivers on their SMLT port.
- The multicast sender S sends data that is received by both A and B through the interface connected to NETWORK. Since both have a local receiver on SMLT port, the Layer 2 switch will receive data from both the routers causing R to receive duplicate traffic.

Assume that the source network is 10.10.10.0/24, Switches A and B have the DVMRP metric on the IST interface, the interface towards NETWORK are all configured as 10, and that the total cost to the source is the same.

You can view DVMRP route details with the following command:

```
show ip dvmrp route
```

- Switch-A has DVMRP route 10.10.10.0 with metric 10 and upstream neighbor through interface connecting NETWORK.
- Switch-B has DVMRP route 10.10.10.0 with metric 10 and upstream neighbor through interface connecting NETWORK.

In this configuration, both A and B forward traffic to the SNOOP switch and the receiver gets duplicate traffic.

The solution to this issue is to configure metrics on the DVMRP interfaces so that either A or B learn the source network route through IST. In this way, the router that receives traffic from IST blocks them from being sent on SMLT (receiver) port so the SNOOP switch receives traffic from only one router.

In the above example, configure the metric of the DVMRP interface towards NETWORK on any one switch. For example, configure Switch B so that the route metric through that interface on B will be greater than the metric through the IST interface. Therefore, the NETWORK interface on B should be greater than 2.

To configure the DVMRP interface metric, use one of the following commands:

For an interface, use

```
config ip dvmrp interface <ip_addr> metric <cost>
```

For a VLAN, use

```
config vlan <vlan_id> ip dvmrp metric <cost>
```

For a brouter port, use

```
config ethernet <port_num> ip dvmrp metric <cost>
```

If the metric of NETWORK interface on B is configured to 3, then B can learn route 10.10.10.0 through the NETWORK interface with metric 12 (because its metric is incremented by 2) and through IST with metric 11. So B will learn route 10.10.10.0 with cost 11 and the upstream neighbor through the IST link.

Now, traffic from S would follow via A to B on the IST link only. Since traffic received on the IST cannot go to the SMLT link, the SNOOP switch will not receive from B. Therefore, R no longer receive duplicate traffic any more and receives its traffic from switch A only.

DVMRP versus PIM

DVMRP and PIM have some major differences in the way they operate and forward IP multicast traffic. This section provides guidelines for network designers to choose which protocol is better adapted to their environment, and what are the factors to consider when using one protocol and not the other or even using a mix of the two protocols in different sections of the network and linking them together with the MBR functionality.

Flood and prune versus shared and source trees

The first difference is related to the flood and prune type of operation of DVMRP versus the use of the Shared tree and SPT of PIM-SM. These differences make DVMRP more adapted to a dense environment where receivers are present in most parts of the network, hence its flood and prune mechanism is well adapted to this environment. On the other hand, PIM-SM is better suited for a sparse environment where few receivers are spread over a smaller part of the network and a flooding mechanism wouldn't be beneficial and efficient. This is particularly true if the protocol is used in places where the network is quite large.

In a network where there are few receivers for multicast streams, DVMRP results in a lot of initially flooded traffic that gets pruned. Not only is there initial flooding here, there is also periodic flooding (once all of the prunes have expired). This occurs if the source continues sending multicast data and adds unnecessary traffic in the network, especially on those branches where there are no receivers. It also adds additional state information on switches with no receivers.

With PIM-SM all initial traffic has to flow to the RP before reaching the destination switches so that they can get to the SPT path to receive data directly from the source. This makes PIM-SM vulnerable to the RP failure, hence the support for redundant RPs in the Passport 8600 implementation. Even with redundant RPs, the convergence time can be faster in DVMRP than in PIM depending on where the failure occurs.

Another drawback of having initial traffic flowing through the RP is that the RP may become a bottleneck if too many streams are initialized at the same time, which will result in too many register data that will reach the RP switch before receiver switches get data on the SPT. In applications like TV applications or where streams are high bandwidth consumers, this can cause some performance issues on the RP switch and may result in longer stream initialization times. To reduce the effect of this PIM-SM inherent operation, the Passport 8600 implementation allows immediate switching to the SPT with the first received packet and this is true for the RP receiving register packets or the switches receiving packets on the shared tree.

Unicast routes for PIM versus DMVRP own routes

DVMRP uses its own RIP-2 based routing protocol that allows it to have a different routing table than unicast, hence giving the flexibility to build different paths for multicast than for unicast traffic. PIM-SM relies on the unicast routing protocols to build its routing table, hence its paths are always linked to unicast paths. So, even though PIM-SM is independent on the unicast routing protocol as such it has to rely on one to get its routing information. With DVMRP, the decoupling of the routing table from the unicast has another advantage: route policies can be applied to DVMRP regardless of what the unicast route policies are. In the case of PIM, it has to follow the unicast routing policies limiting the flexibility in “tuning” the way routes for PIM are to be handled.

One advantage with PIM-SM routing to DVMRP is that PIM-SM scales to the unicast routing table which is several thousands, while DVMRP has limited route scaling (two to three thousand maximum) because of the nature of its RIP-2 based route exchange. This make PIM-SM more scalable in large networks where the number of routes exceed the number supported by DVMRP in the case where DVMRP policies cannot be applied to reduce the number of routes.

Convergence and timers

DVMRP includes configurable timers that provide you with more control on the network convergence in case of failures. PIM requires unicast convergence before it can converge, thus, it may take longer for PIM to converge as compared to DVMRP.

Traffic delay with PIM while rebooting peer SMLT switches

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 for some reason, the router with the next highest IP address becomes the DR.

Rebooting the DR in an SMLT VLAN may result in data loss because of the following actions:

- When the DR is down, the non-DR 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 non-DR 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 2-3 minutes (since the DR learns the RP set after OSPF converges).

A workaround is to configure static rendezvous point (RP) on the peer SMLT switches. This feature avoids the process of selecting an active RP from the list of candidate RPs and dynamically learning about RPs through the BSR mechanism. Then when the DR comes back, traffic resumes as soon as OSPF converges. This reduces the traffic delay to approximately 10 seconds.

Enabling multicast on network interfaces

In some cases, if you have to disable PIM on some interfaces, you need to ensure that all paths to the RP, BSR, and multicast traffic sources for any receiver on the network have PIM enabled. This is due to the fact that 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 BSR. A PIM-enabled switch only accepts the BSR message from the RPF neighbor towards BSR.

Note that DVMRP does not have the same constraint since the existence of one path between a source and a receiver is enough to obtain the traffic for that receiver. In [Figure 92 on page 266](#), 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.

Reliable multicast specifics

This section includes some specific design tips you can follow for networks that require the support of highly reliable multicast. In addition to the inherent reliability of the Passport 8600, and features like SMLT, you need to take some specific considerations into account for the multicast protocols. These include protocol tuning for faster convergence and some tips for using the PGM reliable transport-level protocol. The main application of these design tips is in the financial space, where data has to be error-free and scalable. In addition, the network has to converge rapidly to handle any failure that can occur.

Protocol timers

DVMRP lets you configure several protocol timers to help in obtaining faster convergence. Even though Nortel Networks recommends that you use the default protocol timers, some networks require fast convergence after failures or state changes. If this is the case, you should be careful how you configure the timers for the protocol. Refer to [“DVMRP timers tuning” on page 241](#) for additional information.

PGM-based designs

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 either receives all data from transmissions and retransmissions, or is capable of detecting unrecoverable data packet loss.

The Passport 8600 implements the Network Element part of PGM. Hosts running PGM implement the other features in PGM. PGM operates on a session basis, so every session requires state information in the Passport 8600. Therefore, it is important for you to control both the number of sessions that the Passport 8600 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 guidelines help you design PGM-based applications and parameters for better scalability. Note that they are based on memory consumption calculations for sessions with a given window size and assume that a maximum of 32MB is used by PGM:

- Memory usage of each PGM session and its association with `window_size`

The examples that follow are based on observations that occur when a session is created with a `window_size` of 5000 and a given amount of system memory is used. The number of bytes allocated in the system for each session = 4 bytes x (`win_size*2`) + overhead (=236 bytes)

— Example 1:

if 32Mb of system memory is available for PGM, the number of sessions you can create is $32\text{Mb} / 40,236 = 800$ sessions. Allowing more than 800 sessions (in this particular case), results in using more system memory and may impact other protocols running on the switch.

— Example 2:

if 1.6Mb of system memory is available for PGM, the number of sessions you can create is $32\text{Mb} / 16,236 = 100$ sessions.

- The recommendation here is that you 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 the window size, you should also limit the total number of PGM sessions in the system to control the amount of memory PGM uses.

Specifically, you should ensure that PGM does not consume the memory required by the other protocols running on the Passport 8600 switch. The default value for the maximum number of sessions on a Passport 8600 is 100.



Note: These guidelines will 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 the previous numbers.

Multicast stream initialization

At stream initialization, initial packets reach the CPU that programs the hardware records- based on IGMP membership and neighborhood information for DVMRP. With PIM, the Join and RP information is used to determine the ports to forward to.

During the initialization phase, there may be some data loss depending on how quickly the initial data reaches the switch. In Video-related applications, there may be some frame loss in the first milliseconds of the stream.

TV delivery and multimedia applications

The Passport 8600 provides you with a flexible and scalable multicast implementation for multimedia applications. Several features are dedicated to multimedia applications and in particular, to television distribution.

This section describes the main features in use here and explains the best way for you to use them when designing networks that support interactive TV applications in service provider environments. All other design tips provided in this chapter also apply to these applications.

Static (S,G)s with DVMRP and IGMP static receivers

The static forwarding features that apply to DVMRP allow you to configure static mroutes. This feature is useful in cases where streams must flow continuously and not become aged. You should be careful in using this feature, however, since you need to 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 is that these entries allow you to configure group address only. Note that you can use them as edge configurations, or on interconnected links between switches.

Join/leave performance

In TV applications, you can attach several TV sets directly or through BPS2000 switches to the Passport 8600. You base this implementation on IGMP and the set-top boxes use IGMP reports to join a TV channel and an IGMP leave to exit the channel. When a viewer changes channels, an IGMPv2 leave for the old channel (IP Multicast group) is issued, then a membership report for the new channel is sent to start obtaining traffic for the new channel. If viewers “channel surf” and change channels continuously, the joins and leaves can become large, particularly when there are sizeable numbers of viewers attached to the switch.

The Passport 8600 supports more than a thousand Joins/leaves per second, which is well adapted to scale in this type of application.



Note: For IGMPv3, Nortel Networks recommends a Join rate of 250 per second or less. If the Passport 8600 has to process more than 250 Joins per second, the user may have to re-send the Join.

When you use the IGMP proxy functionality in the BPS, you reduce the number of IGMP reports received by the Passport 8600, thus providing better, overall performance and scalability.

Fast leave

The Passport 8000 Series software, release 3.1 and above, supports the fast leave feature for IGMP. You use this feature along with IGMPv2, IGMPv3 and IGAP, which provide the leave functionality. Release 3.5 introduced the support for several users on an interface configured with Fast Leave, while prior releases did not have support for more than one user per interface. If you use fast leave on a port, there is no Group-Specific-Query sent on a port after a leave message is received on this port.

Release 3.5 provides two modes of operation for Fast Leave: Single User Mode and Multiple User Mode.

- Single User Mode works like the way it was before release 3.5. That is, if more than one member of the left group is on the port, everyone stops receiving traffic for this group as soon as one of the group members leaves the group. There is no Group-Specific-Query sent before allowing the effective leave to take place.
- Multiple User Mode allows you to have several users on the same port/VLAN, and one user leaving the group does not result in having the stream stop if there are other receivers for the same stream. The Passport 8600 achieves this by tracking the number of receivers that have joined a given group. This works properly under the condition that receivers send their joins regardless of the others sending joins (i.e., do not do a report suppression). This ensures that the Passport 8600 properly tracks the correct number of receivers on an interface.

You will find this feature particularly useful in IGMP-based TV distribution applications, where only one receiver of a TV channel is connected to a port. In the event a viewer changes channels quickly, considerable savings can be made in terms of bandwidth demand arising from the consequent IGMP leave process.

With release 3.1 and above of the Passport 8000 Series software, you can implement fast leave on a VLAN and port combination — a port belonging to two different VLANs can have the feature enabled on one VLAN but not on the other. This provides you with the ability to connect several devices on different VLANs, but on the same port with the Fast Leave feature enabled. You can do so without impacting the traffic for the devices when one of them is leaving a group that another is subscribed to. For example, you can use this feature when two TVs are connected to a port through two set-top boxes even if you are using the Single User Mode of the feature.

LMQI tuning

When an IGMPv2 host leaves a group, it notifies the router with a leave message. Because of the IGMPv2 report suppression mechanism, the router is unaware of other hosts that require the stream. Thus, it broadcasts a group specific query message with a maximum response time equal to the last member query interval (LMQI).

Since this timer affects the latency between the time the last member leaves and the stream actually stops, you must tune this parameter properly for TV delivery, or other large-scale, high-bandwidth multi-media applications. For instance, if you assign a value that is too low, it 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 problems too. It can result in too many unwanted high-bandwidth streams being propagated across the network when users change channels rapidly. Note that leave latency is also dependent on the robustness value, so a value of two equates to a leave latency of twice the LMQI.

The LMQI accepts values from 1 to 255 where the increment is in tenths of seconds. Since many other factors affect performance, you should determine the proper setting for your particular network through testing. If there are a very large number of users connected to a port (e.g., a port connected to an L2 switch), assigning a value of three might lead to a storm of report messages when a group specific query is sent. On the other hand, if streams are frequently started and stopped in short intervals, as in a TV delivery network, assigning a value of ten might 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 affect the proper choice of LMQI values. For links that have high loss characteristics, 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 go unanswered before the stream is pruned. The number of unanswered queries is equal to the robustness variable (default is two). Assigning 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 picking an LMQI value, you need to consider all of these factors to determine the best setting for the given application and network. You should then test that chosen value to ensure the best performance.



Note: In networks that have only one user connected to each port, it is recommended that you use the fast leave feature instead of LMQI since no waiting period is required before the stream is stopped. Similarly, the robustness variable has no impact on the fast leave feature, which is an additional benefit for links with high loss

IGAP

Internet Membership Group Authentication Protocol (IGAP) is an authentication and accounting protocol for clients receiving multicast streams. With IGAP's authentication and accounting features, service providers and enterprises have more control over their networks and can better manage the 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 like RADIUS with extensions for IGAP.

The Passport 8600 processes messages according to the following rules:

- On IGAP-enabled interfaces, the Passport 8600 processes IGAP messages and ignores all other IGMPv1, IGMPv2 or IGMPv3 messages.
- On IGMP interfaces, the Passport 8600 processes IGMP messages and ignores IGAP messages.
- IGAP operates with Fast Leave only and does not generate Group-Specific-Queries like in IGMPv2. The Passport 8600 supports the Single User and Multiple User fast leave modes for IGAP.

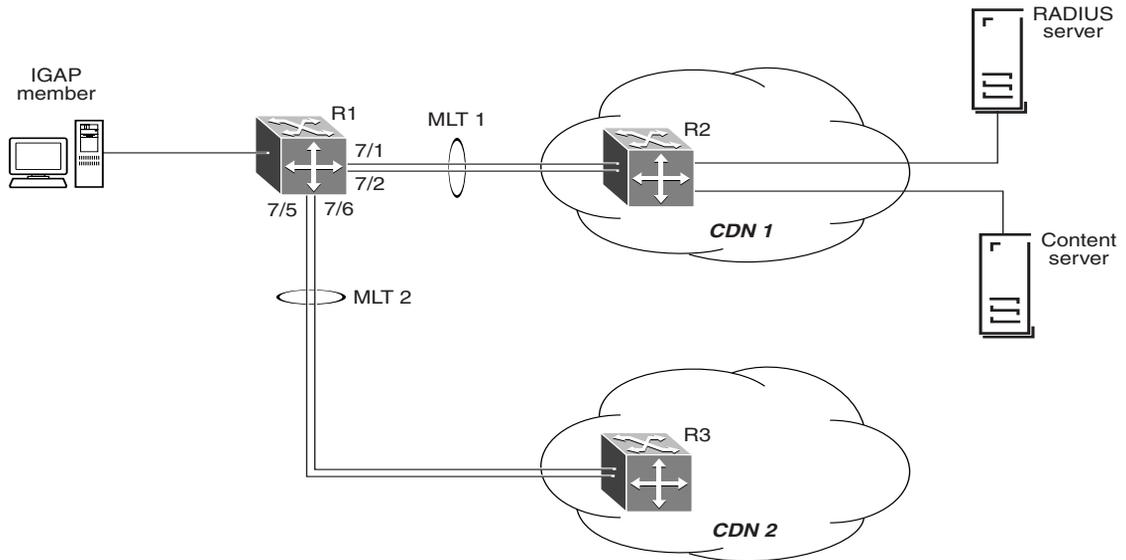
For more information about IGAP, see *Configuring Internet Membership Group Authentication Protocol (IGAP)*.

IGAP with an MLT link down

When an MLT link goes down, it can potentially interrupt IGAP traffic.

Figure 97 shows an IGAP member connected to a Passport 8600 edge switch (R1) that has two MLT links. The MLT links provide alternative routes to the RADIUS server and the Content Delivery Network (CDN) server.

Figure 97 Avoiding an interruption of IGAP traffic



Legend

-  Passport 8600 switch
- IGAP = Internet Group membership Authentication Protocol
- CDN = Content Delivery Network
- RADIUS = Remote Authentication Dial-In User Services

11057FA

The following scenario shows how a potential traffic interruption can occur:

- 1 IGAP member, who has already been authenticated, is receiving multicast traffic and accounting has started.
- 2 Passport 8600 (R1) uses MLT1 to transfer data and accounting messages.
- 3 MLT1 goes down.

Since the S,G was deleted, this triggers an accounting **stop** message.

- 4 MLT2 re-distributes the traffic that was on MLT1.

Since a new S,G was created with a different session ID, this triggers an accounting **start** message.

MLT1 is down so both the accounting stop and accounting start messages were sent to the RADIUS server on MLT2. If the accounting stop message is sent before OSPF had time to re-calculate the route change and send an accounting start message, the Passport 8600 drops the UDP packets.



Note: The above scenario will **not** cause an accounting error because RADIUS uses the session ID to calculate accounting time. Even though the route loss and OSPF re-calculation caused the packets to be sent out of sequence, IGAP and RADIUS processes the events in the correct order.

Workaround

To avoid any potential traffic loss in situations where an MLT link has to be brought down, use the following workaround:

- Enable 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 so there is no traffic interruption and accounting does not have to be stopped and re-started.

PIM-SSM and IGMPv3

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 of a 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.

The following sections describe design considerations for implementing PIM-SSM and IGMPv3 in the Passport 8600.

IGMPv3 and PIM-SSM design considerations

Release 3.5 introduced an SSM-only implementation of IGMPv3. It is not a full implementation, and it processes messages according to the following rules:

- When an IGMPv2 report is received on an IGMPv3 interface, the switch drops the IGMPv2 report. There is no backward compatibility.
- In dynamic mode, when an IGMPv3 report is received with several sources (not SSM) but matches a configured SSM range, the switch does not process the report.
- When an IGMPv2 router sends queries on an IGMPv3 interface, the switch downgrades this interface to IGMPv2 (backward compatibility).

This may cause traffic interruption, but the switch will recover quickly.

- When an IGMPv3 report is received for a group with a different source than the one in the SSM channels table, the switch drops the report.

PIM-SSM design considerations

Keep the following considerations in mind when designing an SSM network:

- When SSM is configured, it takes effect for SSM groups only. The switch handles the rest of the groups in sparse mode (SM).
- You can configure PIM-SSM only on switches at the edge of the network while core switches use PIM-SM, if the core switches do not have receivers for SSM groups.
- For existing networks where group addresses are already in use, you can change the SSM range to match the groups to cover with SSM.

- One switch has a single SSM range.

- You can have different SSM ranges on different switches.

You should configure the core switches that relay all the 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 locations in a network may want to receive from a physically closer server for the same group, hence receivers will listen to different channels (still same group).

Chapter 7

Designing secure networks

Networking and security are often completely at odds with one another. True security can only be achieved when information is isolated, locked in a safe environment, surrounded by guards, and rendered inaccessible. In the early days of computing, networks were created to connect isolated computers and allow them to share information. Thus, networks became the communication bridges to connect these isolated machines.

Security and privacy are the antithesis of sharing and distribution (which is the first goal of networking machines). As a result, network security will always be a compromise between providing access and at the same time, safeguarding access to the information. As a network administrator, the most important thing for you to be aware of is that threats and possible new attacks can happen every day. Thus, you must build an effective strategy against these attacks.

This chapter provides descriptions of some of the most common types of attack, and information on Passport 8000 security and protection measures that you can use to guard against them. Specifically, the following sections are included here:

Topic	Page number
Denial of service attacks	next
Malicious code	288
Attacks to resiliency and availability	289
Implementing security measures with the Passport 8600	290

Denial of service attacks

Denial-of-service (DoS) attacks prevent a target server or victim device from performing its normal functions through the use of flooding, irregular sizes of certain types of protocols such as *ping* requests aimed at the *victim* server, application buffer overflows, and many others. Normally, such attacks are launched from a single machine to a specific server. Their purpose is to overload the processor or monopolize the bandwidth for that server so legitimate users cannot use the resource.

Distributed denial of service (DDoS) attacks operate in much the same way, except that they are launched from multiple machines. Most of the DDoS attacks are done through pre-positioned code on the offending machine, also known as a slave, so that the remote or master machine can command the slave to launch the attack at any time.

Malicious code

The most costly threat to networks today is that of malicious code, primarily viruses, worms, and Trojan horses. Viruses alone were the most prolific and costly threat to corporate networks in 2001 with 94% of enterprises reporting virus attacks.

Viruses and worms are programs which replicate throughout files, file systems or networks. In particular, viruses usually replicate by inserting code into an otherwise legitimate and benign program. Viruses perform their functions through a piece of code inserted into the chain of command of a legitimate program so that when the infected program runs, the piece of viral code is executed. Worms do their damage by typically attacking operating systems and networks, rather than individual files or objects.

Trojans are somewhat different than viruses and worms in that they have no replication function. They are programs that appear to and may perform a legitimate, desirable function. However, they also perform a function that the person running the program does not expect or desire. An example of a Trojan may be one in which you may download a software program, such as a browser. The IDS-SLB switch copies all incoming packets to this group of intrusion detection servers. For each session entry created on the switch, an IDS server is

selected based on the IDS server load-balancing metric. The IDS server receives copies of all processed frames that are forwarded to the distribution devices. Session entries are maintained so that all frames of a given session are forwarded to the same IDS server.

You must connect each IDS server directly to a different switch port or VLAN because you cannot substitute any fields in the packet header. Substituting a field corrupts the packet that must also be forwarded to its real destination.

Nortel Networks provides other equipment, like the Contivity VPN product suite, the Shasta 5000 BSN, and the Alteon Switched Firewall System. They offer differing levels of protection against DoS and DDoS through either third party IDS partners, or through their own high performance, state-aware firewalls.

Attacks to resiliency and availability

In the event that links become congested due to attacks, you can immediately halt end user services. During the design phase, you should study availability very carefully, for each layer, from the physical to the upper layers. See [Chapter 2, “Designing redundant networks,” on page 53](#) for more information. Without redundancy, all services can be brought down very easily.

Additional information and references

The following organizations provide you with the latest, most updated information about network security attacks and some recommendations about good practices to follow:

- The Center of Internet Security Expertise (CERT)
Their web site is http://www.cert.org/nav/index_main.html
- The research and education organization for network administrators and security professionals (SANS)
Their web site at <http://www.sans.org> features security alerts, news items, research materials, white papers and educational and certification opportunities.
- The Computer Security Institute (CSI)

Their web site at <http://www.gocsi.com> provides a very good CSI computer crime and security survey.

Implementing security measures with the Passport 8600

This section explains how to use the Passport 8600 to guard against those common network attacks (DoS, DDoS, malicious code etc.) described previously in this chapter.

Passport 8600 DoS protection mechanisms

The Passport 8600 is protected against DoS attacks by internal mechanisms and some specific features, such as:

- Broadcast/multicast rate limiting
- Directed broadcast suppression
- Prioritization of control traffic
- Control traffic limitation (CP limit)
- ARP limitation (ARP request limit)
- Multicast learning limitation

Each of these are explained in the subsections that follow.

Broadcast/Multicast rate limiting

To protect the Passport 8600 switch and other stations from a high number of broadcasts, the switch has the ability to limit the broadcast/multicast rate. You can configure broadcast/multicast rate limiting on a per-port basis. This feature was introduced in release 3.0. See *Configuring Network Management* for more information on setting the rate limits for broadcast or multicast packets on a port.

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.

By default, this feature is enabled on the Passport 8000 Series switches. This feature has been available since release 3.1.0. See *Configuring and Managing Security* for more information.

Prioritization of control traffic

A very sophisticated prioritization scheme has been implemented on the Passport 8600 to schedule received control packets (BPDU, OSPF Hellos etc.) on physical ports. This scheme involves two levels with both hardware and software queues and guarantees proper handling of these control packets regardless of the load on the switch. In turn, this guarantees the stability of the network. More specifically, it guarantees that the applications that heavily use broadcasts (typically IPX) are handled with a lower priority.

Note that you cannot use the CLI to view, configure, or modify these queues. Setting the queues and determining the type of packets entering each queue is Nortel Networks confidential. For more information, please contact your Nortel Networks representative.

Control traffic limitation

This feature is also known as CP limit and it prevents the CPU from being overloaded by excessive multicast or broadcast control or exception traffic. For example, traffic generated by a network loop introducing broadcast storms in a network will not impact the stability of the system. By default, it protects the CPU from receiving more than 14,000 broadcast/multicast control or exception packets per second within a duration exceeding 2 seconds.

You can disable CP limit and configure the amount of broadcast and/or multicast control or exception frames per second 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 [Table 22](#).

Table 22 Recommended CP limit values

Severe:		
	Broadcast	Multicast
Workstation (PC)	1000	1000
Server	2500	2500
Non-IST Interconnection	7500	6000
Moderate:		
Workstation (PC)	2500	2500
Server	5000	5000
Non-IST Interconnection	9000	9000
Relaxed:		
Workstation (PC)	4000	4000
Server	7000	7000
Non-IST Interconnection	10000	10000

This feature was introduced in the 3.2.2 release. See *Configuring and Managing Security* for more information.

ARP limitation

The ARP request threshold limits the ability of the Passport 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 help customers experiencing excessive amounts of subnet scanning caused by some virus (like Welchia), it is recommended that you change the ARP request threshold to a value between 100 to 50. This will help protect the CPU from causing excessive ARP requests, help protect the network, and lessen the spread of the virus to other PCs.

- Default: 500
- Severe Conditions: 50
- Continuous scanning conditions: 100
- Moderate: 200
- Relaxed: 500

Between release 3.2.2.2 and 3.5.0.0 software, you can access this feature only through VxWorks shell. Within the shell, the designation is *arq_threshold*. Contact a Nortel Networks support engineer for more information. From the 3.5.0 release onwards, you can access the feature through the CLI. See *Configuring IP Routing Operations* for more information on the **config ip arp arpreqthreshold** command.

Multicast learning limitation

This feature protects the CPU from multicast data packet bursts generated by malicious applications such as viruses. Specifically, it protects against those viruses which cause the CPU to reach 100% utilization or which prevent the CPU from processing any protocol packets or management requests. If more than a certain number of multicast streams enter the CPU through a port during a sampling interval, the port will be shutdown until the user or administrator takes appropriate action. See the *Release Notes for the Passport 8000 Series Switch Release 3.5.1* for more information and detailed configuration instructions.

Passport 8600 damage prevention mechanisms

To further reduce the chances of your network being used to damage other existing networks, take the following actions:

- 1 Stop spoofed IP packets
- 2 Prevent your network from being used as a broadcast amplification site
- 3 Enable the `hsecure` flag (bootconfig) and High Secure mode to block illegal addresses. For more information, refer to [“High secure mode \(CLI\)” on page 295](#).

Stopping spoofed IP packets

You stop spoofed IP packets by configuring the Passport 8600 to ensure that only IP packets are forwarded that contain the correct source IP address of your network.

A spoofed packet is one which 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. The basic idea of anti-spoofing protection is for you to have a filter rule/configuration assigned to the external interface, which examines the source address of all outside packets crossing that interface. If that address belongs to internal network or firewall itself, the packet should be dropped.

The correct source IP address(es) consist of the IP network addresses that have been assigned to the site/domain. It is particularly important that you do this throughout your network, especially at the external connections to the existing Internet/upstream provider. By denying all invalid source IP addresses, you minimize the chances that your network will be the source of a spoofed DoS attack.

This will not prevent DDoS attacks coming from your network with valid source addresses, however. In order to implement this, you need to know the IP network blocks that are in use. You then create a generic filter which:

- permits your sites' valid source addresses
- denies all other source addresses

To do so, configure the filters as ingress filters that deny (drop) 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. [Table 23](#) lists the source addresses that you should filter.

Table 23 Source addresses that need to be filtered

Address	Description
0.0.0.0/8	Historical Broadcast ¹
10.0.0.0/8	RFC1918 Private Network

Table 23 Source addresses that need to be filtered

Address	Description
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	Broadcast ¹

¹ 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. For more information, see the [“High secure mode \(CLI\)”](#) section below.

Preventing the network from being used as a broadcast amplification site

To prevent the flooding of other networks with DoS attacks, such as the “smurf” attack, the 8600 Series switch is protected by directed broadcast suppression. This feature is enabled by default. You should *not* disable it.

High secure mode (CLI)

To protect the Passport 8600 against IP packets with an illegal source address of 255.255.255.255 from being routed (per RFC 1812 Section 4.2.2.11 and RFC 971 Section 3.2), the Passport 8600 supports a configurable flag, called *high secure*.

By default, this flag is disabled. Note that when you enable this flag, the desired behavior (not routing source packets with an IP address of 255.255.255.255) is applied on all ports belonging to the same OctaPid (group of 8 10/100 ports (8648), 1 Gig port (8608) or 2 Gig ports (8616)). (See [Appendix D, “Tap and OctaPID assignment,”](#) on page 403).

The command syntax is:

```
config ethernet (slot/port) high-secure [true/false]
```

This feature was introduced in the 3.5 release. For more information, see **XXX**.

Passport 8600 security against malicious code

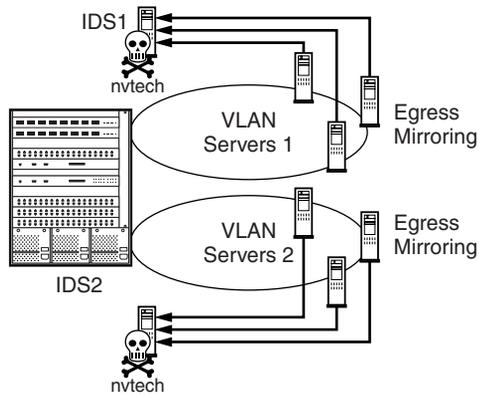
An intrusion detection system (IDS) exists to warn you about the possibility of a security incident. They form yet another defensive tool in a layered security solution. You can broadly categorize them as follows:

- Incident detection timeframe: real-time or off-line
- Type of installation: network-based or host-based
- Mechanism used for detection: signatures or usage patterns
- Reaction to incidents: whether the IDS actively intervenes to head off attacks, or simply notifies staff or other network elements of the problem.

You can then use these categories to separate IDS products into:

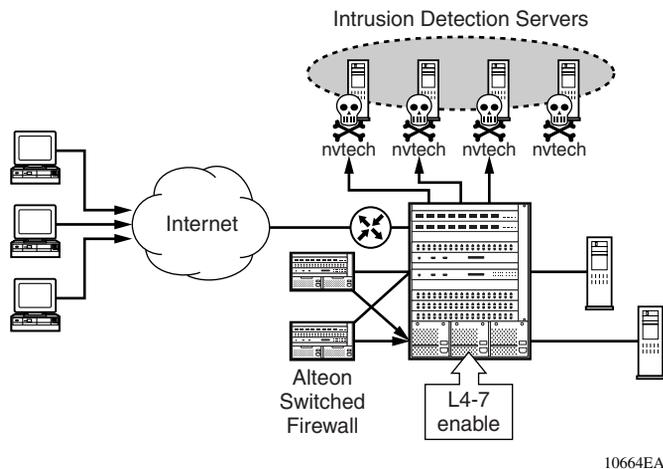
- more easily implemented features, such as off-line detection, network-based installation, signature-based detection, and notification of staff for incidents
- more advanced features, such as real-time detection, host-based installation, usage pattern monitoring, and active intervention for incidents

The Passport 8600 does not currently provide any IDS service. However, you can connect the Passport 8600 to external IDS servers. By using either mirroring, or remote mirroring, as introduced in 3.7, you can utilize external IDS servers to analyze activities, recognize typical patterns of attacks, and analyze abnormal activity patterns ([Figure 98](#)). For more information on remote mirroring, see [page 390](#).

Figure 98 IDS server configuration

Note that some rules apply to egress mirroring. See the *Release Notes for the Passport 8000 Series Switch Release 3.5* for more information.

The Alteon web switch family, including the WSM module available on the Passport 8600 provides load balancing of IDS products which you can leverage to further protect your network ([Figure 99](#)).

Figure 99 Alteon web switch family IDS server configuration

10664EA

With the WebOS (WSM module), the Passport switch forwards the IP packets to an external IDS server at the end of the filtering process, or at the end of client processing (when filtering is not enabled). You must enable IDS server load balancing (SLB) on the port and allocate a real server group for IDS SLB.

The IDS-SLB switch copies all incoming packets to this group of intrusion detection servers. For each session entry created on the switch, an IDS server is selected based on the IDS server load-balancing metric. The IDS server receives copies of all processed frames that are forwarded to the distribution devices. Session entries are maintained so that all frames of a given session are forwarded to the same IDS server.

You must connect each IDS server directly to a different switch port or VLAN because you cannot substitute any fields in the packet header. Substituting a field corrupts the packet that must also be forwarded to its real destination.

Nortel Networks provides other equipment, like the Contivity VPN product suite, the Shasta 5000 BSN, and the Alteon Switched Firewall System. They offer differing levels of protection against DoS and DDoS through either third party IDS partners, or through their own high performance, state-aware firewalls.

Passport 8600 security against resiliency and availability attacks

Redundancy in hardware and software is one of the key feature of the Passport 8600. High availability is achieved by eliminating single points of failure in the network, and using the unique features of the Passport 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/CPU, 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)

Passport 8600 access protection mechanisms

When implementing access protection measures in the Passport 8600, there are three primary areas that you should focus on:

- The data plane
- The control plane
- Other platforms and equipment

For descriptions of these, see the sections that follow.

Data plane

The data plane includes the extended authentication protocol, 802.1x, traffic isolation: VLANs, filtering capabilities, routing policies (announce/accept policies) and routing protocol protection. Each of these is described in the sections that follow.

Extended authentication protocol- 802.1x

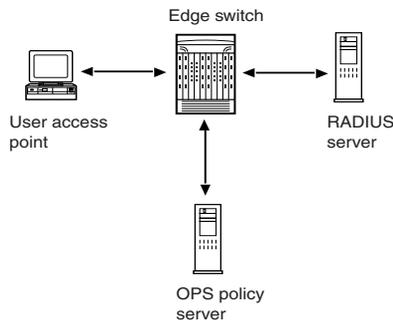
To protect the network from inside threats, the Passport 8600 3.7 software supports the 802.1x standard which separates user authentication from device authentication. In this case, the switch itself requires that end-users securely log into the network before obtaining access to any resource.

Interaction between 802.1x and Optivity Policy Server v4.0

User-based networking goes one step beyond the 802.1x EAP authentication in the network to ensure that users have access to only authorized services. It then links that authorization to individual user-based security policies based on individual policies. As a result, network managers can define corporate policies configured on every port and add another level of precision based on a login and password.

Nortel Networks OPS (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.

[Figure 100](#) shows the interaction between 802.1x and OPS v4.0. First, the user initiates a login from a user access point and receives a request/identify request from the switch (EAP access point). The user is then presented with a network login. Prior to DHCP, the user has no network access since the EAP access point port is in EAP blocking mode. The user then provides User/Password credentials to the EAP access point via EAPoL. Note that the client PC is considered both a RADIUS peer user and an EAP supplicant.

Figure 100 802.1x and OPS interaction

For a specific example of this interaction, consider the following:

1 User *Joe* attempts a login.

The EAP access point initiates RADIUS authentication to the RADIUS server and presents user credentials to the authentication server.

2 The Userid and password for *Joe* are passed to the RADIUS server.

The RADIUS server examines the directory service via LDAP to determine if user/password combination is valid and returns the user role attribute.

3 The RADIUS server looks up *Joe* based on the userid/password.

4 If *Joe* is found:

- The role associated with *Joe* is retrieved from the RADIUS server data store and passed to the edge switch.
- If *Joe* is found but has no role, the authentication with no role result is passed to the edge switch.

In both instances, *Joe* is logged on and the edge switch allows him to access the network.

5 If *Joe* is not authorized, he is notified that he is not found and access is denied.

6 If *Joe* is authorized, authentication pass/fail and role data are sent to the switch.

The switch opens the port with basic access enabled and simultaneously triggers a communication with OPS to provision attributes of the port (based on user role/policy combinations). These attributes remain valid during the lifetime of that *session*. (Note that session is tightly defined at the MAC layer in 802.1X and is not the application session).

- 7 The edge switch communicates with OPS by sending the role and port details for *Joe* to the policy server.

In the *no role* case, *Joe* will have default corporate policies applied. The OPS downloads the user-specific policy to the port. The policy server applies the role for *Joe* to the device/port combination.

The implemented process for 802.1x and OPS v4.0 includes support for a limited number of RADIUS servers; Preside (Funk) and Microsoft IAS. Additional RADIUS servers supporting the EAP standard should also work here.

LAN Enforcer

In environments where there is less control over where and how the end-user device is located and used, processes to check for viruses, worms and other threats on devices connected to the corporate environment are more critical than ever before. This is particularly true for users working from home and using their Internet connection. Such processes help protect the corporation from uneducated users, such as mobile employees and business partners, who may not even be aware of the extent of such threats. The enforcement mechanism here checks for the latest anti-virus, firewall definitions, or software patches before allowing authorized access to the network.

Nortel Networks LAN Enforcer enables the Passport 8600 to use the 802.1x standard to ensure that someone connecting inside the corporate network is in fact a legitimate user. The solution then goes one step further by verifying and checking the endpoint security posture, including anti-virus, firewall definitions, Windows registry content, and specific file content (plus date and size). Non-compliant systems attempting to obtain switch authentication may 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.

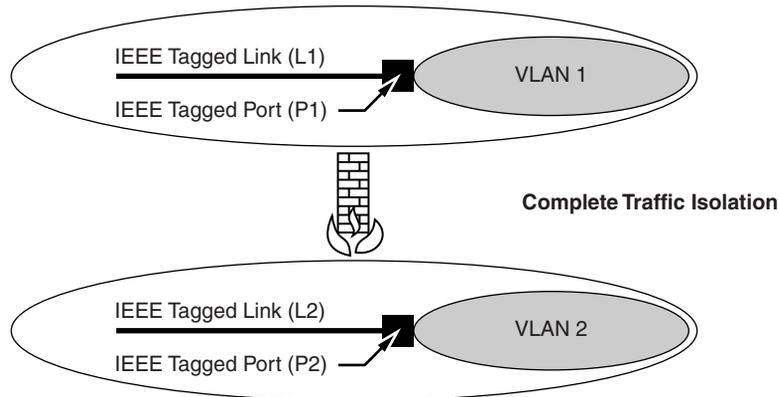
Traffic isolation: VLANs

The internal architecture of the Passport 8600 lets you build secure VLANs. When you configure port-based VLANs, each VLAN is completely separated from each other (broadcast domain).

By its unique hardware architecture (distributed ASICs with local forwarding decision), each packet is analyzed independently of the preceding one. This mode, as opposed to the cache mode that some Passport competitors use, allows complete traffic isolation.

By allowing you to discard untagged traffic on tagged port or tagged traffic on untagged ports, you have this guarantee. Even if a tagged port receives traffic with a VLAN Id which identifies a VLAN from another customer configured on the switch, this traffic is completely discarded (Figure 101).

Figure 101 Traffic discard process



10665EA

Filtering capabilities

A brief description of the Passport 8600 filtering capabilities follows. For more details on filtering, see Chapter 9, “Provisioning QoS networks,” on page 341.

Layer 2

At Layer 2, the Passport 8600 provides the following security mechanisms:

- Filtering

The Passport 8600 provides L2 filtering based on the MAC destination (available since Passport release 3.0), and the MAC source (available since release 3.5.0).

- Unknown MAC Discard

This allows the customer to secure the network by learning allowed MAC addresses during a certain time and then lock these MACs in the FDB. After the learning has occurred, the switch does not accept any new MAC addresses on this specific port.

With the Passport 3.5.1 release, it is possible to globally filter a MAC address. Regular filters work at the VLAN level, while the Passport 8600 has an FDB per VLAN and not per switch. For more information and configuration examples, see the *Release Notes for the Passport 8000 Series Switch Release 3.5.1*.

With the 3.5.2 release, the following two features are available:

- Limited MAC learning

This feature limits the number of FDB-entries learned on a particular port to a user-specified value. Once 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 now 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 Passport 8000 Series Switch Release 3.5.2*.

- Global MAC filtering

This feature eliminates the need for you to configure multiple per-VLAN filter records for the same MAC address. It provides you with the ability to discard ingress MAC addresses that match a global list stored in the switch.

By using a global list, you do not need to configure a MAC filter per VLAN. You can also apply global MAC filtering to any multicast MAC address. However, you cannot apply it to Local, Broadcast, BPDU MAC, TDP MAC, and All-Zeroes MAC addresses. Once a MAC address has been added to this Global list, it cannot be configured statically or learned on any VLAN. In addition, no bridging or routing will be performed on packets to or from this MAC address on any VLAN.

For more information and configuration examples, see the *Release Notes for the Passport 8000 Series Switch Release 3.5.2*. For more information on the Layer 2 MAC filtering process, see *Configuring IP Multicast Routing Operations*.

Layer 3 and 4 filters

At Layer 3 and above, the Passport 8600 provides advanced filtering capabilities as part of its security strategy to protect the network from different attacks.

You can configure two types of filters on the Passport 8600:

- global filters
- source/dest filters



Note: These filters are *always* executed in the hardware on the ASICs, and do not impact the CPU in any way- except during configuration loading and modification.

The matching criteria for filters in Passport 8600 can be any of the following:

- Destination address
- Source address
- Exact IP protocol match (TCP, UDP or ICMP)
- TCP or UDP port numbers
- Established TCP connections (from within the network or bi-directionally)
- ICMP request
- DS (DiffServ) field
- IP frame fragment

You can perform the configuration actions listed in [Table 24](#):

Table 24 Configuration actions

Action	Description
Drop	Discards the traffic.
Forward	Forwards the traffic.
Forward to Next Hop	Forwards traffic to the next hop address
Mirror	Copies the traffic to another port
Police	Enforces a Service Level Agreement (SLA) at ingress
TCP connect	Prevents incoming TCP sessions
Stop on match	Stops the filtering process if the condition matches
Modify the DS field	Remarks the traffic's CoS at L3 (IP) using the DiffServ field
Modify the IEEE 802.1p bit	Remarks the traffic's CoS at L2 (Ethernet) using the .p bits as part of 802.1q

Filter action modes

Each filter has an associated action mode which determines whether packets matching this filter are forwarded or routed through the switch. Each filtered port on the Passport switch has an associated default action of forward or drop. When the filtering action mode matches the port default action, the default action is performed.

When the port default action is drop, a packet is forwarded only if a matching filter is set with a forward action mode. If a single match occurs with a forward action mode, it does not matter how many matching filters are found with a drop action mode. The frame is forwarded. Thus, if a packet matches multiple filters and any one of them has a forward action mode, the packet is forwarded.

When the port mode is set to forward, a packet is dropped only if a matching filter is found with a drop action mode. Again, if a single match occurs with a drop action, it does not matter how many matching filters have forwarding actions. The packet is dropped. Thus, if a packet matches multiple filters and any one of them has a drop action mode, the packet is dropped.

Customer Support Bulletins (CSBs) available on the Nortel Networks web site provide information and configuration examples on how to block some attacks. Go to the www.nortelnetworks.com/documentation URL for more information or contact your local Nortel Networks representative.

Routing policies (announce/accept policies)

You can use route policies to selectively accept/announce some networks, and to block the propagation of some routes. This is one tool you can use to enhance the security in a network, by *hiding* the visibility of some networks (subnets) to other parts of the network.

In previous releases, separate policy databases for RIP accept, RIP announce, OSPF accept, and OSPF announce filtering purposes were used. With release 3.2, a unified database of route policies is available for RIP or OSPF to use for any type of filtering task.

You identify a policy by name or ID. Under a given policy, you can define several sequence numbers, each of which is equal to one policy in the old convention. Each policy sequence number contains a set of fields. Only a subset of those fields are used when the policy is applied in a certain context. For example, if a policy has a set-preference field set, it is used only when the policy is applied for accept purposes. This field is then ignored when the policy is applied for announce/redistribution purposes.

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 (i.e., a lower sequence number is preferred).

OSPF

You can protect the OSPF updates with an MD5 key on each interface, according to RFC 2178 (OSPF cryptographic authentication with the MD5 algorithm). See [Table 25](#) for details.

Table 25 OSPF packet

Version	Type	Packet Length
	Router ID	
	Area ID	
	Checksum	Authentication Type
	Authentication	
	Authentication	

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.

BGP

RFC 2385 describes how to protect BGP sessions via the TCP MD5 Signature option. This option allows BGP to protect itself against the introduction of spoofed TCP segments in the connection stream. Every segment sent on the TCP connection is protected against spoofing by a 16-byte MD5 digest.

The MD5 algorithm is applied to these items in the following order:

- the TCP pseudo-header (in the order: source IP address, destination IP address, zero-padded protocol number, and segment length)
- the TCP header, excluding options, and assuming a checksum of zero
- the TCP segment data (if any)
- an independently-specified key or password, known to both TCPs and connection specific

As per RFC 2358, the format is as follows:

```

+-----+-----+-----+
| Kind=19 |Length=18| MD5 digest... |
+-----+-----+-----+
|           |           |           |
+-----+-----+-----+
|           |           |           |
+-----+-----+-----+
|           |           |           |
+-----+-----+-----+
|           |           |           |
+-----+-----+-----+

```

Note that the MD5 digest is always 16 bytes in length. In addition, tests have been successfully conducted to verify interoperability with other vendors. Refer to *Configuring BGP Services* for more information and configuration examples.

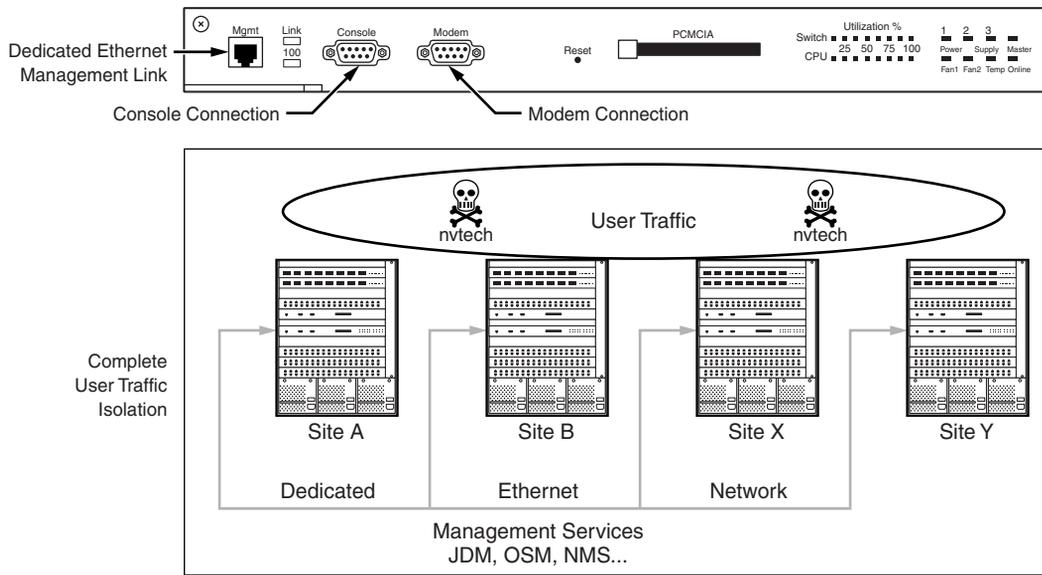
Control plane

The control plane physically separates management traffic using the out of band (OOB) interface. It includes management, high secure mode (bootconfig), management access control, access policies, authentication, secure shell and secure copy, and SNMP, each of which is described in the sections that follow.

Management

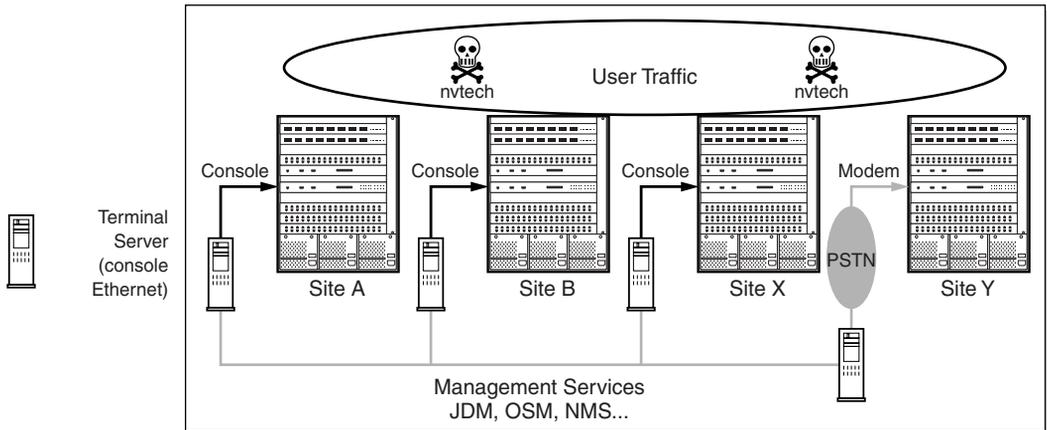
Both the Passport 8100 and 8600 provide an isolated management port on the SSF (switching fabric/CPU). They do so in order to completely separate user traffic from management traffic in highly sensitive environments, such as brokerages and insurance agencies. By using this dedicated network to manage the switches, and by configuring access policies (when routing is enabled (See [“Access policies” on page 314](#))) you can be sure that you will always be able to manage the switch in a secure fashion. See [Figure 102](#).

Figure 102 Dedicated Ethernet management link



You can also use the terminal servers/modems to access the console/modems ports on the switch (Figure 103).

Figure 103 Terminal servers/modem access



10667EA

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

High secure mode (bootconfig)

This flag in bootconfig mode allows network managers to disable all unsecured application and daemons, such as FTP, TFTP, rlogin etc. It is strongly recommended that you *not* use any unsecured protocols, with the exception of Secure Shell (SSH). Note here that SSHv2 is recognized as being more secure than SSHv1.

You should also plan to use Secure Copy (SCP), rather than FTP or TFTP. For more information, see “[SSSHv1/v2 and SCP](#)” on page 318.

Management access control

The 8600 Series switch has the following levels of management access ([Table 26](#)):

Table 26 8600 Series switch management access levels

Access level	Description
Read only	This level lets you view the device settings, but you cannot change any of the settings.
Layer 1 Read Write	This level lets you view switch configuration and status information and change only physical port parameters.
Layer 2 Read Write	This level lets you 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	This level lets you view and edit device settings related to Layer 2 (bridging) and Layer 3 (routing). You cannot change the security and password settings.
Read Write	This level lets you view and edit most device settings. You cannot change the security and password settings.

Table 26 8600 Series switch management access levels (continued)

Access level	Description
Read Write All	<p>This level lets you 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.</p> <p>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. For information about community string encryption, see “SNMP community string encryption” on page 320.</p>
ssladmin	<p>This level lets you login to connect to and configure the SAM (ssl acceleration module).</p> <p>Caution: The ssladmin users are granted a broad range of rights that incorporate the 8600 read/write access. Users with ssladmin access can also add, delete, or modify all 8600 configurations and the WSM software image and configuration. For more information, see <i>Configuring the SSL acceleration module</i>.</p>
<p>Note: On the 8600 switch, the following access levels are equivalent to read-only access. Use these logins only if you intend to connect to the WSM. These levels have been added to the 8600 CLI to provide the granularity required for mapping 8600 to WSM access levels. Each level provides a different level of access to the Web OS command line interface.</p>	
User	<p>This level gives you no direct responsibility for switch management. The User can view all switch state information and statistics, but cannot make any configuration changes to the switch.</p>
SLB Operator	<p>This level lets you manage Web servers and other Internet services and their loads. In addition to being able to view all switch information and statistics, the SLB Operator can enable/disable servers using the Server Load Balancing operation menu.</p>
Layer 4 Operator	<p>This level lets your manage traffic on the lines leading to the shared Internet services. This user currently has the same access level as the SLB operator.</p>
Operator	<p>This level lets you manage all functions of the switch. In addition to SLB Operator functions, the Operator can reset ports or the entire switch.</p>

Table 26 8600 Series switch management access levels (continued)

Access level	Description
SLB Administrator	This level lets you configure and manage Web servers and other Internet services and their loads. In addition to SLB Operator functions, the SLB Administrator can configure parameters on the Server Load Balancing menus, with the exception of not being able to configure filters or bandwidth management.
Layer 4 Administrator	This level lets you configure and manage traffic on the lines leading to the shared Internet services. In addition to SLB Administrator functions, the Layer 4 Administrator can configure all parameters on the Server Load Balancing menus, including filters and bandwidth management.
Administrator	This level gives you complete access to all menus, information, and configuration commands on the WSM, including the ability to change both the user and administrator passwords.

The following are some rules you should follow in choosing passwords:

- Do not use your first, middle, or last name in any form
- Do not use other information easily obtained about you (for example, license plate numbers, telephone numbers, social security numbers, etc.)
- Do not use a password of all digits, or all the same letter
- Do not use a word contained in English or foreign language dictionaries, spelling lists, or other lists of words
- Do not use a password shorter than six characters



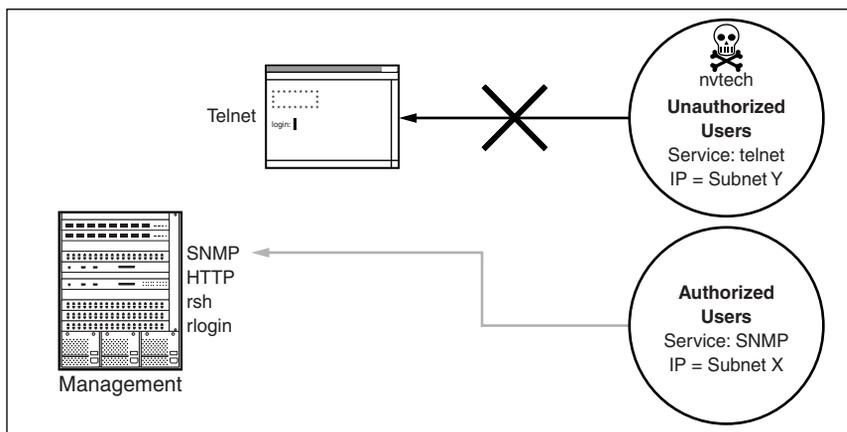
Note: Using the `hsecure` flag, the Passport 8000 enforces an eight-character length password. Nortel Networks recommends that you configure this flag. For more information, refer to the *Release Notes for the Passport 8000 Series Switch Software Release 3.5*.

- Do use a password with mixed-case alphabets, and with non-alphabetic characters (for example, F8Rt34X6)
- Do use a password that is easy to remember, so you do not have to write it down.

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 which is spread out among all of the switches in the network, this feature allows you to build a full, L3 routed network and securely manage the switch with any of the in-band IP addresses attached to any one of the VLANs (Figure 104). For more information on access policies, see XXX.

Figure 104 Access levels



10668EA

It is *highly* recommended that you use access policies for in-band management when securing access to the switch. By default, all services are accessible by all networks.

Authentication

Authentication involves RADIUS and its enhancements for release 3.3 and 3.5, each of which is described in more detail in the sections that follow.

RADIUS

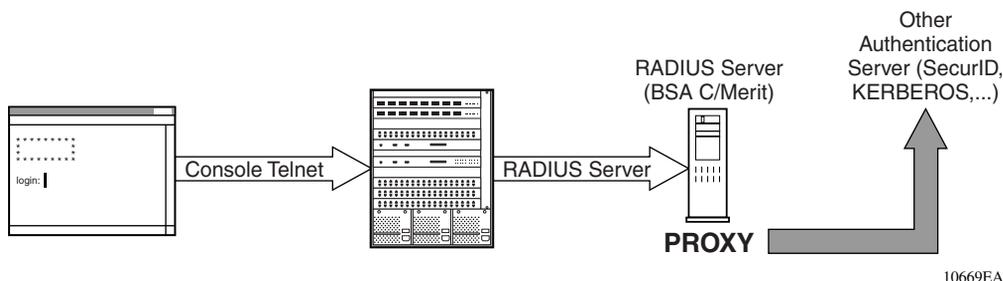
You can enforce access control by utilizing RADIUS (Remote Authentication Dial-in User Service). RADIUS is designed to provide a higher degree of security against unauthorized access to Passport 8000 switches, and to centralize the knowledge of the security access, based on a client/server architecture. The database within the RADIUS server stores list of pertinent information about client information, user information, password, and access privileges including the use of the *shared secret*.

Acting as a Network Access Server, the Passport 8000 switch operates as a client of RADIUS. The switch is then responsible for passing user information to the designated RADIUS servers. Since the Passport switch operates in a LAN environment, it allows user access via Telnet, Rlogin, and Console log-in.

The supported RADIUS servers are Nortel Networks' BaySecure Access Control (BSAC) Version 2.2 and MERIT Networks' RADIUS server. You can configure a list of 10 RADIUS servers on the client. If the first server is unavailable, the Passport 8600 tries the second and so on until it establishes a successful connection.

You can use the RADIUS server as a proxy for stronger authentication, such as:

- SecurID cards
 - KERBEROS
- or
- other systems like TACACS/TACACS+ ([Figure 105](#))

Figure 105 RADIUS server as proxy for stronger authentication

10669EA

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 the RADIUS feature
- Provide the IP address of the RADIUS server that is to be used
- 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 will be used. (This is essential when there is more than one RADIUS server in the network).
- Specify the UDP port that will be 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.

Note that there are other customizable parameters in the switch RADIUS configuration that require careful planning and consideration on your part. These include such things as the 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 will accept before sending an authentication request failure.

It is strongly recommended that you use the default value in the attribute-identifier field. The value on the attribute identifier field ranges from 192 to 240, with the default set to 192. If you change the set default value, you must alter the dictionary on the RADIUS server with the new value. The dictionary is set to an attribute value of 192, thus matching the default attribute value on the switch. When configuring the RADIUS feature, you need to configure Read-Write-All access to the switch.

RADIUS enhancements for release 3.3

With release 3.3, two major enhancements have been developed for RADIUS support.

- 1** RADIUS authentication now supports the RADIUS authentication challenge message as described in the RFC. This allows more effective interoperability with some RADIUS servers. Release 3.3 has modified authentication to be performed in two steps:
 - a** If you have RADIUS configured, the switch attempts the RADIUS authentication method. If RADIUS authentication fails (Auth Reject), local authentication is not performed. If the first RADIUS server is not reachable, you can configure up to 10 servers, and the switch tries authentication on these 10 servers (if configured).
 - b** If all of the servers that you have configured are unreachable (IP connectivity), then local authentication occurs.
- 2** RADIUS accounting support is now provided in release 3.3. This feature allows network managers to track every management session. It provides information not only on the packets, bytes, and duration of the session, but all the CLI commands executed during the session.

RADIUS enhancements for release 3.5

Release 3.5 introduces the following RADIUS features that enable you to:

- 1** Define which commands are accessible by each login level (CLI/RADIUS). This allows a network administrator to “tune” which feature can be configured by each user.
- 2** Do some accounting for SNMP (SNMP/RADIUS accounting). This gives a network administrator a view of who is connected and the duration of the pseudo SNMP session.

Encryption of control plane traffic

Control plane traffic involves SSHv1/v2, SCP, and SNMPv3. Each of these is detailed in the sections that follow.

SSSHv1/v2 and SCP

SSH is a protocol used to conduct secure communications over a network between a server and a client. The Passport 8600 switch supports *only* the server mode. (You need an external client to establish communication). If 3 versions of SSH exist, the Passport switch supports version 1 and version 2. Note that these versions are *not* compatible. Nortel Networks recommends that you configure the switch in version 2 *only* since Version 1 has some well-known security issues related *not* to the Passport implementation, but to the protocol (SSH version 1) itself.

The SSH protocol covers many security aspects including:

- Authentication

SSH determines in a reliable way someone's identity. During the login process, the SSH client asks for a digital proof of this identity.

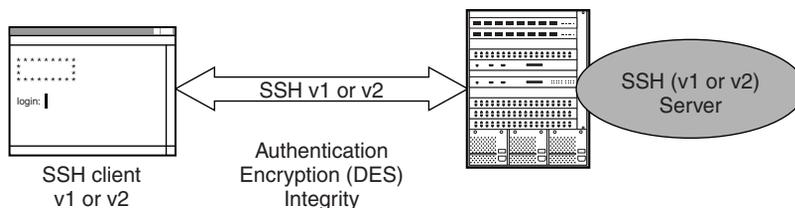
- Encryption

SSH uses some encryption algorithms to scramble data. This data is rendered unintelligible except to the receiver.

- Integrity

This guarantees that the data is transmitted from the sender to the receiver without any alteration. If any third party captures and modifies the traffic, SSH will detect this alteration (Figure 106).

Figure 106 Authentication encryption



10670EA

The Passport 8000 switch acts as a server (the client mode is not currently supported), and secures the communication between a client (PC, UNIX) and the switch. It supports:

- SSH version 1, with password and RSA authentication

- SSH version 2 with password and DSA authentication
- 3DES to encrypt the data



Caution: Due to export restrictions, the encryption module is separated from the code itself (loadable module), and can be downloaded with a special procedure. See the *Release Notes for the Passport 8000 Series Switch Software Release 3.5* for more information.

Modifying the RADIUS/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. For more information, refer to *Configuring and Managing Security*.

SNMPv3 support in release 3.3 and 3.7

SNMP is listed as one of the top ten security holes by the SANS institute. (For more detail, see <http://www.sans.org/topten.pdf> and the section titled “[Additional information and references](#)” on page 289). SNMP version 1 and version 2 are open to hackers since communities are NOT encrypted.

With the support of SNMPv3 in release 3.3, Nortel Networks strongly recommends that you use SNMP version 3 (SNMPv3) as defined by RFC 2571 through RFC 2576. SNMPv3 allows stronger authentication and the encryption of data traffic for network management. For instructions on configuring SNMPv3, see *Configuring and Managing Security* in the Passport 8000 Series release 3.3 documentation set.



Note: With release 3.7, be aware that Nortel Networks has introduced a revised version of the SNMPv3 agent. It is strongly recommended that you refer to the *Release Notes for the Passport 8000 Series Switch Software Release 3.7* for instructions on upgrading the SNMPv3 agent from a previous release. Setting community strings and trap receivers now uses a different syntax and in some cases, requires special handling.

SNMP community string encryption

With release 3.5, SNMP community strings are encrypted, stored in a hidden file, and are not displayed on the switch. They are no longer stored in the configuration file.

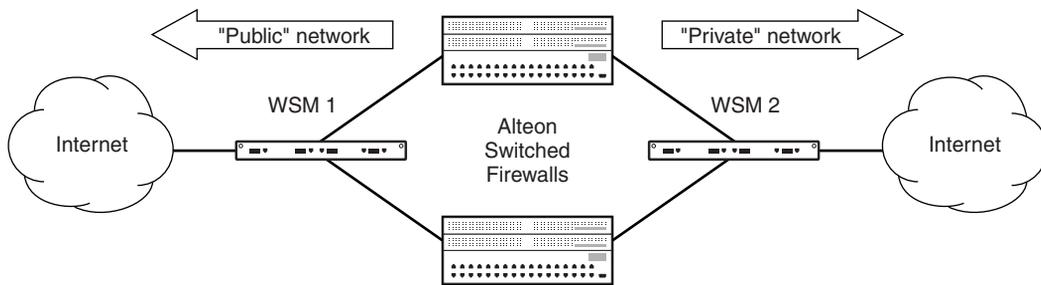
Other platforms and equipment

The Passport 8600 is one of the elements in the security architecture as defined by Nortel Networks. As mentioned previously, Nortel offers other equipment that lets you increase the security of your network. The following section discusses two of these elements: firewalls and VPN service (Contivity).

Firewalls

For sophisticated state-aware packet filtering (Real Stateful Inspection), you can add an external firewall to the architecture. Nortel (Alteon, Contivity, and Shasta) provides the newest generation in firewall technology. 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, audio and video streaming. For every packet, the state-aware firewall finds a matching flow and conversation.

This is the typical configuration used in firewall load balancing ([Figure 107](#)).

Figure 107 Firewall load balancing configuration

10671EA

This configuration enables you to redirect incoming and outgoing traffic to a group of firewalls (Alteon provides a firewall), and automatic load balance across multiple firewalls. The WSM can also filter packets at ingress port, so that firewalls see only relevant packets. The benefits of such a configuration are:

- Increase firewall performance
- Reduce response time
- Redundant firewalls ensures Internet access

On the WSM WebOS, the concept of firewall load balancing is an extension of application redirection where all IP is redirected. It uses hash metric, ICMP or HTTP for health checks. You can also use static routes for health check. With firewall load balancing, both IP source and destination addresses are used in the hashing algorithm, to maintain session persistence and symmetry.

Virtual Private Networks

Virtual private networks (VPN) replace the physical connection between the remote client and access server with a logical connection- a tunnel over a public network. Included here are elements such as:

- IPSEC

This is the security suite of protocols designed by the IETF to provide full security services to IP datagrams. The IPSEC security suite provides standardized cryptographic security mechanisms for authentication, confidentiality, data integrity, anti-replay protection, and protection against traffic flow analysis.

IPSEC is an optional overlay for IP version 4, and a mandatory component of IP version 6. IPSEC provides cryptographic protection at the network layer (Layer 3), and hence secures all higher layer applications without having to modify the applications themselves in any manner.

- Secure socket layer (SSL)

SSL technology has been widely used as a method for protecting web (HTTP) communications. SSL was originally developed by Netscape Communications Corporation, so SSL is built into most browsers and web servers. SSL is a layer residing between the TCP/IP protocol and the application layer. It is a set of protocols built on top of TCP for sending encrypted information over the Internet.

SSL provides data encryption, server authentication, message integrity, and optional client authentication for a TCP/IP connection at the transport layer (Layer 4). The most recent version of SSL is version 3. Transport Layer Security (TLS) is the IETF standards-based successor to SSL and is currently at version 1. TLS has added some enhancements to the SSL protocol to make it more cryptographically secure.

At another level, secure communications are needed for the web traffic in use among business partners and retailers and customers. The use of the web as a tool for retailers and companies for e-business transactions has led to the use of SSL to secure web traffic. With the growth in online retail and e-business transactions among partners, both the SSL traffic and the SSL acceleration market are primed for dramatic growth.

Nortel Networks' IPSEC and SSL solutions

Several Nortel products support IPSEC and SSL. Contivity and Shasta products support IPSEC. Contivity can support up to 5000 IPSEC tunnels and scales easily to support customer operational requirements. Shasta has the capability to support up to 30,000 tunnels.

For SSL needs, Nortel Networks offers several solutions.

- With Release 3.5, the Passport 8600 now gives you access to a fully integrated solution with the SSL module. Please refer to the SSL documentation for a complete description.
- Alteon Integrated Service Director (iSD) SSL Accelerator - The Accelerator functions as a transparent SSL proxy for the web switch to decrypt sessions, making encrypted cookies and URLs visible to the web switch for complete Layer 7 switching.

The Accelerator provides you with the ability to:

- Secure session content networking at wire speed.
- Off-load the web servers for better performance.
- Optimize web traffic of secure websites.
- Realize cost savings since there are fewer servers enabled in a server farm.

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, it encrypts the server response according to the negotiated encryption rules and forwards it to the requesting client using the established, secure HTTPS session. You can load balance up to 32 iSD-SSL units transparently with an Alteon web switch.

Chapter 8

Connecting Ethernet networks to WAN networks

A Passport 8672ATM module supports a choice of configuration options for your ATM networks. This chapter highlights some general design factors and techniques you need to be aware of when you are configuring a Passport 8672ATM module. These factors and techniques are organized according to the following general categories:

Topic	Page number
Engineering considerations	next
Feature considerations	329
Applications considerations	332

Engineering considerations

You should keep the engineering considerations described in the following sections in mind when configuring a Passport 8672ATM module:

- [“ATM scalability,”](#) next
- [“ATM resiliency”](#) on page 327

ATM scalability

The maximum number of Passport 8672 modules supported per chassis is as follows:

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

The maximum supported ELANs, PVCs, and VLANs are as follows:

- 256 RFC1483 BRIDGED/ROUTED ELANs per MDA
- 500 RFC1483 BRIDGED/ROUTED ELANs per switch
(12 more 1483 BRIDGED ELANs can be configured per switch)
- 64 PVCs per RFC1483 BRIDGED ELAN
- 1 PVC per RFC1483 ROUTED ELAN

Performance

The Passport 8672 ATM interface exhibits throughput of less than 50% of link bandwidth when dealing with a continuous stream of small packet sizes (less than 512 bytes).

You may notice some variations, however, in more of a real network scenario. In particular, this applies to those instances where the traffic stream is made up of a good mix of different packet sizes with large packet sizes contributing more to the link bandwidth than the small packet sizes. In such instances, the Passport 8672 ATM interface throughput is close to line rate. Such a scenario has been simulated in the Passport test lab by simultaneously sending multiple packet sizes over the ATM link. Throughput numbers observed during this testing process are as follows:

- For OC-3 bridged, the throughput is 125.9 Mb/s
- For OC-3 Routed, the throughput is 126.9 Mb/s
- For OC-12 Bridged, the throughput is 520.6 Mb/s
- For OC-12 Routed, the throughput is 507.4 Mb/s



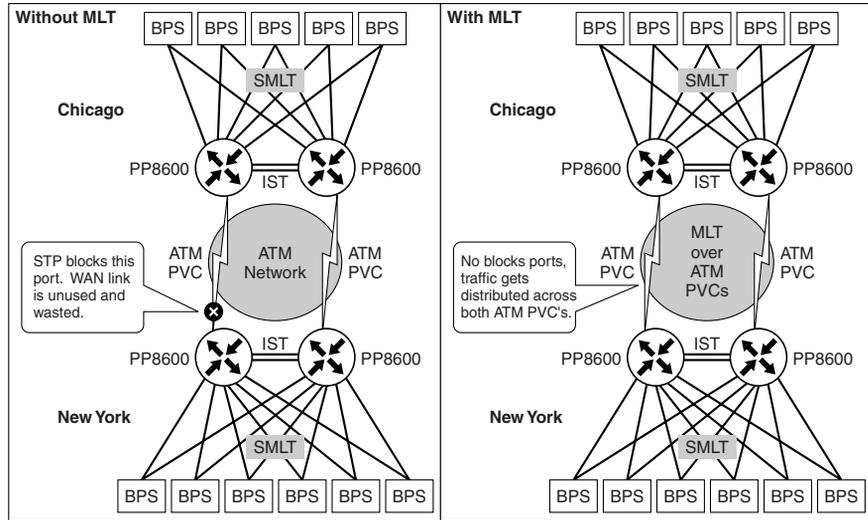
Note: Tests involved simultaneously sending 64, 128, 512, 1024, 1280, and 1518 bytes of traffic from a Gig Smartbit port over an ATM link in *both directions*. The Smartflow application used in these tests automatically adjusts the Transmit rate, so that the same number of packets for each size are sent over the ATM link without a single packet in any of the packet sizes being dropped. In this way, Smartflow also ensures that large packet sizes contribute much more to the link bandwidth than small packet sizes (i.e., simulating a real network scenario).

ATM resiliency

With MLT and SMLT over an ATM port, the Passport 8672 ATM module provides ATM resiliency. For example, MLT provides enormous value by presenting an opportunity to remove spanning tree protocol from the WAN links. When multiple paths exist between two bridges, STP will automatically block one of the links to prevent bridging loops. If the port that is blocked happens to be an expensive ATM WAN link, the customer ends up paying for a resource that is utilized only as a backup link. MLT removes the need to use spanning tree, and utilizes the trunk group in a load-sharing manner based on a MAC address hash algorithm.

SMLT introduces greater redundancy by allowing an MLT to terminate at two separate co-located switches that appear as one large switch. This protects against failure of an entire switch at the core. The PP8672 ATM module supports MLT and SMLT, with the restriction that the IST link between two switches cannot be ATM. [Figure 108](#) shows a network that uses STP, and one that uses SMLT across ATM WAN links.

Figure 108 Network with and without MLT



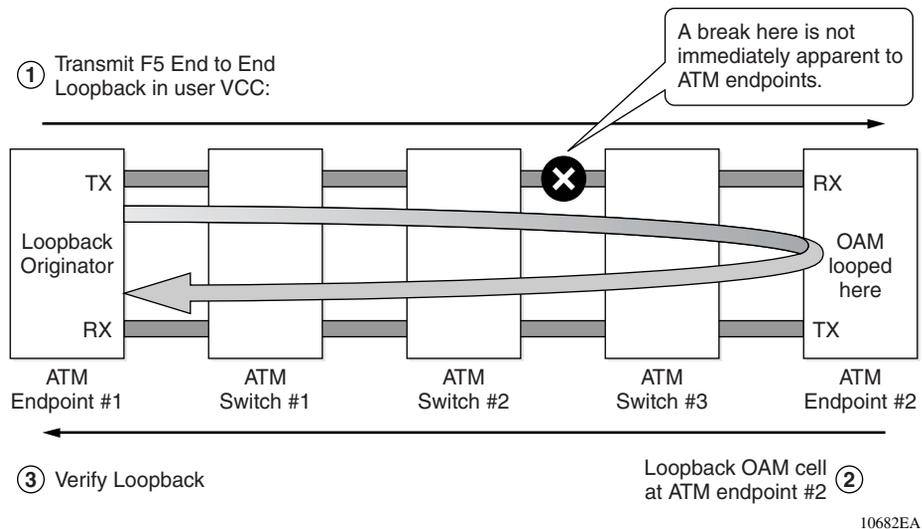
10681EA

When designing an MLT link, or any trunk, a separate PVC is required for each VLAN to be carried on the trunk.

F5 OAM loopback request/reply

The F5 OAM loopback request/reply feature (introduced in release 3.1.1) provides a mechanism to detect a failure within an ATM network that is downstream of the MDA ports. Without this feature, there is no notification to the 8672 of a broken PVC within the ATM network, i.e., between ATM switches 2 and 3 (Figure 109). The link is never perceived to go down and therefore, packets continue to be sent down the link, to be lost in the network. This feature utilizes an ATM loopback cell that gets sent end to end to determine the health of the link. If a configured number of loopback cells is not returned by the far end, the PVC is declared down.

Figure 109 ATM network broken PVCs



One important thing to note is that this feature requires that all PVCs on the port have the F5 OAM loopback feature enabled, and that all PVCs must fail the loopback in order to propagate the failure up to higher layers such as OSPF and MLT.

The result here is that while the feature may work well in point-to-point scenarios with all PVCs originating terminating at the same place, it may not work for other topologies. For example, if a port has multiple PVCs terminating at different locations (ie hub and spoke), failure of a single PVC due to a bad ATM switch at one remote location will not propagate up to higher levels, and features such as MLT will not work – traffic will continue to be sent out this PVC and be lost until a bridging or routing timer expires (‘black hole’).

By default, the F5 OAM loopback request feature is disabled. It must be enabled on each VC via CLI configuration commands or via Device Manager.

Feature considerations

You should keep in mind the feature considerations described in the following sections when configuring a Passport 8672ATM module:

- [“ATM and MLT,”](#) next
- [“ATM and 802.1q tags”](#) on page 329
- [“ATM and DiffServ”](#) on page 330
- [“ATM and IP multicast”](#) on page 330
- [“Shaping”](#) on page 332

ATM and MLT

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

ATM and 802.1q tags

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

ATM and DiffServ

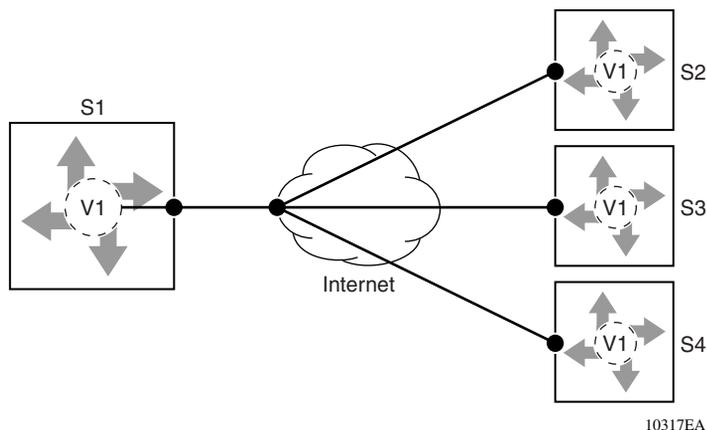
The Passport 8672ATM module can function as a DiffServ core port for the Passport 8600 switch. DSCP values placed by Ethernet ports behaving as DiffServ access ports, are preserved over ATM links.

The DSCP for Quality of Service (QoS) does not map directly to ATM class of service on the Passport 8672ATM module. To map QoS to ATM class of service, assign a QoS level to the Ethernet VLAN and configure a variable bit rate (VBR) class of service for a PVC in that VLAN.

ATM and IP multicast

If you configure a network in point-to-multipoint mode (hub to spoke mode), connecting a central Passport 8600 switch to several switches with PVCs on the same port in the same VLAN, multicast traffic will be flooded on these PVCs if just one of them has a member of an active multicast group. IGMP does not distinguish between different PVCs on the same VLAN when they are configured on the same port (Figure 110).

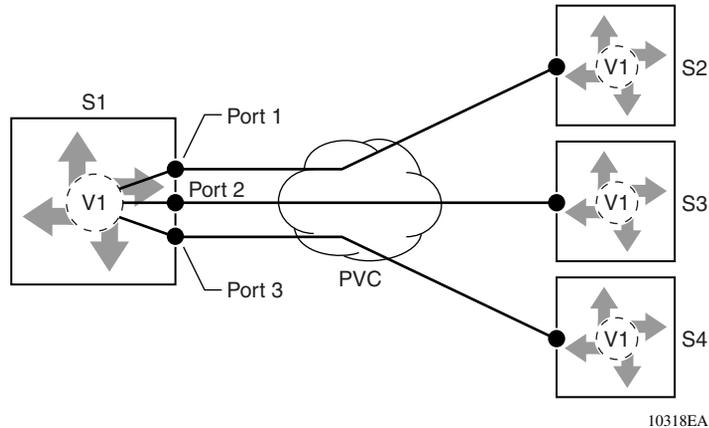
Figure 110 Point-to-multipoint IP multicast



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

If IP multicast over ATM is an essential requirement use PVCs on different VLANs connecting to the central switch, and traffic between these PVCs should be routed. If PVCs are required to be on the same VLAN, you can use different ports for these PVCs so that IP multicast traffic will flow only on the ports/PVCs with receivers (Figure 111).

Figure 111 IP multicast traffic over ATM



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

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

When using the IGMP fast leave feature with PVCs on the same port or VLAN flooded with traffic for a given group, if one member leaves the group all traffic for this group stops on all PVCs on the port or VLAN.

Shaping

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

The Passport 8672 ATM module supports shaping on a per-VC basis. The PCR, SCR, and burst size are all configurable for each PVC. For VBR service, a channel can burst at the PCR for MBS cells. If the MBS is exhausted, the channel will reduce to SCR as credits are accumulated to support another burst. The minimum PCR or SCR for a channel is 86 cells/sec or 36.67 Kbps. The maximum shaping rate per PVC is ½ of the link rate (i.e., 353207 cells/sec for OC-3/STM-1 and 733490 cells/sec for OC-12/STM-4).

Applications considerations

You should keep in mind the applications considerations described in the following sections when configuring a Passport 8672ATM module:

- [“ATM WAN connectivity and OE/ATM interworking,”](#) next
- [“Transparent LAN services”](#) on page 337
- [“Video over DSL over ATM”](#) on page 338
- [“ATM and voice applications”](#) on page 339

ATM WAN connectivity and OE/ATM interworking

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

- point-to-point leased line
- frame relay,
- POS, or ATM

The Passport 8672 provides the option of using ATM for WAN connectivity for sites that have ready access to an ATM network. This is a more economical solution than leased lines and allows access to higher bandwidths than frame relay.

For carriers, the 8672 can be used as an interworking point between new Optical Ethernet architectures and still prevalent and revenue generating ATM networks. It can provide a bridge to join these two types of networks until a full gigabit Ethernet core with MPLS is realized. An example of how this might work with a Passport 15000 ATM network is shown later in this section. In addition, the 8672 can be used as one of many options to bring services into an aggregation POP. Access sites that have been traditionally served with ATM, or are not serviced by dark fiber, can still use ATM to reach the aggregation site, which may have already migrated to a gigabit Ethernet/dark fiber access architecture.

Point-to-point WAN connectivity

This is the simplest and most common application for the 8672. Various sites within an enterprise network can be interconnected over a service provider's ATM network. Note that using the Spanning Tree Protocol (STP) can potentially require that one of the expensive WAN links be in a blocked state, wasting valuable resources. To avoid such a situation, Nortel Networks recommends that you instead use MLT or SMLT to provide Layer 2 redundancy.

SMLT also provides a load-sharing mechanism across the two WAN links, again, making the most effective use of the resources. Across high priority sites, running the two links across different service providers adds another layer of resiliency and reduces dependency on any one provider's network.

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

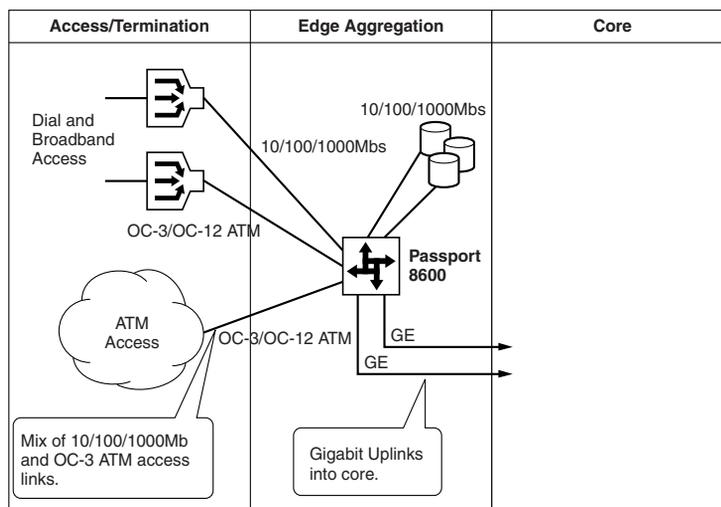
Service provider solutions – OE/ATM interworking

When discussing service provider applications, the focus will be on existing ATM networks, and new Optical Ethernet networks. Today, the majority of the revenue from data networks still comes from frame relay and ATM networks. Optical Ethernet is a new architecture which will eventually allow for seamless Layer 2 Ethernet connectivity for enterprise customers across both MANs and WANs. The Passport 8600 with the 8672 ATM module can be viewed as a way to extend the reach of ATM networks and services into an Optical Ethernet arena, and vice versa. These next sections will illustrate with some examples.

Carrier PoP aggregation

In [Figure 112](#), the Passport 8672 is being used as a method of bringing remote sites into an aggregation PoP over public ATM services.

Figure 112 Bringing remote sites into an aggregation PoP



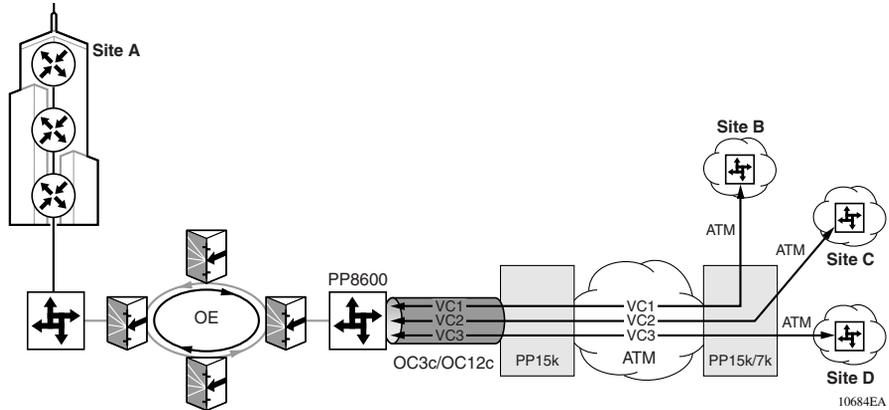
110683FA

This is most useful for sites that may not be reachable via gigabit Ethernet due to lack of dark fiber, but are readily serviced by public ATM. In this scenario, high speed gigabit Ethernet uplinks should be used to connect the 8600 to the core, rather than ATM links. ATM to ATM switching through a PP8600 is not recommended, as PP8672 ATM is a UNI device and not an NNI.

OE/ATM interworking- A detailed look

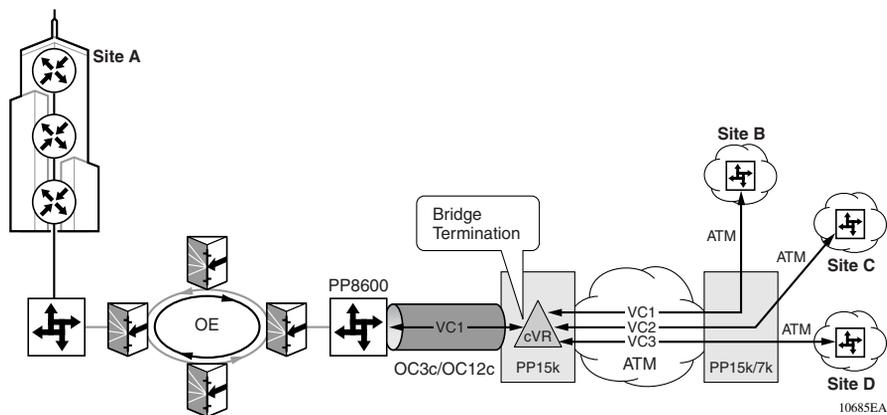
Figure 113, Figure 114, and Figure 115 provide a detailed view of how the Passport 8672 can facilitate the interworking of OE and ATM networks based on Passport 15000.

Figure 113 OE/ATM interworking- using home run PVCs



In Figure 113, ATM PVCs are mapped one to one from each remote site, through the ATM network (via PP15000), and terminate on the Passport 8672 ATM module. Bridging can be accomplished from each remote ATM site to anywhere on the OE ring, but remote ATM sites cannot bridge to each other. This should not be an issue when most remote sites need to access a data center or HQ on the OE core.

This solution works well for a customer that has a number of key sites (HQ, data center) on the OE ring, but needs to bring in some remote sites that do not have direct access to the ring. The ATM network could span a wide area (across MANs), or it might be a local ATM access network within a metropolitan area. For large numbers of remote sites, keep in mind the engineering rules for 8672 with regards to maximum number of PVCs, ELANs, and PVCs per ELAN (these are summarized in ATM Scalability section). A one-to-one mapping of PVCs to remote sites means that for any single ELAN, up to 64 remote sites (PVCs) can be supported.

Figure 114 OE/ATM interworking- using RFC 1483 bridge termination

In [Figure 114](#), routed PVCs in the ATM network bring the remote sites into a customer VR on PP15k at the interworking point. A single bridged PVC per customer is used to interconnect the PP8600 to the PP15k. Different customers would home to a different cVRs at the interworking point, with a different PVC between the PP8600 and PP15k. Bridged frames are terminated on PP15k and routed over ATM PVCs to their destination. This solution scales better by reducing the number of PVC's between the PP8600 and PP15k, but is a hybrid routed/bridged solution (no bridging end-to-end). ***Note that rfc1483 bridge termination feature on Passport 15000 is available starting at PCR 3.0.***

If necessary, multiple ATM ports can be used between PP8600 and PP15000. (They would need to be on different MDAs as ports on an MDA share the same forwarding engine). Egress shaping on the PP15k should be applied as the 8672 does not support policing. Without shaping, a burst of traffic from the PP15k could easily overwhelm the capabilities of the 8672, leading to lost cells and poor throughput.

Figure 115 OE/ATM interworking- using RFC 1483 bridge termination with cVRs

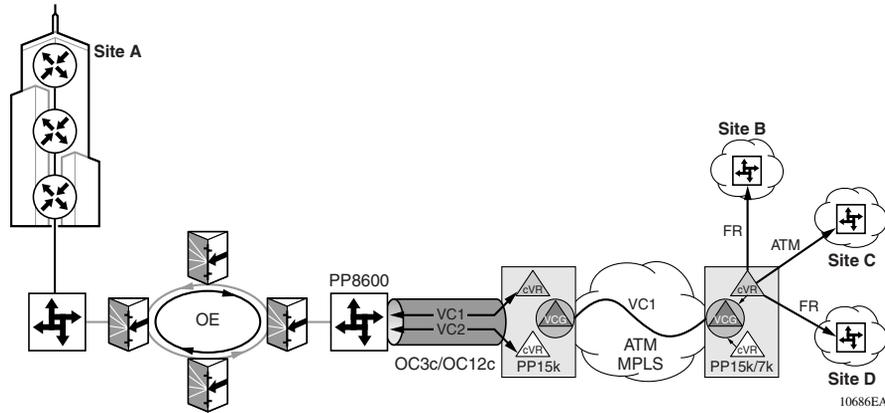
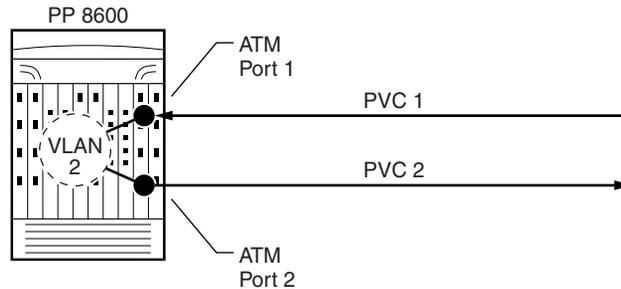


Figure 115 is similar to the routed/bridged model shown in Figure 114. However it uses cVRs at each customer POP to allow for a variety of connectivity options onto the customer premise, including frame relay. Multiple customers can be supported at each site using multiple cVRs. In this example, the cVR's are aggregated with a virtual connection gateway (VCG) into the ATM core. This simplifies engineering, but the cVRs must share bandwidth of a single VC. Another option is to provide premium service by using an individual mesh for each set of cVRs, which provides a reserved VC for each cVR.

Transparent LAN services

Figure 116 displays the supported transparent LAN services (TLS) configuration.

Figure 116 Supported TLS configuration

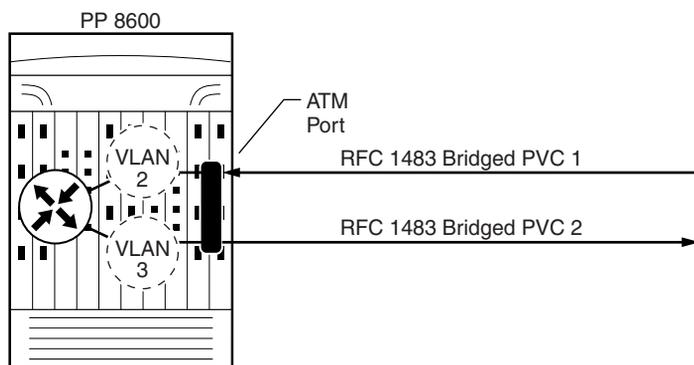


10319EA

The Passport 8672ATM module does not allow bridging between PVCs in the same VLAN on the same ATM port. Note that this is generally not an issue if the application requires many remote sites to access a central resource such as a data center, rather than exchanging traffic between different remote sites. Also, this can enhance security because you can place multiple customers in the same VLAN on the same ATM port and they will not be able to see each others' traffic, including broadcast traffic.

Alternatively, you can place two bridged PVCs in different VLANs on the same ATM port, and route between them using the Passport 8600 switch routing capabilities (Figure 117).

Figure 117 Configuring PVCs in different VLANs on the same ATM port



10321EA

Video over DSL over ATM

Nortel Networks recommends that you do not configure the Passport 8672ATM module to support video over DSL over ATM applications due to limitations imposed by current network designs, including the following:

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

Point-to-multipoint configuration for video over DSL over ATM

There are security issues relating to point-to-multipoint bridged ATM PVC mode because multicast is treated as broadcast for that VLAN on the ATM port. This characteristic limits the application to a single customer or subscriber per Ethernet VLAN on the same ATM port.

Point-to-point configuration for video over DSL over ATM

The limitation of one customer per Ethernet VLAN limits to 500 the number of subscribers to a Passport 8600 Series switch because this is the maximum number of routable VLANs currently supported. A Passport switch supports 1980 bridged VLANs. The number of ATM PVCs available on the Passport 8672ATM module is a secondary issue with respect to video over DSL applications.

ATM and voice applications

Before moving forward, note the following Passport 8600 switch and voice application characteristics:

- The Passport 8600 routing switch architecture is optimized for frame based applications
- Introducing ATM cells based interfaces such as OC-3, OC-12 or DS3 in such a system is a challenge in itself
- One pays for the overhead of converting Ethernet frames into fixed size ATM cells of 53 bytes. This overhead can have a significant impact especially at small Ethernet frame sizes (less than 512 bytes).
- The inherent characteristics of voice traffic include:
 - The average Ethernet frame size of voice traffic is 120 Bytes
 - The bursting nature of the traffic- in other words, there could be a continuous stream of small packets and links being idle afterwards
- Due to its nature, voice applications have very small delay and jitter tolerance

Design recommendations

When using Passport 8672 ATM for voice applications, Nortel Networks recommends that you:

- Under provision the PP8672 ATM link bandwidth when using it for voice over IP or ATM applications (i.e., provision your network in such a way that ATM link is not oversubscribed at all and is not running line rate almost all the times)
- Designate only 20% of link bandwidth for voice traffic to ensure that there is a good mix of small and large packet sizes traversing the ATM link
- Use separate VLANs for Voice and Data traffic
- Assign higher QoS level to Voice traffic VLANs
- Map voice traffic VLANs to VBR PVCs with SCR value carefully chosen based on your network design and the application. In this way, the Sustained Cell Rate is guaranteed for the voice traffic

ATM latency testing results

The actions of ATM Segmentation and Reassembly do incur some additional delay in transferring a packet through a switch, ranging from 65 microseconds to 232 microseconds for very large packets. However, the total latency introduced is well within the tolerance (in the order of milliseconds) required for time sensitive applications such as voice over IP. Testing was performed with a single flow of traffic between back to back 8600's connected with gigabit Ethernet trunks and then ATM trunks, with traffic at 50% or below to avoid congestion

Chapter 9

Provisioning QoS networks

This chapter describes a number of features to consider when configuring IP filtering, DiffServ and Quality of Service (QoS) on a Passport 8600 switch. The following topics are described:

Topic	Page number
Combining IP filtering and DiffServ features	next
IP filtering and ARP	342
IP filtering and forwarding decisions	343
IP filter ID	344
Per-hop behaviors	344
Admin weights for traffic queues	344
DiffServ interoperability with Layer 2 switches	345
DiffServ access ports in drop mode	346
Quality of Service overview	346
Nortel Networks QoS strategy	348
Passport 8600 QoS mechanisms	350
Passport 8600 network QoS	358
QoS summary	364
QoS and filtering	366
QoS flow charts	371
QoS and network congestion	375
QoS network scenarios	380

Combining IP filtering and DiffServ features

You can use IP filtering with DiffServ features in the following combinations:

- For IP routed traffic on DiffServ access ports, use source/destination filters
- For IP bridged traffic on DiffServ access ports, use global filters
- For filtering IP Multicast traffic, use global filters.

IP filtering and ARP

IP filters only affect the flow of IP traffic which has an Ethertype of 0800 and do not affect traffic from other Ethertypes, such as ARP, which has an Ethertype of 0806. When you configure a physical interface to have a default action of drop, it drops all traffic for which there is no matching forwarding filter. Non-IP traffic, particularly ARP packets, ingressing a port with a default action of drop, is never replied to by the Passport 8600 switch.

When you configure a port in drop mode, Nortel Networks recommends that you:

- statically configure the ARP entry related to the gateway on the end stations
or
- configure a protocol-based VLAN (using the 0x806 type) to capture the ARP traffic

If you configure the port in “forward” mode and define global filters to specify which traffic must be forwarded and a global filter to block all the traffic (0.0.0.0/0.0.0.0), it is possible to get the same result as configuring the port in “drop” mode. The next section provides information about global filters and IP filtering in general.

IP filtering and forwarding decisions

The following sections describe the two types of filters you can configure on the Passport 8600 switch:

- [“Global filters,”](#) next
- [“Source/destination filters”](#) on page 343

Global filters

Global filters are executed in a hardware component, called an ARU (address resolution unit). This ARU is an ASIC that makes the forwarding decision which does not require CPU activity and, therefore, does not have an adverse impact on forwarding speeds.

The maximum number of global filters you can configure per interfaces is:

- Eight global filters per group of eight 10BASE-T/100BASE-TX ports
- Eight global filters per Gigabit Ethernet port

Global filters can be applied to the following types of traffic:

- IP bridged traffic, i.e. traffic within the same VLAN
- Routed traffic if there are no DiffServ operations
- Multicast traffic with no minimum/maximum mask length

Source/destination filters

Source/destination filters are stored in memory associated with the ARU. The time required for the Passport 8600 switch to make a forwarding decision for a given IP routable packet is determined by the following factors:

- The number of source and destination filters configured for and associated with the source/destination IP addresses of this packet
- Any IP route that constitutes a less-specific match for one or both of these addresses

You can configure up to 3071 source/destination filters, and each associated IP address must have a minimum mask length of 8 bits.

By minimizing the number of source/destination filters associated with IP addresses, you minimize the lookup time necessary for the Passport 8600 switch to complete a forwarding decision for this packet.

When you are configuring source/destination filters, Nortel Networks recommends the following guidelines:

- In general, minimize the number of source/destination filters in your configuration to avoid multiple filter lookups
- Design your source/destination filters to be as specific as possible.
- Use the longest possible source/destination masks to avoid multiple filter lookups for different IP traffic flows)

IP filter ID

Filter IDs from 3072 to 4096 are reserved for internal usage (system filters). Do not use these filters IDs.

Per-hop behaviors

In a DiffServ network, traffic is classified as it enters the network and is assigned a per-hop behavior (PHB) based on that classification. On the Passport 8600 switch, the PHB dictates the drop precedence and latency flow experiences.

Admin weights for traffic queues

The Passport 8600 switch offers eight egress QoS levels (or queues) for traffic. The admin weights are currently non-configurable. These queues are administratively configured so that all the queues are serviced fairly. The weights are assigned through packet transmission opportunities (PTO) to each queue. More PTOs or higher admin weight is assigned to the high priority queues so that

the time sensitive transmissions are forwarded with minimum latency. The less time-sensitive or low priority traffic goes to the low priority queues, which are assigned lesser PTOs or lower admin weight. The configured default admin weights allow fair servicing of all low and high priority queues.

DiffServ interoperability with Layer 2 switches

Differentiated services (DiffServ) provides a mechanism by which discrete flows of IP traffic may be classified (marked) and forwarded with different levels of service. Any IP traffic can be marked for:

- IP routed use SRC/DST filters
- IP bridged use global filters

The BayStack 450 switch and the Passport 8100 Edge switch are both Layer 2 devices that cannot perform classification at Layer 3 level, and so cannot function as DiffServ edge devices.

However, these devices are capable of performing traffic classification at Layer 2, and may forward this traffic using a particular internal PHB based on this classification. Also, if tagging is enabled, this classification may determine the egressing IEEE 802.1p value for a given flow of traffic, which can be used to communicate the priority of a given flow to an adjacent device.

When you design a DiffServ network with a combination of Layer 2 devices in conjunction with the Passport 8600 switch, Nortel Networks recommends that the Passport 8600 switch be allowed to perform the DiffServ Layer 3 classification. This means that devices such as the BayStack 450 switch and Passport 8100 Edge switch should generally be connected to a DiffServ access port on the Passport 8600 switch, where Layer 3 classification can be performed.

Use the Layer 2 classification functionality of the BayStack 450 switch or the Passport 8100 Edge switch where appropriate to provide any desired per-hop behaviors in network devices outside of the DiffServ domain, but do not rely on those modules to provide formal DiffServ edge classification functionality.

DiffServ access ports in drop mode

When IP routable packets ingress on a DiffServ access port with a default action of drop, the packets are forwarded by matching an IP filter. If this filter does not modify the QoS level or the DSCP marking for a matching packet, then the packet is forwarded with the QoS based on the highest of port-, VLAN-, or MAC-configured QoS, and the DSCP is remarked using the egress QoS to DSCP table.

Quality of Service overview

Speed and performance have long been the hallmarks of LAN technology as the industry has developed increasingly faster and more intelligent networking solutions. Regrettably, the benefits of speed and performance have been primarily limited to single-purpose IP data networks.

The enterprise data network is quickly becoming the lifeblood of successful companies as more mission-critical applications are thrust upon the IP space. With the introduction of delay-sensitive and packet-loss-sensitive applications, such as telephony and video services transported over IP networks, today's network planners are focusing on traffic prioritization strategies as a new, mission-critical priority.

QoS in IP-based networks is essential because of IP's connectionless nature. IP provides what is called a *best effort* service. Thus, a simple IP network makes no guarantees about when data will arrive, or how much it can deliver. For example, the Internet was originally designed as a fairly *dumb* medium, using IP as a simple means of source and destination addressing, and TCP as a datagram ordering and re-ordering mechanism. The public Internet suffers from its lack of guaranteed performance and reliability.

To make the IP world offer a transparent service venue when compared to more traditional leased-line services or virtual circuit services such as Frame Relay and ATM, network planners have two choices:

- 1 Over-provision bandwidth
- 2 Adopt a QoS strategy

The over-provisioning of bandwidth no longer appears to be a workable solution. Over-provisioning is expensive and may not even solve all networking problems. Without a QoS mechanism in place to prioritize traffic across a big, fat, gigabit Ethernet pipe, the broadcasts generated by even low bandwidth, low priority IP data may cause service disruption to an IP video stream or an IP voice call by introducing jitter.

The focus of QoS is to provide predictable service in an increasingly IP-based world, especially during periods of congestion. It is these periods of congestion that are the target of QoS's traffic prioritization mechanisms. Several standards have evolved to address QoS, which are outlined in the section that follows.

There are a number of key parameters that define QoS in a network, including:

- Latency – Also referred to as propagation delay, latency represents the time between a node sending a message and receipt of the message by another node.
- Jitter – It is the variance of delay when packets do not arrive at the destination address in consecutive order or on a timely basis and the signal varies from its original reference timing. This distortion is particularly damaging to multimedia traffic. For example, the playback of audio or video data may have a jittery or shaky quality.
- Bandwidth – This parameter can be visualized as the pipe that delivers data, usually expressed in kilobits per second (Kbps) or megabits per second (Mbps). The bigger the pipe; the greater volume of data that can be delivered.
- Packet Loss – This parameter can be expressed as a percentage of packets, which could be dropped over a specified interval. Note that packet loss must be kept at a minimum in order to deliver effective IP telephony and IP video services. Specifically with IP Telephony, the selection of CODEC compression algorithm is important with respect to packet loss. For example, G.729 is more susceptible to the service impairment with packet loss than the G.711 algorithm.
- Availability – High overall availability is fundamental to delivering effective QoS. IP networks must be engineered to be telephony-grade IP networks in order to make delay-sensitive or jitter-sensitive applications successful over IP.

Nortel Networks QoS strategy

QoS is a fairly complex subject matter with many variables, many standards, and many permutations. For these reasons the basis and fundamentals of QoS are sometimes misunderstood or confused. Nortel Networks has developed a QoS strategy and roadmap with a primary objective to simplify this subject matter and provide a simple and easy-to-understand and implement solution. Nortel's QoS strategy also offers a unified and common "look and feel" to QoS across all Nortel Networks products, and provides a seamless end-to-end implementation that is easy to engineer, implement, and support. This common QoS strategy has been implemented on the recently developed products such as the Passport 8600 and it will be implemented on new/emerging products, but it is important to note that some of the older Nortel products may not be fully synchronized with this strategy.

This section provides a very brief look at Nortel's QoS strategy. For more information on the Nortel Networks Quality of Service roadmap, direction, vision, strategy, product status, etc. refer to the Nortel Networks QoS/Policy Resource Center at <http://qos.simi.baynetworks.com>.

Traffic classification

Nortel Networks' strategy simplifies QoS by providing a mapping of various traffic types and categories to a "Traffic Class of Service" using an intuitive easy-to-understand nomenclature. [Table 27](#) provides a summary of this mapping.

Table 27 Nortel Networks QoS traffic classification

Traffic Category		Application Example	Class of Service
Network Control		Alarms and heartbeats	Critical
		Routing table updates	Network
Real-Time, Delay Intolerant		IP Telephony Inter-Human comm.	Premium
Real-Time, Delay Tolerant		Video conferencing Inter-Human comm.	Platinum
		Audio/video on demand Human-Host comm.	Gold

Table 27 Nortel Networks QoS traffic classification

Traffic Category		Application Example	Class of Service
Non-RT Mission Critical	Interactive	eBusiness (B2B, B2C) Transaction processing	Silver
	Non-Interactive	Email Store and Forward	Bronze
Non-Real Time, Non-Mission Critical		FTP Best Effort	Standard
		PointCast Background/Standby	Custom/best effort

With this strategy, you simply select the Class for Service (i.e. Premium, Gold, Bronze, etc.) for a given device or for a group of devices and the network will take care of mapping the traffic to the right QoS level, mark the DSCP accordingly, set the 802.1p bits accordingly, and send the traffic to the appropriate emission queues for processing.

QoS-ready products, such as Passport 8600, provide default configurations that are streamlined for Quality of Service. Nortel Networks recommends, however, that you use extreme caution when changing any of these default QoS configurations. More details about these configurations are provided in the following sections.



Note: With release 3.3, the Passport 8600 supports a new set of filters, called multimedia filters. These allow you to automatically prioritize the traffic sent/received by Nortel VoIP platforms. See *Configuring IP Routing Operations* in the Passport 8000 Series documentation set for more details and instructions on configuring these filters.

Class of service mapping to standards

Table 28 provides a brief summary of how Nortel's QoS traffic classes of service (i.e. premium, platinum, gold, silver, etc.) map to some of the main QoS standards, both in the Local Area Network and Wide Area Network environments.

Table 28 Class of service mapping to standards

Nortel Networks Class of Service	DSCP DiffServ Code Point	802.1p user priority	ATM ¹ Class of service
Critical	CS7	7	CBR
Network	CS6	7	nrt-VBR
Premium	EF, CS5	6	CBR ²
Platinum	AF4x, CS4	5	rt-VBR
Gold	AF3x, CS3	4	rt-VBR
Silver	AF2x, CS2	3	nrt-VBR
Bronze	AF1x, CS1	2	nrt-VBR
Standard	DE, CS0	0	UBR
Custom/best effort	User Defined	1	UBR

¹ If a single ATM VC is used with different traffic classes, then the ATM class of service will be CBR.

² CBR is used when there is no VAD. Rt-VBR can be used when there is VAD.

Passport 8600 QoS mechanisms

QoS is without a doubt becoming an absolute necessity rather a luxury in today's IP networks. The driving reason for QoS is the migration of real-time and mission-critical applications such as voice, video, financial, e-commerce, etc. to IP networks. The Passport 8600 has a solid, well-defined architecture to handle QoS in an efficient and effective manner. This section provides in-depth details on the QoS features and capabilities of the Passport 8600.

There are 5 QoS mechanisms on Passport 8600 that will be explained in detail in the following paragraphs. These should be fundamental to your knowledge of designing QoS on Passport 8600:

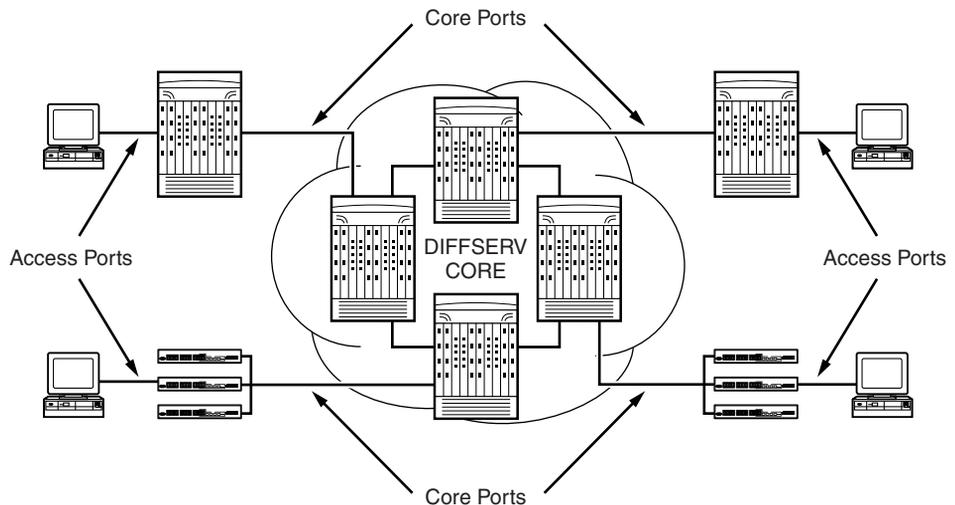
- 1 Internal QoS level
- 2 Emission Priority Queuing and Drop Precedence
- 3 Packet Classification
- 4 Filtering
- 5 Policing and Rate Metering

QoS highlights

The Passport 8600 provides a hardware-based Quality of Service platform through hardware packet classification. Packet classification is based on examining the various QoS fields within the Ethernet packet, primarily the DSCP and the 802.1p fields. Unlike legacy routers that require CPU processing cycles for packet classification, which could grind the router to a halt, the beauty of hardware-based QoS is that packet classification is performed in hardware at switching speeds.

The Passport 8600 provides the network administrator with a lot of flexibility to configure and customize QoS parameters on the Passport 8600; however, the default QoS settings are configured for optimum operation, and therefore caution should be used when changing these default values. The main objectives for simple nomenclature such “premium, gold, platinum, etc.” are to define the various QoS levels on the Passport 8600, simplify the user interface and avoid the complexities of configuring DSCP and 802.1p priority bit mapping.

There are two basic concepts in regards to QoS on the Passport 8600: The concept of a core port, and the concept of an access port. A core port, also known as a “trusted” port, will trust the QoS marking of the incoming packets and will pass the packets through the network based on their DSCP or 802.1p marking. A trusted port does not allow DSCP re-marking since it assumes DSCP marking was done prior to ingressing the port. An access port, also known as an “untrusted” port, will modify the QoS marking before the packet is sent to the network. [Figure 118](#) provides an illustration of how core ports and access ports are typically implemented in a network; these concepts are explained in more detail later in this section.

Figure 118 Passport 8600 core vs. access ports

10638EA

Internal QoS level

An internal QoS level is assigned to each packet that enters a Passport 8600 port based on different criteria. Once the QoS level has been set, the emission queue is determined and the packet is transmitted on the outgoing port. Note that the mapping of QoS levels to queues is a hard-coded 1-to-1 mapping.

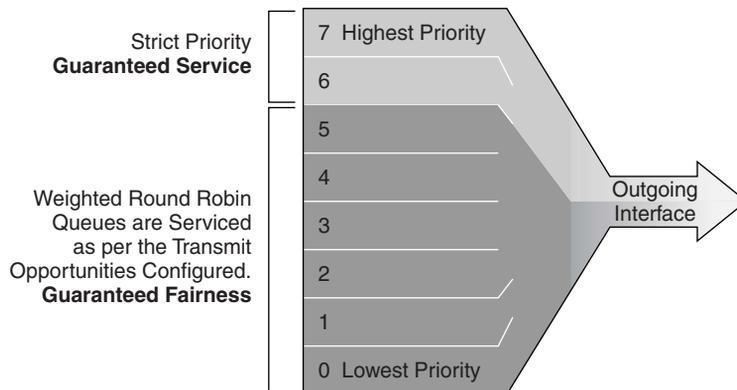
Emission priority queuing and drop precedence

There are two concepts that define the foundation for traffic management: emission priority, and discard priority. These concepts are two separate and orthogonal concepts, but they are indirectly related (namely through management of traffic in emission queues). Emission priority defines the urgency of the traffic, whereas discard priority defines the importance of the traffic. A packet with high emission priority means service this packet first through the emission queues. A packet with high discard priority means that this packet is to be the last to be discarded under congestion.

For example, delay sensitive traffic such as voice and video should be classified with high emission priority, whereas traffic that is sensitive to packet-loss such as financial information should be classified with high discard priority. In the Passport 8600 the emission priority and discard priority are commonly referred to as latency and drop precedence, respectively.

The Passport 8600 provides 8 hardware queues, essentially providing 8 different levels of emission priorities, or 8 different levels of QoS. Figure 119 provides an illustration of the hardware queues on the Passport 8600. The diagram is followed by an explanation.

Figure 119 Passport 8600 queue structures



10639EA

Each Ethernet port on the Passport 8600 has 8 emission queues with queue 0 having the lowest emission priority and queue 7 having the highest emission priority. These queues map directly to QoS levels 0 to 7 respectively. Queues 6 and 7 are known as “Strict Priority” queues, which means, they will be guaranteed service, and queues 1 to 5 are known as Weighted Round Robin (WRR) queues, which means that each queue will be serviced according to its queue weight after the strict priority traffic has been serviced.

Basically, the queue scheduler will service queue 7 first to completion, then will drop down and service queue 6 to completion, then it will drop down to service the WRR queues in a round robin scenario according to their weights. If one of the strict priority queues receives a packet while the scheduler is servicing one of the

WRR queues, the scheduler will complete servicing the current packet then it will immediately jump up to service the strict priority queue. Once the packet(s) in the strict priority queue has been serviced, it will then return to where it left off in the servicing of the WRR queues.

The weight of each queue is determined by what is known as its PTO. Table 29 provides a summary of the default queue configurations along with their packet transmission opportunities (PTOs) and effective queue weights.

Table 29 Passport 8600 QoS defaults

Class of Service	Emission queue	Type	Packet transmission opportunity	Percentage weight
Network	7	Strict priority	2	6%
Premium	6	Strict priority	32	100%
Platinum	5	WRR	10	31%
Gold	4	WRR	8	25%
Silver	3	WRR	6	18%
Bronze	2	WRR	4	12%
Standard	1	WRR	2	6%
Custom	0	WRR	0	0%

Table 30 Passport 8600 PTO settings

QoS Lev	Pto	% Wt	Time Slots																																	
			0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31		
7	2	6	x	x	Fixed (not configurable)																															
6	32	100	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x			
5	10	31			x	x	x			x	x	x														x										
4	8	25								x	x																									
3	6	18																																		
2	4	12																																		
1	2	6																																		
0	0	0																																		

There are a total of 32 PTOs. Queue 7 has 2 opportunities out of the 32 PTOs, giving it approximately a 6% weight. Queue 6 has 32 opportunities out of the 32 PTOs, giving it essentially 100% weight. Queue 5 has 10 opportunities out of the 32 PTOs, giving it about 31% weight. Queue 4 has 8 opportunities out of the 32 PTOs, giving it a 25% weight. Queue 3 has 6 opportunities out of the 32 PTO's, giving it approximately 18% weight.

Queue 2 has 4 opportunities out of the 32 PTOs, giving it about 12% weight. Queue 1 has 2 opportunities out of the 32 PTOs, giving it approximately a 6% weight. Queue 0 has 0 opportunities out of the 32 PTOs, giving it a 0% weight. The main reason for queue weights and WRR operation is to prevent the starvation of the lower priority queues. [Table 30](#) provides an illustration of the Passport 8600 PTO structure.

It is important to note that even though the Passport 8600 has different number of emission queues the DSCP and 802.1p priority bits are preserved across the network.

Packet classification

Ingress interfaces can be configured in two ways:

- 1 Where the interface is not configured to classify traffic, but merely forward it based on the packet settings as they ingress the Passport 8600 service provider network.

This mode of operation uses *Trusted Interfaces*, since the DSCP or 802.1p field is trusted to be correct and the edge switch performs the mapping without any classification taking place.

- 2 Where the Passport 8600 edge node classifies traffic as it enters the port and sets or modifies the DSCP or 802.1p field for further treatment as it traverses the Passport 8600 network.

For a DSCP classification scheme, [Table 28](#) shows the recommended configuration that the service provider should deploy. It should be noted that this should be used as a starting point, since the actual traffic types and flows are not yet known at this stage. As such, the mapping scheme referred to here may not be optimal for the service provider.



Note: Nortel Networks recommends not changing the default values. If you change the values, make sure that the values are consistent on all other Passport switches and other devices in your network. Inconsistent mapping of table values can result in unpredictable service levels. See [Appendix A, “QoS algorithm,” on page 393](#) for information on QoS default values.

Two standard PHBs are defined in IETF RFCs 2597 and 2598. The first RFC describes the Assured Forwarding PHB group, which further divides delivery of IP packets into four independent classes. The Assured Forwarding PHB group offers different levels of forwarding resources in each DiffServ node. Within each Assured Forwarding PHB group, IP packets are marked with one of three possible drop precedence values. In case of network congestion, the drop precedence of a packet determines its relative importance with the Assured Forwarding group.

RFC 2598 defines the Expedited Forwarding PHB group as the “Premium” service, the best service your network can offer. Expedited Forwarding PHB is defined as a forwarding treatment for a particular DiffServ microflow when the rate of the microflow’s packets from any DiffServ node ensures that it is the highest priority and experiences no packet loss for in-profile traffic.

Although the major QoS focus is on IP, the Passport 8600 also allows traffic prioritization for non-IP traffic. This is achieved by assigning various QoS levels to the VLANs, or to the physical pots, or a priority based on the MAC address. In cases where more than one level of QoS is defined i.e. the VLAN QoS is different from the port’s QoS, which is different from the MAC address QoS, the highest level of QoS will be honored.

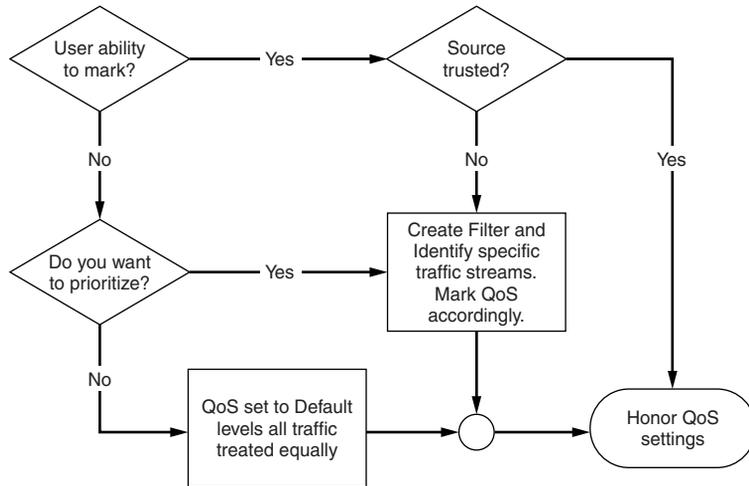
Note that this prioritization is local to each Passport 8600 and will only prioritize data as it egresses the switch, however. It is up to the subsequent switches in the path to ensure that the traffic will be prioritized through them accordingly in order to get end-to-end Quality of Service.

Filtering

When using filters to design QoS on Passport 8600, there is a decision-making process to determine whether to filter or not. This decision-making process is outlined in [Figure 120](#). The key questions here are:

- 1 Does the user/application have the ability to mark QoS on the data packets?
- 2 Is the source of the traffic trusted? In other words, is the QoS levels set to where they should be for each data source? There is the possibility that users may maliciously set QoS levels on their devices to take advantage of higher queues.
- 3 Do you want to prioritize traffic streams?

Figure 120 QoS filtering decisions



10640EA

Policing and rate metering

As part of the filtering process, the administrator or service provider has the ability to police the amount of traffic that each user is able to transmit. 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 criteria assigned by network managers.

For service providers policing at the edge is used to provide different bandwidth options as part of a Service Level Agreement (SLA). For example, in an enterprise network you may choose to limit the traffic rate from one department to give mission-critical traffic unlimited access to the network. In a service provider network you may want to control the amount of traffic customers sent through the network to ensure that they comply with the SLA they signed for. In terms of QoS, it will ensure that users do not exceed their traffic contract for any given QoS level. Rate metering will give the administrator the ability to limit the amount of traffic coming for a specific user in two ways.

- 1 Drop out of profile traffic.
- 2 Re-mark out of profile traffic to a lower (or higher) QoS level to be treated as necessary upon the onset of port congestion.

As a function of the filtering mechanisms in the Passport 8600, rate metering can only be performed on a Layer 3 basis.

Passport 8600 network QoS

There are 4 QoS design considerations for a Passport 8600 network that will be explained in detail in the following sections. Each of these is very important in properly designing QoS in the network:

- 1 Trusted vs. untrusted interfaces
- 2 Access vs. core ports
- 3 Bridged vs. routed traffic
- 4 Tagged vs. untagged packets

Trusted vs. untrusted interfaces

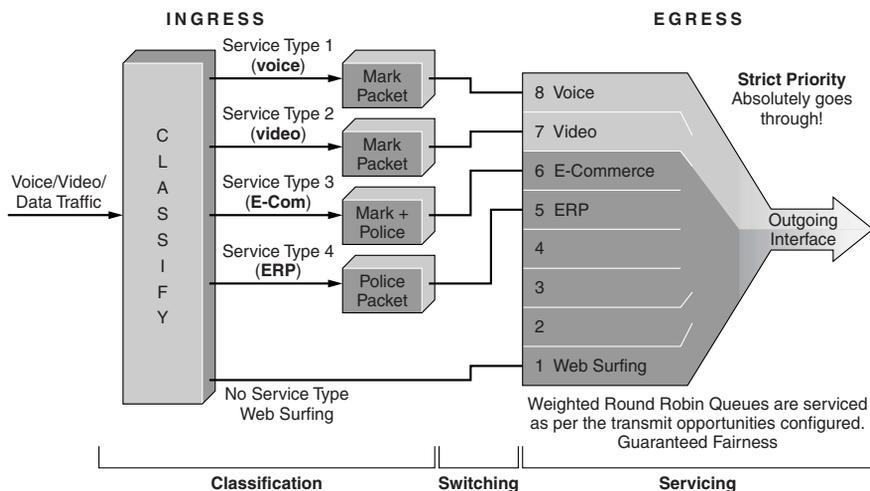
Using trusted interfaces allows the service providers and end users to mark their traffic in a specific way and the service provider will ensure those packets are treated according to the service level for those markings. Trusted Interfaces are used when the service provider wants to have control over the traffic prioritization through the network. If a service provider was to use 802.1p as the method of applying desired CoS attributes before forwarding to the access node, the service provider is able to classify other protocol types such as IPX ahead of IP packets if that were required.

Untrusted interfaces are where service provider wishes to control the classification and mapping of traffic for delivery through the network. This would be done where a service provider has no interest in controlling these functions, or is unable to do so. Untrusted interfaces require the configuration of more complicated filter sets to be created and managed by service provider in order to classify and re-mark traffic on ingress to the network. For untrusted interfaces in the packet forwarding path, the DSCP in the IP header will be configured to be 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.

Access vs. core port

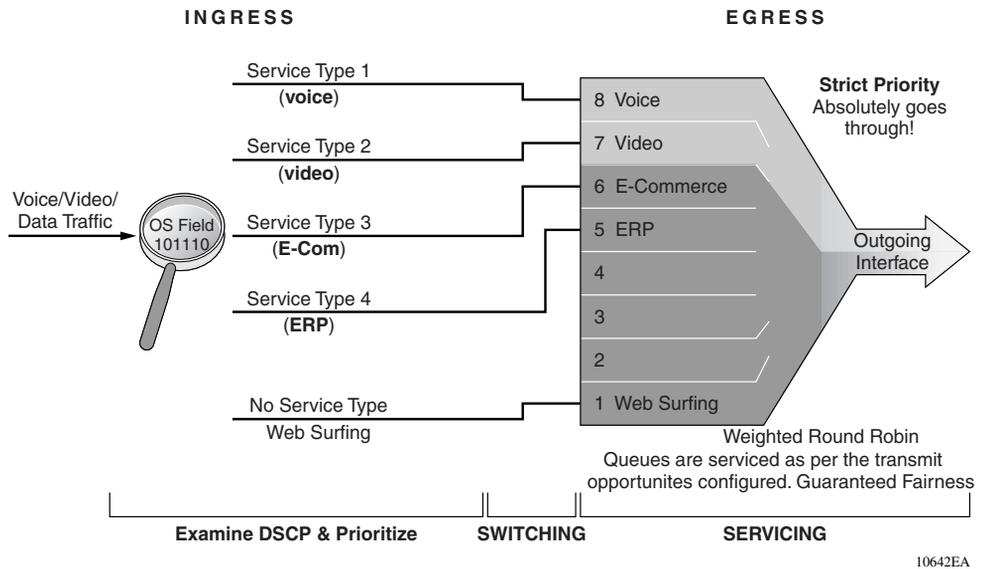
By definition, an access port on a Passport 8600 is an untrusted port from a QoS perspective, and therefore traffic entering an access port is re-marked with the appropriate DSCP and 802.1p and given an internal QoS level. This re-marking is done based on the filters and traffic policies that are enabled.

[Figure 121](#) provides an illustration of how packets are processed through an access port. The diagram is then followed by a summary of how the various packet types are processed through an access port.

Figure 121 Passport 8600 access port

10641EA

A core port on a Passport 8600 is classified as a trusted port from a QoS perspective. A core port preserves the DSCP values and the 802.1p priority bits as they enter the switch, uses these values to assign a corresponding QoS level to the packets, and sends the packets to the appropriate emission queues for servicing. [Figure 122](#) provides an illustration of how packets are processed through a Core port.

Figure 122 Passport 8600 core port

Bridged vs. routed traffic

In a service provider network the access and nodes consist of the Passport 8600 configured for bridging. The Passport 8600 also uses DiffServ to manage network traffic and resources, except many of the features are unavailable in bridging modes of operation. For bridging, ingress traffic is mapped from IEEE 802.1p bits to the appropriate QoS level and egress traffic is mapped from QoS level to the appropriate IEEE 802.1p bits as per [Table 31](#).

Table 31 IEEE 802.1p bits to QoS level mapping

IEEE 802.1p	QoS Level	Nortel Networks Class of Service
7	7	Critical
		Network
6	6	Premium
5	5	Platinum
4	4	Gold
3	3	Silver
2	2	Bronze

Table 31 IEEE 802.1p bits to QoS level mapping

IEEE 802.1p	QoS Level	Nortel Networks Class of Service
1	0	Standard
0	1	Custom/best effort

In an enterprise network, the access nodes consist of the Passport 8600 configured for bridging and the core nodes consist of the Passport 8600 configured for routing. For bridging, ingress and egress traffic is mapped as per [Table 31](#). For routing, ingress traffic is mapped from DSCP to the appropriate QoS level and egress traffic is mapped from QoS level to the appropriate DSCP as per [Table 32](#).

Table 32 DSCP to QoS level mapping

DSCP DiffServ Code Point	QoS Level	Nortel Networks Class of Service
CS7	7	Critical
CS6		Network
EF, CS5	6	Premium
AF4x, CS4	5	Platinum
AF3x, CS3	4	Gold
AF2x, CS2	3	Silver
AF1x, CS1	2	Bronze
DE, CS0	0	Standard
User Defined	1	Custom/best effort

Tagged vs. untagged packets

In a service provider network, the customer must utilize an 802.1Q tagged encapsulation on the CPE equipment in order for the Passport 8600 to map the 802.1p user priority to queue. This is required since the Passport 8600 will not provide the necessary classification functionality when operating in bridging mode. If the customer is not using 802.1Q tagged encapsulation to connect to the Passport 8600 switch, traffic can only be classified based on VLAN membership, port or MAC address.

To ensure consistent layer-2 QoS boundaries within the service provider, customers should be forced to utilize 802.1Q encapsulation when connecting their CPE directly to the Passport 8600 access node. If the customer is not interested in performing packet classification, then a Business Policy Switch (BPS) 2000 should be used for the customer to connect to. The BPS 2000 is then connected to the access node. In this case, service provider would configure the traffic classification functions in the BPS 2000.

At the egress access node, packets are examined to determine if their IEEE 802.1p bits or DSCP will be re-marked before leaving the network as per [Table 33](#).

Table 33 QoS level to IEEE 802.1p and DSCP mapping

DSCP DiffServ Code Point	IEEE 802.1p	QoS Level	Nortel Networks Class of Service
CS7	7	7	Critical
CS6			Network
EF, CS5	6	6	Premium
AF4x, CS4	5	5	Platinum
AF3x, CS3	4	4	Gold
AF2x, CS2	3	3	Silver
AF1x, CS1	2	2	Bronze
DE, CS0	0	1	Standard
User Defined	1	0	Custom/best effort

Upon examination, if the packet is egressing as a tagged packet, the IEEE 802.1p tag is set based on the QoS level-to-IEEE 802.1p bit mapping. In bridged packets, the DSCP is reset based on the QoS level.

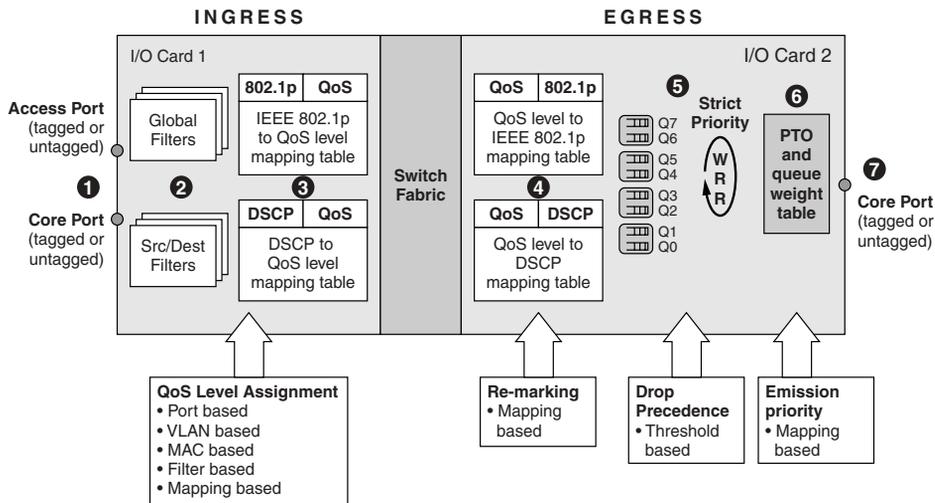


Note: Nortel Networks recommends not changing the default values. If you change the values, make sure that the values are consistent on all other Passport switches and other devices in your network. Inconsistent mapping of table values can result in unpredictable service levels. See [Appendix A, “QoS algorithm,” on page 393](#) for QoS default values.

QoS summary

Figure 123 illustrates the QoS mechanisms described in this chapter.

Figure 123 Passport QoS summary graphic



10643EA

The main QoS function of the ingress card (i.e., Card 1) is to assign a QoS level to each packet that enters on a Passport 8600 port. Depending on the type of port (access or core) and whether tagged or untagged, the packet is treated according to the flow charts shown in “QoS flow charts” on page 371.

There are 3 levels of QoS level assignment:

- 1 Port and packet (VLAN or MAC) based
- 2 Filter based
- 3 802.1p/DSCP mapping based

The egress card (i.e., Card 2) has 3 QoS functions:

- 1 Re-marking
- 2 Drop precedence treatment
- 3 Emission priority queueing

The required QoS parameters (i.e. latency, jitter, bandwidth and packet loss) for different service class (i.e. voice, video/audio, email, etc) are provided by the drop precedence and emission priority functions. For example consider 3 services classes like voice, video/audio and email are being sent on an ingress access port. Every service class has specific requirements to ensure traffic is transmitted properly and quality is maintained through the network.

You want the voice traffic to have low latency, jitter and packet loss. However the video/audio traffic is tolerant to delay but requires low packet loss and bandwidth guarantee, but you do not want the email traffic that is tolerant to delay and packet loss to affect the other traffic types in order to have better quality. And you want to maintain the quality even during periods of congestion in the network. With a distributed architecture Passport 8600 switches traffic from ingress ports to one or more egress ports where packets are queued for transmission.

This is the most critical point where depending on the QoS level assigned on ingress the emission queue is chosen accordingly. What it means is that each queue has its own latency and jitter values which makes it a lot easier to choose the right queue for a specific application. For example the voice traffic should always go in queue 6, the video/audio traffic should go in queue 5 or 4 and the email traffic should go in queue 3 or 2. Furthermore, each emission queue has its own weight that guarantees a percentage of the transmission time that translates to bandwidth. For example the voice traffic should always get 100% of the bandwidth, the video/audio traffic should get between 25% and 31% of the bandwidth and the email traffic should get between 13% and 19% of the bandwidth. This assumes the bandwidth is fully utilized otherwise each traffic type can use the entire bandwidth on its own. Packet loss is directly related to the drop precedence during periods of congestion.

Depending on congestion levels one of the four thresholds is crossed which correspond to a pair of emission queues. This means that traffic in lower emission queues is partially discarded in order to protect the traffic in the higher emission queues. For example the voice traffic would not be discarded until the 4th threshold is crossed, the video/audio traffic would not be discarded until the 3rd threshold is crossed and the email traffic would not be discarded until the 2nd threshold is crossed.

QoS and filtering

In the most typical of cases, filters are applied to act as firewalls or Layer 3 redirection. In more advanced cases, traffic filters can be used to identify traffic streams (Layer 3 and Layer 4) and re-mark and classify them for a specific QoS Level at both Layer 2 (802.1p) and Layer 3 (DSCP). Filtering on the 8600 being hardware-based (RAPTARU internal registers for global filters, and associated RAPTARU memory for source/destination filters), there is never a need for CPU intervention which means that there is no load on the CPU and that filtering can be achieved at wire-speed resulting in virtually no performance impact on users.

Traffic filtering is a key feature that complements the 8600 QoS capabilities. The switch by default can read incoming 802.1p or Diffserv markings and forward traffic based on their assigned QoS levels. However there are situations where these 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, filtering can be used to prioritize (or de-prioritize) specific traffic streams.

This section outlines key situations where filtering will play a large part in ensuring that users and applications receive the correct level of service on an end-to-end basis.

Filtering

Filtering is the key process used to finish QoS. The Passport 8600 has the ability to re-mark the 802.1p bits and the DSCP field. The Passport 8600 uses an internal value, called *Internal* QoS to make the mapping between all the parameters that can be used (802.1p bits, MAC, port, VLAN QoS level, DSCP).

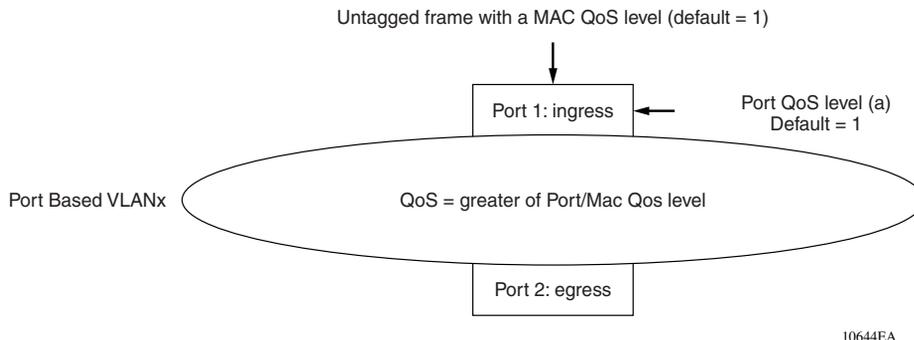
DiffServ access port (IP bridged traffic with DiffServ enabled)

To enable DiffServ on a port, use the following CLI command:

```
config eth <port> enable-diffserv true
```

By default, the mode is “core”, so we have to configure the mode to access: **config eth <port> access-diffserv true**. With Device Manager, double click on a port, and check the button “DiffServ enable” and select the mode. [Figure 124](#) shows the untagged ingress traffic on the port-based VLANs.

Figure 124 Untagged ingress traffic on the port-based VLANs:



In [Figure 124](#), the QoS internal level is determined by taking the greater value of Port QoS level or Mac QoS level. At the egress, DSCP and 802.1p bits are re-marked according to this internal QoS value. The mapping is based on [Table 31](#).

- a** The port QoS level is accessible via the following CLI command:

```
config ethernet <port> <qos-level>
```

or via Device Manager. Open the DM and double click on a port. The “QoS level” is directly accessible. Note that the value “7” is not acceptable (only “0” to “6” values are acceptable).

- b** The MAC level is accessible via the following CLI command:

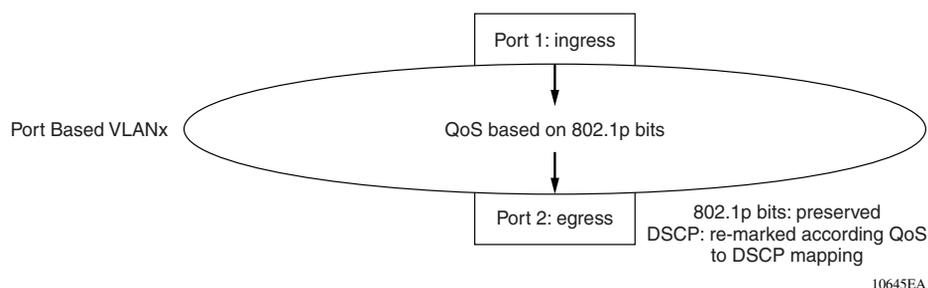
```
config vlan <vlan-id> fdb-entry qos-level <mac> status <value>
```

where <mac> is the mac address (in the format 0x00:0x00:0x00:0x00:0x00:0x00), <value> the QoS value (from “0” to “6”) and <status> the way the MAC has been learned (other | invalid | learned | self | mgmt).

In this configuration, you can use global filters (because it is IP bridged traffic) to modify the DSCP and the 802.1p bits. The QoS internal level is based on the greater value of the port QoS level, the MAC QoS level and the value determined by the filter. At egress, for DSCP, if the MAC QoS level is greater than the port or the VLAN QoS level, DSCP is re-marked according the MAC QoS level, or else marked with the value determined by the DSCP filter. If the egress port is tagged, 802.1p bits are as the internal QoS.

Figure 125 shows the tagged ingress traffic on the port-based VLANs.

Figure 125 Tagged ingress traffic on the port-based VLANs:



In Figure 125, the QoS internal level is determined by the 802.1p bits value of the ingress tagged frame. These 802.1p bits are preserved at the egress. The QoS value can be used to re-mark DSCP at egress.

In this configuration, you can use global filters (because it is IP bridged traffic) to modify the DSCP and the 802.1p bits. The internal QoS level is based on the greater value of the ingress 802.1p bits and the value determined by the filter (modification of the 802.1p bits with the filter). DSCP can be re-marked by a filter. If the egress port is tagged, 802.1p bits are re-marked based on greater of ingress 802.1p bits and the value determined by the filter.

Source MAC-based VLANs

For an ingress untagged port, the internal QoS level is based on the greater of port level QoS and destMAC QoS level. The srcMAC QoS level and VLAN QoS level are not effective. At egress, the DSCP is re-marked based on the internal QoS level. Egress 802.1p bits are defined as per internal QoS. For ingress tagged port, the internal QoS level is based on the 802.1p bit value. At egress, the DSCP is re-marked based on the internal QoS level. 802.1p bits are preserved.

Protocol-based/IP subnet-based VLANs

For an ingress untagged port, the internal QoS level is based on the greater VLAN QoS level and MAC QoS level. At egress, the DSCP is re-marked based on the internal QoS level. Egress 802.1p bits are defined as per internal QoS.

For an ingress tagged port, the internal QoS level is based on the 802.1p bit value. At egress, the DSCP is re-marked based on the internal QoS level. The 802.1p bits are preserved.

In this configuration, you can use global filters (because it is IP bridged traffic) to re-mark the DSCP and 802.1p bits. For an untagged port, the internal QoS level is based on the greater of the VLAN QoS, MAC QoS and the level determined by the filter. The DSCP can be re-marked based on:

- MAC QoS if the MAC QoS is greater than the Port and VLAN QoS
- value determined by the DSCP in other cases

At egress, 802.1p bits are defined as per internal QoS. If the ingress port is tagged, the internal QoS level is based on the greater value of the ingress 802.1p bits and the value determined by the filter (modification of the 802.1p bits with the filter). The DSCP is re-marked based on the QoS. If the egress port is tagged, 802.1p bits are re-marked based on greater of ingress 802.1p bits and the value determined by the filter.

Core port (IP bridged traffic)

The core port includes port-based and protocol-based/IP subnet/source MAC-based VLANs. These are described in the subsections that follow.

Port-based VLANs

In the case of port-based VLANs, the internal QoS level is based on the ingress DSCP value. As a core port, there is no modification of this value. DSCP is preserved. If the egress port is tagged, the 802.1p bits are defined according the internal QoS. If the ingress port and the egress port are tagged, the 802.1p bits are preserved.

Non-IP traffic (bridged or L2)

Non-IP traffic includes port-based, protocol-based, and source MAC-based VLANs. These are described in the subsections that follow.

Port-based VLANs

In the case of port-based VLANs, the internal QoS is based on the greater of the port and MAC QoS levels. At egress, if the port is tagged, the 802.1p bits are based on the internal QoS. If the ingress port is tagged, the QoS is based on ingress 802.1p bits. At egress, if the port is tagged, the 802.1p bits are preserved.

Protocol-based VLANs

In the case of protocol-based VLANs, the internal QoS is based on the greater of the VLAN and MAC QoS levels. At egress, if the port is tagged, the 802.1p bits are based on the internal QoS. If the ingress port is tagged, the QoS is based on ingress 802.1p bits. At egress, if the port is tagged, the 802.1p bits are preserved.

Source MAC-based VLANs

In the case of source MAC-based VLANs, the internal QoS is based on the greater of the port and destination MAC QoS levels. Source MAC QoS level and VLAN QoS are not effective. At egress, if the port is tagged, the 802.1p bits are based on the internal QoS. If the ingress port is tagged, the QoS is based on ingress 802.1p bits. At egress, if the port is tagged, the 802.1p bits are preserved.

Note that there is a limitation in such cases. You cannot use the DM to supply the QoS level of the destination MAC. You should use the CLI command instead.

DiffServ access (IP routed traffic)

For IP routed traffic, the DSCP field is set to 0 at ingress. QoS is determined by source/destination filter profiles. The DSCP can be re-marked by source/destination filters only.

DiffServ core (IP routed traffic)

The DSCP field is honored and traffic sent to the queue according the DSCP value.

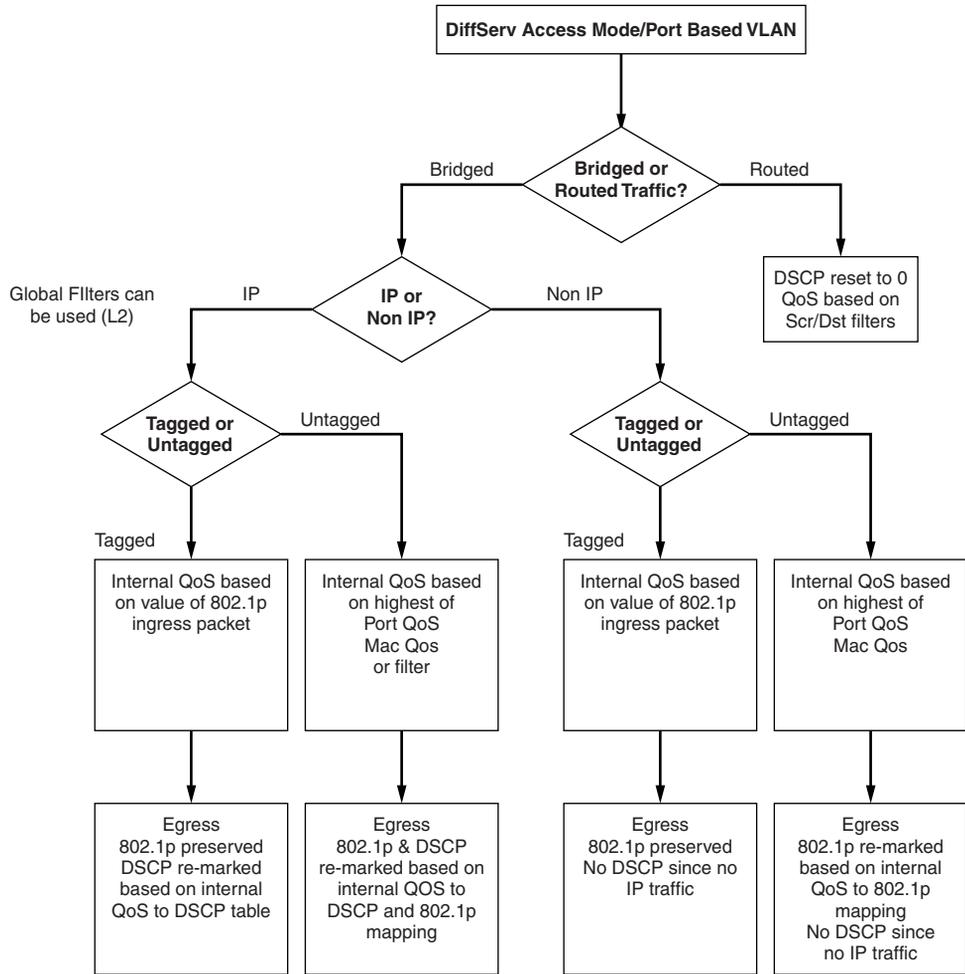
QoS flow charts

The graphics that follow display flow charts for DiffServ access mode:

- Port-based VLANs ([Figure 126](#))
- MAC-based VLANs ([Figure 127](#))
- IP subnet and protocol-based VLANs ([Figure 128](#))

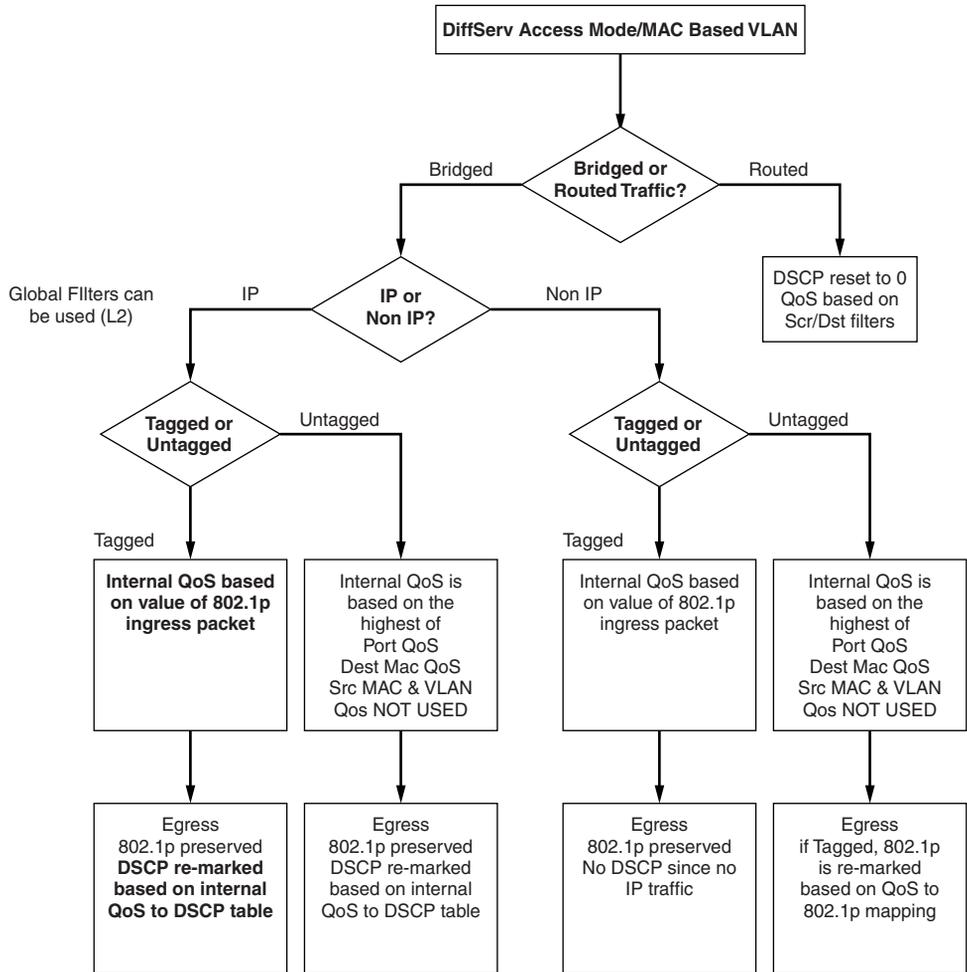
[Figure 129](#) provides a flow chart for DiffServ core mode.

Figure 126 DiffServ access mode- port-based VLANs



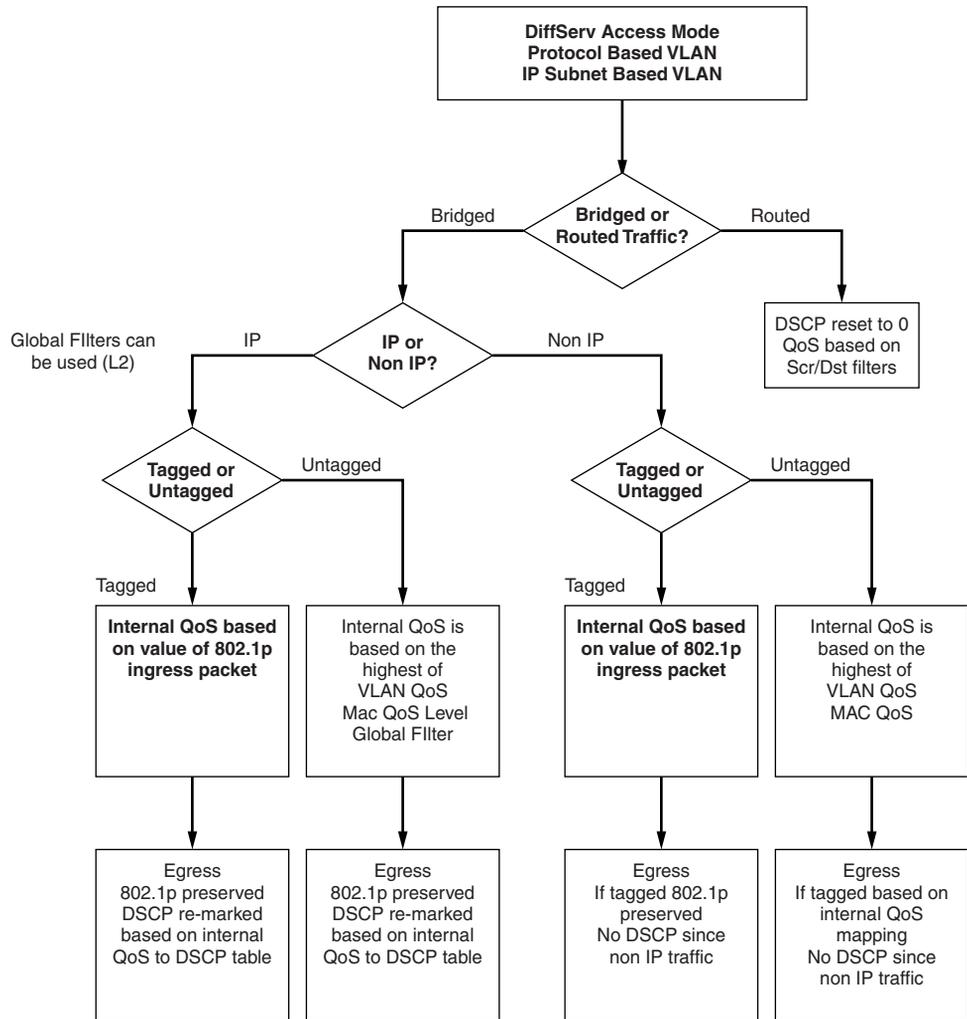
10646EA

Figure 127 DiffServ access mode- MAC-based VLANs



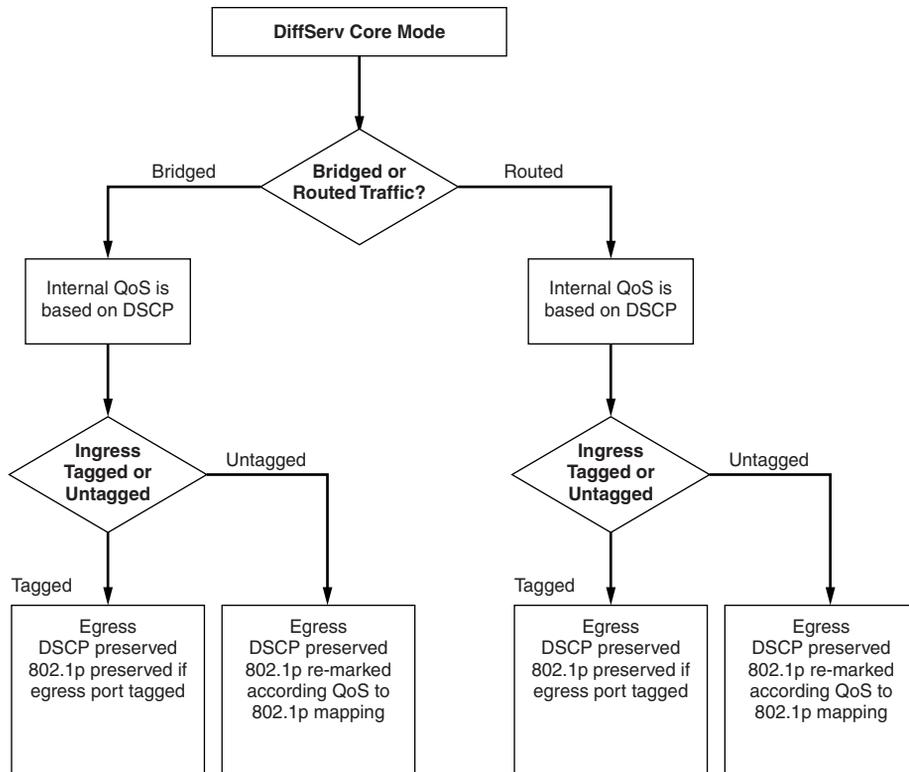
10647EA

Figure 128 DiffServ access mode- IP subnet and protocol-based VLANs



10648EA

Figure 129 DiffServ core mode



10649EA

QoS and network congestion

When engineering Quality of Service in a Network, one of the major elements you need to take into consideration is the various network congestion scenarios, and the traffic management behavior under these congestion scenarios. Congestion in a network can be caused by many different conditions and/or events; including node failures, link outages, broadcast storms, user-traffic bursts, etc.

At a high level, there are 3 main types or stages of congestion:

- 1 No congestion
- 2 Bursty congestion
- 3 Severe congestion

Each of these stages are explained in the paragraphs that follow.

No congestion

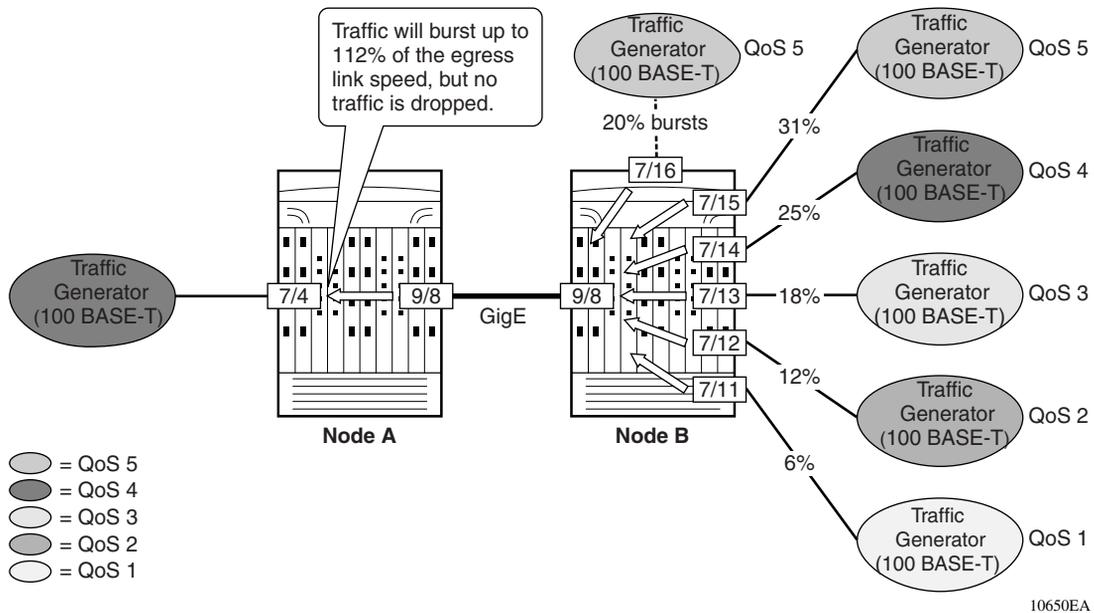
In a non-congested scenario, it is important to understand that QoS is still required and plays a very valuable and crucial role to ensure that delay-sensitive applications, such as real-time voice and video traffic, are emitted before lower-priority traffic. The prioritization of delay-sensitive traffic is essential in order to not only minimize delay, but also reduce or eliminate jitter, which could have a detrimental impact on these applications.

With 8 levels of QoS, and an architecture that supports both strict priority and weighted round robin queuing, the Passport 8600 can provide superior QoS capabilities with minimum delay and negligible or no jitter. In addition, the Passport 8600 has 8 hardware queues, while most of its competitors usually have only 2 to 4.

Momentary bursts of congestion

In a network environment, the network can experience momentary bursts of congestion for reasons such as network failures, rerouting, broadcast storms, etc. The Passport 8600 is designed just for such situations. It has sufficient queue capacity and an efficient queue scheduler to handle bursts of congestion in a seamless and transparent manner. [Figure 130](#) provides an example of a situation in which the traffic can burst over 100% within the WRR queues and yet no traffic is dropped because of the traffic management buffers and the efficiency of the queue scheduler.

Figure 130 Congestion bursts



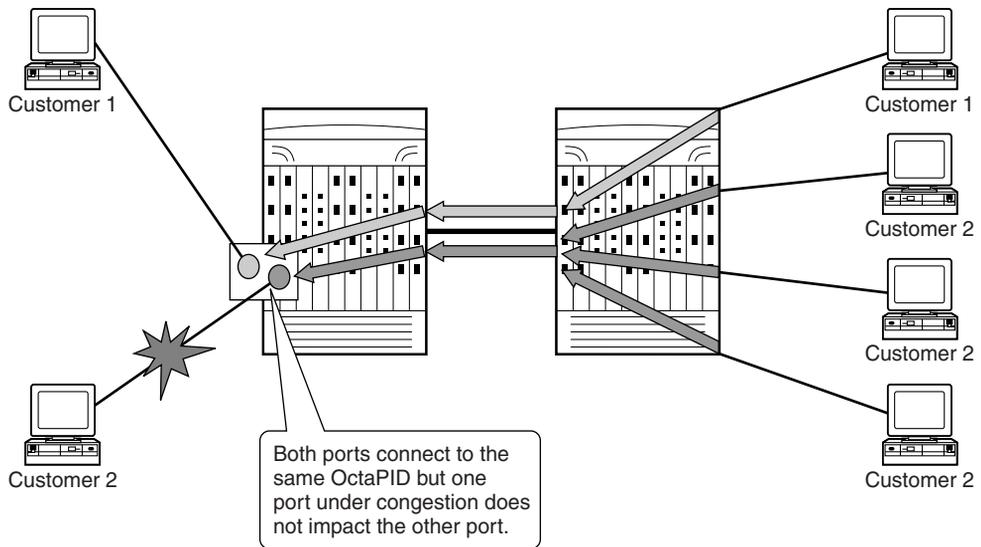
In Figure 130, there are 6 traffic sources, all 100-Base-T links, sending traffic to a single 100-Base-T egress port. Five of these traffic sources are sending constant traffic at different QoS levels (QoS 0 to QoS 5 respectively). Each of these 5 traffic sources is sending traffic at a percentage weight that matches the queue weight for that QoS level.

For example, QoS level 5 has a default queue weight of 31% and hence the traffic source that is sending QoS 5 traffic is set to send traffic at 31% utilization. All the constant traffic sources will add up to 92%. Now, if a sixth traffic source starts sending bursts of traffic such that the link utilization from the bursty source will be 20% during the traffic burst, the total traffic will add up to 92% + 20%, which will equal to 112% during the traffic burst.

If the burst is not a sustained one, then the traffic management and buffering process on the 8600 will allow all the traffic to pass through without any loss of traffic. In this case the bursts were sent at a 1000 packets every 10th of a second.

Note that in the case of the 10/100 I/O module where multiple access ports are shared by the same forwarding engine, each port is essentially assigned its own “logical” queue to ensure that traffic congestion on one port does not affect traffic on another port in the same forwarding engine. This situation is illustrated in [Figure 131](#).

Figure 131 OctaPID queue buffers

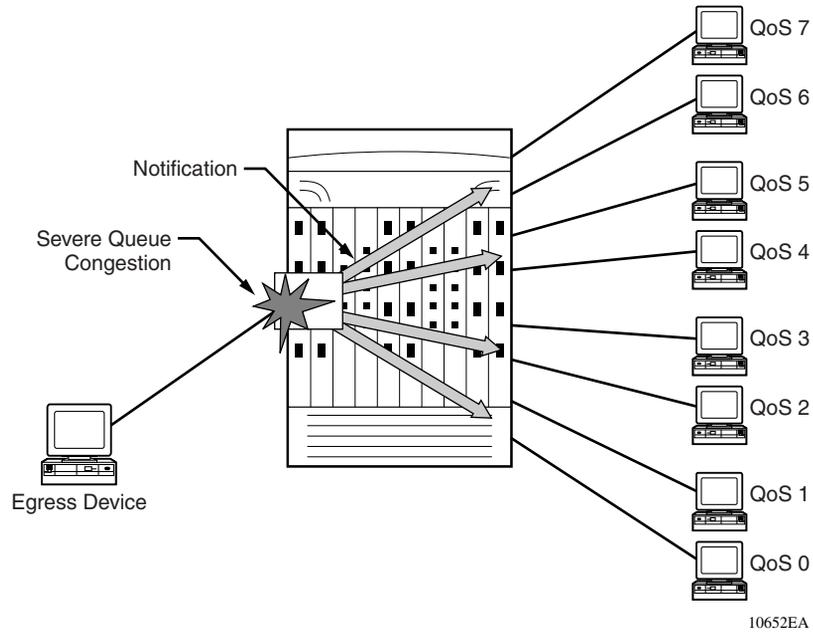


10651EA

Severe congestion

Severe congestion is defined as a condition where the network or certain elements of the network (in some cases it could be a single port) experience a prolonged period of sustained congestion. Under such congestion conditions, congestion thresholds will be reached, buffers will overflow, and a substantial amount of traffic will be lost.

[Figure 132](#) shows an illustration of a very simple example where severe congestion can occur. The diagram is followed by an explanation of how the Passport 8600 handles these scenarios of congestion.

Figure 132 Severe congestion

In [Figure 132](#), all the users on the right-hand side of the graphic are communicating with a single egress device (on the left-hand side). If the combined rate of the traffic to the egress device is sustained over 100% for a prolonged period of time, the egress buffers will overflow, excess traffic will start getting discarded, and *severe congestion* is detected by the Passport 8600.

Upon the detection of severe congestion, the Passport 8600 starts discarding traffic based on drop precedence values. This mode of operation ensures that there is no chance of any high priority traffic being discarded before lower-priority traffic.

When doing traffic engineering and link capacity analysis for a network, the standard design rule is that you should engineer the network links and trunks for a maximum average-peak utilization of no more than 80%. This means that the network will peak up to 100% capacity, but the average-peak utilization should not exceed 80%.

In fact, the network is expected to handle momentary peaks above 100% capacity because of the sophisticated traffic management and traffic-buffering techniques used in high performance switches such as the Passport 8600 as shown in [Figure 130](#).

A network may experience a prolonged period of sustained/severe congestion due to various reasons such as a serious network outage, or bridging loops, or other catastrophic network events.

QoS network scenarios

The sections that follow present QoS network scenarios for bridged and routed traffic over the core network.

Scenario 1 – bridged traffic

Bridging traffic over the core network is typically a service provider type implementation where you keep customer VLANs separate (i.e. the concept of a VPN). Normally, a service provider only implements VLAN bridging (Layer 2), with no routing. This means that the 802.1p determines the level of service assigned to each packet. In those instances where Diffserv is active on the core ports, the level of service received is based on the highest of the Diffserv or 802.1p settings.

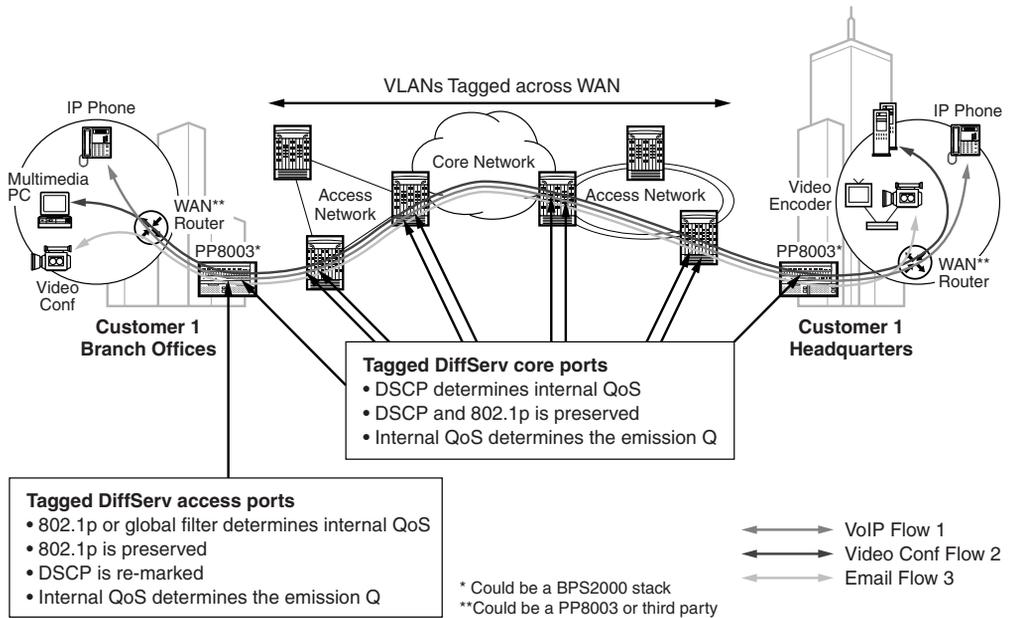
The following cases describe sample QoS design guidelines you can use to provide and maintain high service quality in a Passport 8600 network.

Case 1 – Customer traffic is trusted

In this case, the assumption you make is that for all incoming traffic the QoS setting has been marked properly. All core switch ports are configured for core/trunk ports where they simply read and forward packets. There is no re-marking or re-classification from the switch. All initial QoS markings are done on the customer device, on the edge devices such as PP8003 or BPS2000. (In this case, 8003 treats ingress traffic as trusted).

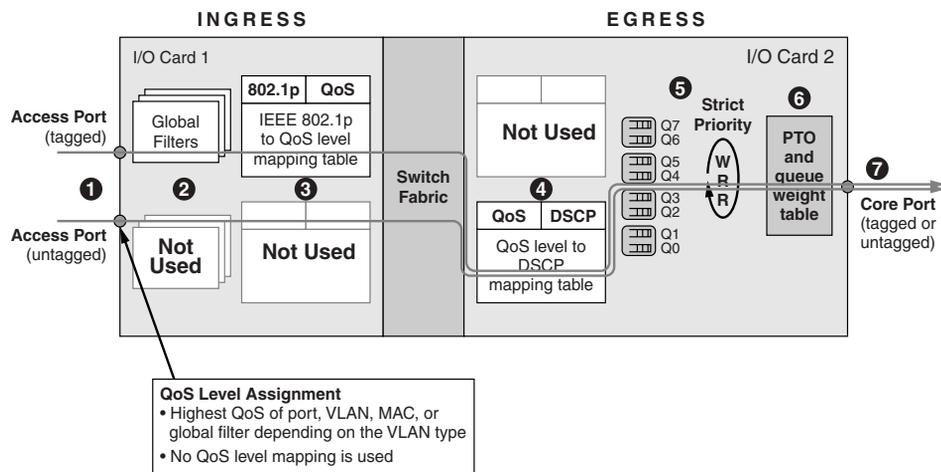
[Figure 133](#) describes the actions performed on 3 different traffic flows (i.e. VoIP, video conference and email) at access and core ports throughout the network.

Figure 133 Trusted bridged traffic



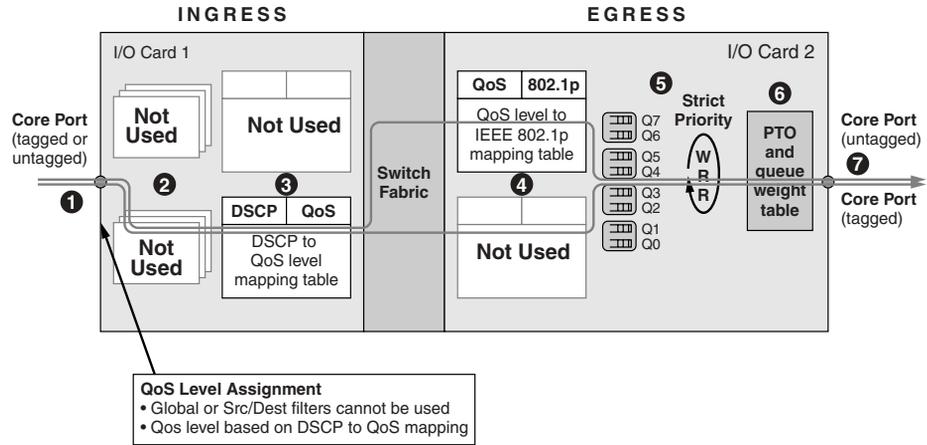
10654EA

Figure 134 shows what happens inside a Passport 8600 access node as packets enter through a tagged or untagged access port and exit through a tagged or untagged core port.

Figure 134 Passport 8600 summary on bridged access ports

10655EA

Figure 135 shows what happens inside a Passport 8600 core node as packets enter through a tagged or untagged core port and exit through a tagged or untagged core port.

Figure 135 Passport 8600 summary on bridged or routed core ports

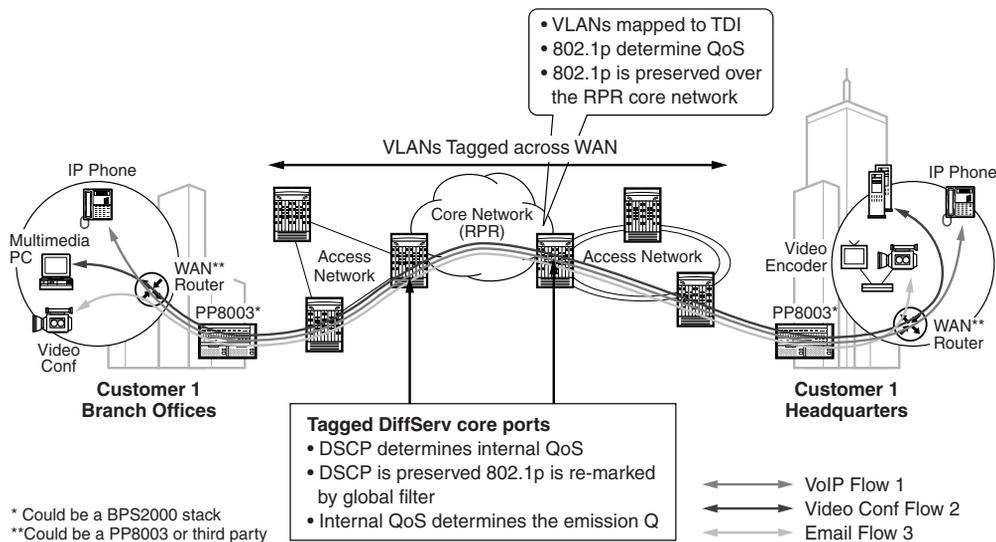
10656EA

Case 2 – Customer traffic is untrusted

For untrusted traffic, the service provider should mark and prioritize customer traffic on the access node using global filters. Service providers would have to re-classify the customer traffic to make sure they comply with the class of service mentioned in the SLA. Once traffic is admitted in the network, it is treated as shown in the Routed traffic over the core network. This is typically an enterprise type implementation where VLANs are not kept separate. The following cases describe sample QoS design guidelines you can use to provide and maintain high service quality in a Passport 8600 network.

Case 3– RPR interworking

In this case, the assumption you make is that for all incoming traffic the QoS setting has been marked properly on the access nodes. The RPR interworking is done on the core switch ports that will be configured for core/trunk ports. These ports will preserve the DSCP and will re-mark the 802.1 p to match the 802.1 p on RPR. [Figure 136](#) shows the actions performed on 3 different traffic flows (i.e. VoIP, video conference and email) over the RPR core network.

Figure 136 Passport 8600 to RPR QoS interworking

10657EA

Scenario 2 – routed traffic

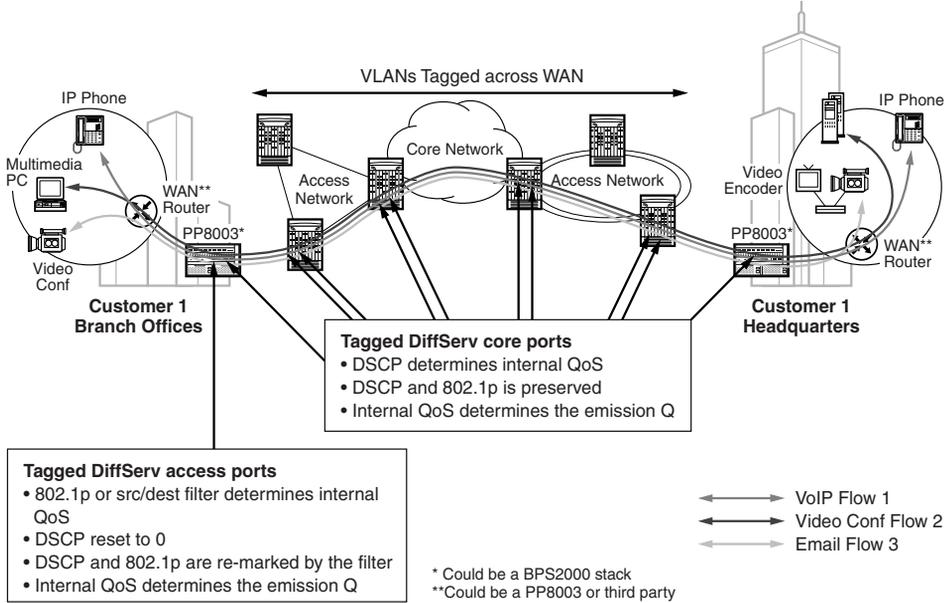
Routed traffic over the core network is typically an enterprise type implementation where VLANs are not kept separate. The following cases describe sample QoS design guidelines you can use to provide and maintain high service quality in a Passport 8600 network.

Case 1 – Customer traffic is trusted

In this case the assumption you make is that for all incoming traffic, the QoS setting has been marked properly. All core switch ports are configured for core/trunk ports where they will simply read and forward packets. There is no re-marking or re-classification from the switch. All initial QoS markings are done on the customer device, on the edge devices, such as the Passport 8003 or BPS2000 (in this case, the 8003 treats ingress traffic as trusted).

Figure 137 shows the actions performed on 3 different traffic flows (i.e. VoIP, video conference and email) at access and core ports throughout the network.

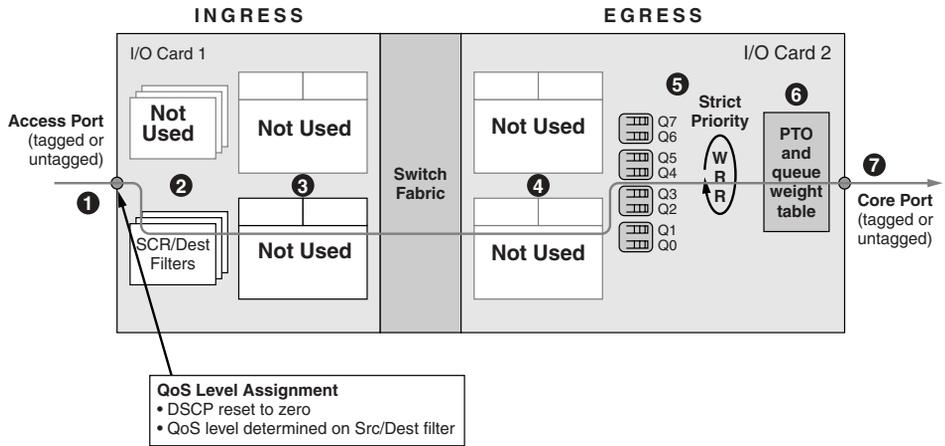
Figure 137 Trusted routed traffic



10658EA

Figure 138 shows what happens inside a Passport 8600 access node as packets enter through a tagged or untagged access port and exit through a tagged or untagged core port.

Figure 138 Passport 8600 QoS summary on routed access ports



10659EA

Chapter 10

Managing Passport 8000 Series switches

This chapter provides design guidelines that will assist you in managing your Passport 8000 switch. It includes the following system management topics:

Topic	Page number
Offline switch configuration	next
Port mirroring	388
pcmbboot.cfg	391
Default management IP address	392
Backup configuration files	392
DNS client	392

Offline switch configuration

The ASCII configuration file of the Passport 8000 Series switch lists configuration commands in specific order, listing MLT information, for example before STG information.

To avoid boot problems, save your running configuration to a flash card and edit the configuration file offline. To save the configuration file in the CLI, enter the following command:

```
save config file <filename>
```

The filename can include the directory structure.

When you finish editing the configuration file, you can upload it to the switch. Use the `source` command to load the changed configuration onto the running switch.

Port mirroring

Port mirroring is a diagnostic tool that can be used for troubleshooting and performing network traffic analysis. When using port mirroring, you have to specify a destination port to see mirrored traffic and specify the source ports from which traffic is mirrored. Unlike other methods used to analyze packet traffic, packets flow normally through the destination port and packet traffic is uninterrupted.

Port mirroring can be divided into two overall categories, local and remote mirroring. Each of these is described in the sections that follow.

Local port mirroring

You can configure the Passport 8000 Series switch to monitor ingress traffic on a port and, in some cases, to monitor egress traffic from a port.

Port mirroring considerations are as follows:

- Ingress mirroring mirrors only packets with valid CRCs
- The Passport 8100 switch:
 - Supports ingress and egress port mirroring
 - Supports egress mirroring in *only* half-duplex mode
- The Passport 8600 switch:
 - Supports ingress mirroring on all modules
 - Supports egress mirroring *only* on Passport E- or M-modules

Identifying E-modules

You can identify Passport 8600 Series E-modules by the letter “e” at the end of the module name. See [Table 34](#).

Table 34 Passport 8600 E-modules

Module name	Part number
Passport 8648TXE	DS1404035
Passport 8608SXE	DS1404036
Passport 8608GBE	DS1404038
Passport 8608GTE	DS1404044
Passport 8624FXE	DS1404037
Passport 8672ATME	DS1304008
Passport 8683POSE	DS1404043

Mirroring scalability

Prior to release 3.2:

- 10 mirrored ports to 1 mirroring port

From release 3.2.2 onwards (including 3.3):

On the Passport 8100:

- 25 mirrored ports to 1 port. Note that Ethernet MDAs can also be a part of these 25 ports
- Ingress and egress mirroring
- Half-Duplex ONLY for egress mirroring

On the Passport 8600:

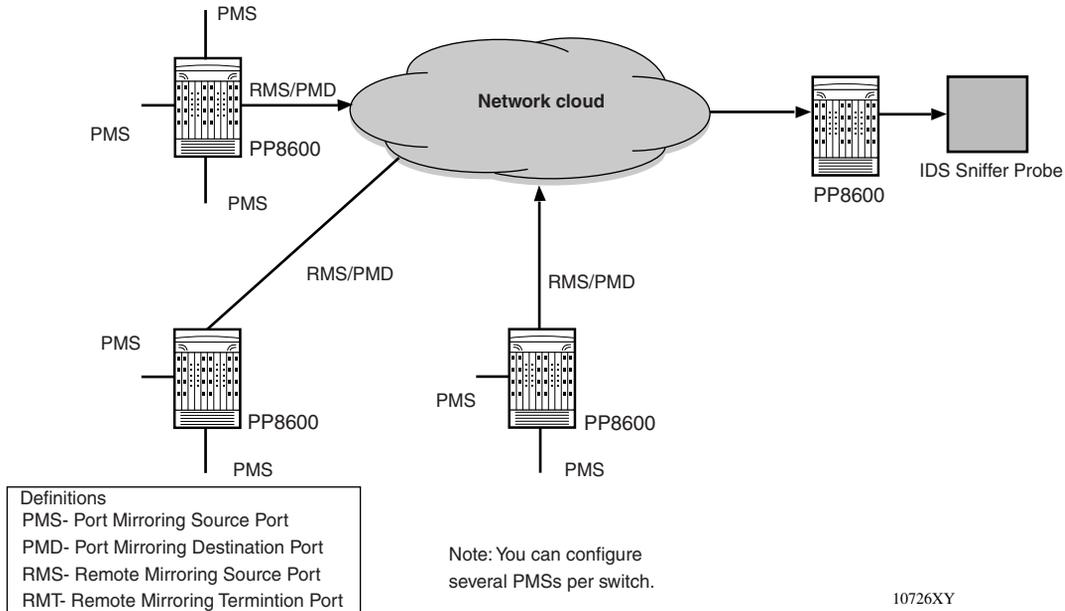
- You can now configure a number of port mirroring entries between 1 and 383, with ALL entries simultaneously active. Note that the number of mirroring ports plus the number of mirrored ports cannot exceed 384.
- You can mirror traffic based on ingress, egress, both, or based on a filter rule (TX, RX, TX&RX, RxFilter). Note that egress mirroring works only on E- or M-modules.
- You can have a mix of TX mirroring, RX mirroring and both (TX + RX = total traffic) on a chassis at the same time.

- You can create a maximum of 100 entries with only 25 entries active at a time. Nortel Networks recommends that you disable mirroring when not in use in order to reduce the load on the switch. (Note that the mirrored traffic has the lowest priority).
- You can mirror multiple active entries to multiple destination ports. However, some hardware limitations apply including:
 - You can only mirror ports supported by the same forwarding engine (group of 8 10/100 ports or 1 Gig port) to the same destination.
 - You cannot mirror one port to multiple destinations.
 - You can configure a maximum of 25 destination ports at one time without violating the hardware limitations. Note that checks have been put in place to inform you if a mirroring rule is broken.

Remote mirroring

Remote mirroring allows you to steer mirrored traffic through a switch cloud to a network analysis probe located on a remote switch. In a network, you can use remote mirroring to monitor many ports from different switches using one network probe device.

With the 3.7 release, the Passport 8600 switch offers a way to reduce the cost of ownership (i.e., traffic analyzer, external probes) by centralizing these relatively expensive devices ([Figure 139](#)). For more information on remote mirroring, see *Configuring Network Management*.

Figure 139 Remote mirroring

pcmboot.cfg

The 3.7 release introduces the *pcmboot.cfg* file for the Passport 8600. When you wish to deploy a large network where several locations are remote, it is not always possible to have a trained employee on-site. With the *pcmboot.cfg* file, you can:

- 1 Send the configuration file by email.
- 2 Copy it to a regular PCMCIA card (as supported by the Passport 8600) using a PC or laptop.
- 3 Plug the PCMCIA card into the switch and boot it.



Note: Ensure that the *pcmboot.cfg* file is the same size before and after sending it through email. If the file is corrupted, the Passport 8600 may use the factory default configuration instead.

If the switch detects the *pcmbboot.cfg* file on the PCMCIA card, it uses this file to boot rather than the *boot.cfg* file on the flash. Once the switch has booted, it is possible for a remote network manager to fully control the switch.

Default management IP address

If the *boot.cfg* file is not present on the flash, the network management port (also referred to as the Out-of-Band [OOB] interface) is assigned a default IP address (192.168.168.168/24 for slot 5 or slot 3 and 192.168.168.169 for slot 6). If you have already configured an IP address for this interface in the *boot.cfg*, the switch uses this address.

Backup configuration files

The Passport 3.7.0 release introduces a backup file feature that allows network managers to rely on a specific configuration file if one of the configuration files is faulty. Thus, the switch does not need to use the default configuration file.

This feature is particularly useful in some carrier environments where the Passport 8600 is shared among several customers. In such environments, it is not acceptable to see some traffic shared between different users if the switch reboots with the factory default configuration.

DNS client

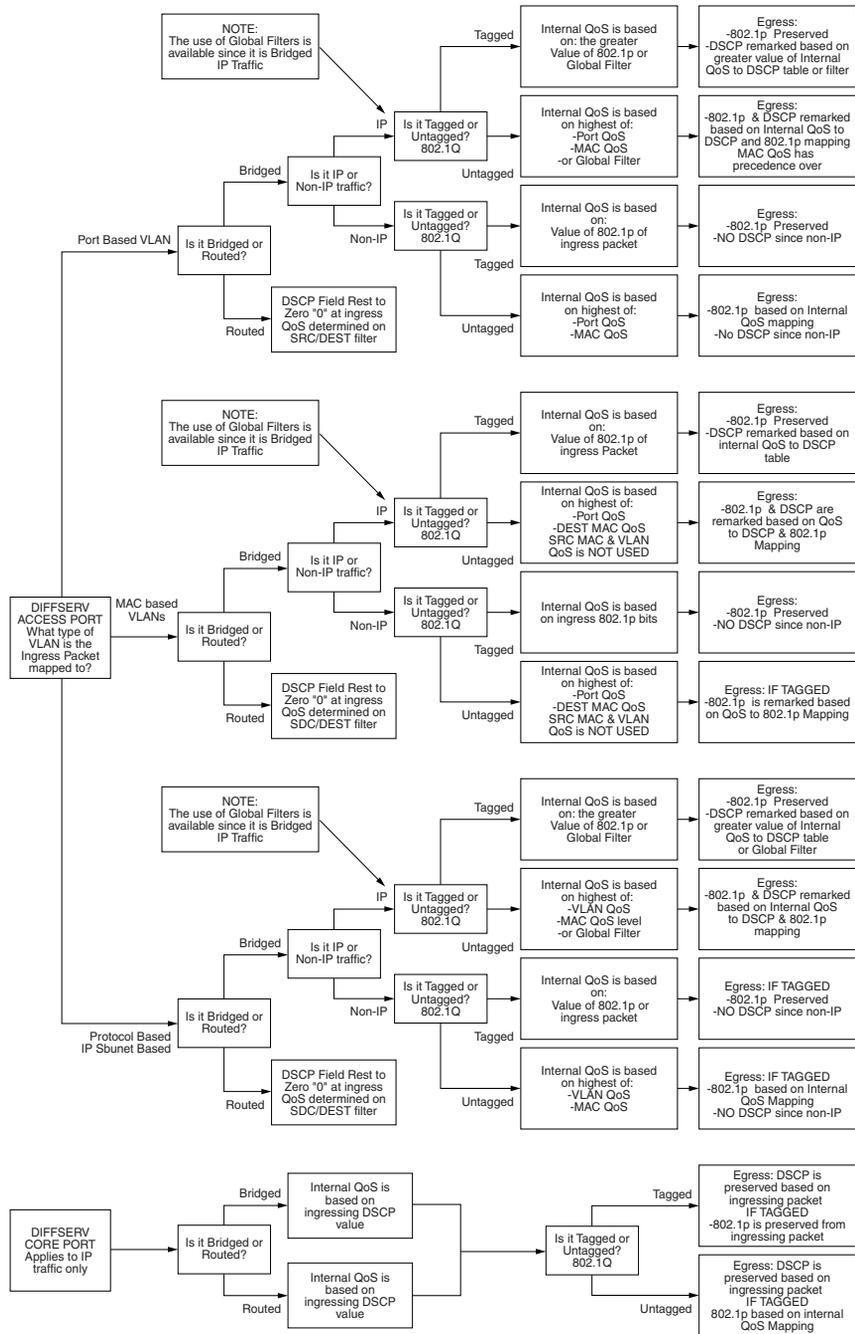
The Passport 8600 3.7 release allows network managers to configure up to 3 DNS servers, thus allowing managers to use names rather than IP addresses. If the DNS servers are not available, the switch relies on a local file */etc/hosts* to find the relationship between the name and the IP address.

Appendix A

QoS algorithm

This appendix provides an illustration of the QoS algorithm. [Figure 140](#) shows the algorithm and the decision-making process it encompasses.

Figure 140 QoS algorithm



10662EA

Appendix B

Scaling numbers

This appendix provides scaling numbers for the Passport 8000 Series switch release 3.7 features, including E- and M-modules. Be aware that [Table 35](#) lists the current values for this release. These capabilities are enhanced in subsequent software releases. Thus, Nortel Networks recommends that you consult your copy of the Release Notes for the latest values.



Note: The capabilities described in [Table 35](#) are supported as individual protocols.

Table 35 Scaling numbers for Release 3.7 features

Feature	Maximum number supported
Hardware records	Non E / E Modules: 25 000 records M Modules : 125 000 records ¹
M Modules	Nortel Networks strongly recommends using 8691SFs or 8692SFs with M Modules
10GE	Release 3.7 does NOT support the combination of the following features and the 10GE Module: <ul style="list-style-type: none"> - IPX routing - SMLT - External MLT (Nortel Networks recommends that you use a Layer 3 routing protocol for resiliency, like OSPF, associated to ECMP, Equal Cost Multi Path) - Egress Mirroring Due to the internal architecture, Nortel Networks strongly recommends using 2 8691SFs/8692SFs per system using a 10GE Module (internal MLT of 8 Gig ports) for load sharing and redundancy.

Table 35 Scaling numbers for Release 3.7 features (continued)

Feature	Maximum number supported
VLANs	Passport 8100: 2013 Passport 8600: 1980
IP subnet based VLANs (Passport 8600 only)	200
IP Interfaces (Passport 8600 only)	<ul style="list-style-type: none"> • 500 (default) • 1980 (requires order number DS1411015: Passport 8000 Chassis MAC Address Upgrade Kit. License for reprogramming the chassis to a block of 4096 addresses for routed VLAN scaling)
RIP Routes (Passport 8600 only)	2500
OSPF Areas per switch (Passport 8600 only)	5
OSPF Adjacencies per switch (Passport 8600 only)	80
OSPF Routes per switch (Passport 8600 only)	Non E / E modules: 15 000 M Module: 20 000
BGP (Passport 8600 only)	Number of peers: 10 Number of routes: - Non E / E Modules : 20 000 - M Modules : 119 000
DVMRP Interfaces (Passport 8600 only)	500
DVMRP Routes (Passport 8600 only)	2500
PIM Interfaces (Passport 8600 only)	500
Multicast source subnet trees (Passport 8600 only)	500
Multicast (S,G) DVMRP	2000
Multicast (S,G) PIM	2000
IPX Interfaces (Passport 8600 only)	100
IPX RIP (Passport 8600 only)	5000
IPX SAP (Passport 8600 only)	7500
VRRP Interfaces (Passport 8600 only)	255
Spanning Tree Groups	Passport 8600: 25 ² Passport 8100: 1
Aggregation Groups - IEEE 802.3ad aggregation groups - Multi Link Trunking group (MLT)	Passport 8600: 32 Passport 8100: 6 Redirection: 3

Table 35 Scaling numbers for Release 3.7 features (continued)

Feature	Maximum number supported
Ports per MLT Note: all the ports MUST be of the same type (no mix of technology will be supported)	Passport 8600: up to 8 Passport 8100: up to 4
Permanent virtual circuits scaling (ATM)	Passport 8600 and Passport 8100: up to 500 permanent virtual circuits (PVCs) per chassis. <ul style="list-style-type: none"> • 256 RFC1483 bridged/routed ELANs per MDA • 500 RFC1483 bridged/routed ELANs per switch (12 more RFC1483 bridged ELANs per switch can be configured) • 64 PVCs per RFC1483 bridged ELAN • 1 PVC per RFC 1483 routed ELAN

- 1 The exact number is 125838. 2162 records are used by the system. With the record reservation feature, 8K records are pre allocated (see the documentation for more information) for some specific types of traffic (for example, MAC and ARP).
- 2 Nortel Networks supports only 25 STGs with Release 3.7. You can configure up to 64 (63 with the WSM Module) STGs, but configurations including more than 25 STGs will not be supported. If you do need more than 25 STGs, contact your Nortel Networks Sales Representative for more information about the support of this feature. With Release 3.7 (8600) and 10.0 (WSM), the WSM Module supports the tagged BPDU from the 8600 only with the default STG (STG ID 1).

Passport 8000 Series Software Release 3.7 does not support configurations of Passport 8100 modules and Passport 8600 modules simultaneously within the same chassis.

The Web Switching Module is not supported in Passport 8100 nor in 8100 module configurations.

- 3 The number of aggregation groups decreases when you install a WSM module into the chassis. Refer to the WSM configuration manual for more information about how to connect through the backplane and the logical configuration (VLAN/STGs)

Appendix C

Hardware and supporting software compatibility

The matrix below describes your hardware and the minimum Passport 8000 Switch Series software version required to support the hardware.

Chassis & switching fabric		Minimum software version	Part number
8010co chassis	10-slot chassis	3.1.2	DS1402004
8010 chassis	10-slot chassis	3.0.0	DS1402001
8006 chassis	6-slot chassis	3.0.0	DS1402002
8003 chassis	3-slot chassis	3.1.2	DS1402003
8690SF	Discontinued, see 8691SF	3.0.0	DS1404001
8691SF	Switching fabric	3.1.1	DS1404025

Upgrade Kits

256MB CPU upgrade kit	This memory upgrade may be required for the 3.5 software to run properly. ¹	3.5	DS1404016
MAC upgrade kit	Use this kit to add Media Access Control (MAC) addresses to your system.	3.5	DS1404015

¹ With the Passport 8000 Switch Series Software Release 3.5 you must upgrade the 8690SF to 256MB. Nortel Networks recommends that you upgrade the 8691SF to 256MB.

8600 modules and components		Minimum software version	Part number
Security module			
8661SSL Acceleration Module (SAM)	High Performance SSL Acceleration Module secures web-based applications and business communications ¹	3.3.1 ²	DS1404070
Layer 4-7 module			
Web Switching Module (WSM)	4-Port Gigabit Ethernet SX or 10/100TX	3.1.3 ³ , 3.2.1 ⁴ , 3.3.0 ⁵	DS1404045

8600 modules and components (continued)		Minimum software version	Part number
Ethernet modules			
8608GB module	Discontinued, see 8608GBE module	3.0.0	DS1404015
8608GT module	Discontinued, see 8608GTE module	3.1.0	DS1404012
8608SX module	Discontinued, see 8608SXE module	3.0.0	DS1404003
8624FX module	Discontinued, see 8624FXE module	3.0.0	DS1404005
8648TX module	Discontinued, see 8648TXE module	3.0.0	DS1404002
Ethernet E-modules:⁶			
8608GBE module	8-port Gigabit Ethernet GBIC	3.1.1	DS1404038
8608GTE module	8-port Gigabit Ethernet 1000TX	3.1.1	DS1404044
8608SXE module	8-port Gigabit Ethernet SX	3.1.1	DS1404036
8616SXE module	16-port Gigabit Ethernet SX	3.1.0	DS1404011
8616GTE module	16-port Gigabit Ethernet TX	3.3.0	DS1404034
8624FXE module	24-port 100FX ⁷	3.1.1	DS1404037
8648TXE module	48-port 10/100 TX	3.1.1	DS1404035
8632TXE module	32-port 10/00TX (2 GBICs)	3.1.2	DS1404024
Ethernet M-modules⁸			
8608GBM module	8-port Gigabit Ethernet GBIC	3.3.0	DS1404059
8608GTM module	8-port Gigabit Ethernet 1000TX	3.3.0	DS1404061
8632TXM module	32-port 10/00TX (2 GBICs)	3.3.0	DS1404055
8648TXM module	48-port 10/100 TX	3.3.0	DS1404056
10 Gigabit Ethernet M modules⁸			
8681XLW module	1-port 10 Gigabit Ethernet (1310nm WAN serial)	3.3.0	DS1404052
8681XLR module	1-port 10 Gigabit Ethernet (1310nm LAN serial)	3.3.0	DS1404053
ATM/ATME/ATMM modules			
8672ATM module	Discontinued, see 8672ATME module	3.1.0	DS1304001
8672ATME module	ATME module ⁶ .	3.1.1	DS1304008
8672ATMM module	ATMM module ⁸ .	3.3.0	DS1304009

8600 modules and components (continued)		Minimum software version	Part number
ATM/ATME/ATMM module components⁹			
DS-3 MDA	2-port 75 ohm coaxial	3.3.0	DS1304002
OC-12c/STM-4 MDA	1-port MMF	3.1.0, 3.1.1, 3.3.0'	DS1304004
OC-12c/STM-4 MDA	1-port SMF	3.1.0, 3.1.1, 3.3.0'	DS1304005
OC-3c/STM-1 MDA	4-port MMF	3.1.0, 3.1.1, 3.3.0'	DS1304006
OC-3c/STM-1 MDA	4-port SMF	3.1.0, 3.1.1, 3.3.0'	DS1304007
POS/POSE/POSM modules			
8683POS module	Discontinued, see 8683POSM module	3.1.0	DS1404016
8683POSE module	Discontinued, see 8683POSM module	3.1.1	DS1404043
8683POSM module	M module ⁸	3.3.0	DS1404060
POS/POSE/POSM MDAs¹⁰			
OC-3c/STM-1 MDA	2-port MMF	3.1.0, 3.1.1, 3.3	DS1333003
OC-3c/STM-1 MDA	2-port SMF	3.1.0, 3.1.1, 3.3	DS1333004
OC-12c/STM-4 MDA	1-port MMF	3.1.0, 3.1.1, 3.3	DS1333001
OC-12c/STM-4 MDA	1-port SMF	3.1.0, 3.1.1, 3.3	DS1333002
8600 compatible GBICs¹¹			
1000BASE-SX GBIC	850 nm, short wavelength, Gigabit Ethernet	3.0.0	AA1419001
1000BASE-LX GBIC	1300 nm, long wavelength, Gigabit Ethernet	3.0.0	AA1419002
1000BASE-T GBIC	Category 5 copper unshielded twisted pair (UTP)	3.5.0	AA1419041
1000BASE-XD GBIC	50k, SC duplex SMF, Gigabit Ethernet	3.0.0	AA1419003
1000BASE-ZX GBIC	70k, SC duplex SMF, Gigabit Ethernet	3.0.0	AA1419004
Gray CWDM GBIC	Discontinued, see Gray CWDM APD GBIC	3.1.2	AA1419005
Violet CWDM GBIC	Discontinued, see Violet CWDM APD GBIC	3.1.2	AA1419006
Blue CWDM GBIC	Discontinued, see Blue CWDM APD GBIC	3.1.2	AA1419007
Green CWDM GBIC	Discontinued, see Green CWDM APD GBIC	3.1.2	AA1419008
Yellow CWDM GBIC	Discontinued, see Yellow CWDM APD GBIC	3.1.2	AA1419009

8600 modules and components (continued)		Minimum software version	Part number
Orange CWDM GBIC	Discontinued, see Orange CWDM APD GBIC	3.1.2	AA1419010
Red CWDM GBIC	Discontinued, see Red CWDM APD GBIC	3.1.2	AA1419011
Brown CWDM GBIC	Discontinued, see Brown CWDM APD GBIC	3.1.2	AA1419012
Gray CWDM APD GBIC	1470nm	3.1.4	AA1419017
Violet CWDM APD GBIC	1490nm	3.1.4	AA1419018
Blue CWDM APD GBIC	1510nm	3.1.4	AA1419019
Green CWDM APD GBIC	1530nm	3.1.4	AA1419020
Yellow CWDM APD GBIC	1550nm	3.1.4	AA1419021
Orange CWDM APD GBIC	1570nm	3.1.4	AA1419022
Red CWDM APD GBIC	1590nm	3.1.4	AA1419023
Brown CWDM APD GBIC	1610nm	3.1.4	AA1419024

- 1 The 8661 SAM is used in conjunction with the Web Switching Module to intelligently accelerate secure business communication and confidential data by off-loading Secure Sockets Layer (SSL) Processing.
- 2 The 8661 SAM and Web Switching Module security solution also require WebOS version 10.0.27.3 or newer. Passport 8000 Series Switch Series Software Release 3.3.1 was specifically designed to introduce the 8661 SAM module. Release 3.3.2 does not support this module.
- 3 Passport 8000 Switch Series Software Release 3.1.3 is the first and only release in the 3.1.x software branch that supports the Web Switching Module.
- 4 Passport 8000 Switch Series Software Release 3.2.1 (and later) supports the Web Switching Module.
- 5 Passport 8000 Switch Series Software Release 3.3.0 introduced support for WebOS 10.0 on the Web Switching Module.
- 6 E-modules are required to support Egress port mirroring and Enhanced Operational Mode.
- 7 The 8624FX and the 8624FXE modules support only full-duplex, and cannot connect to half-duplex devices.
- 8 M modules offer additional memory to support large routing tables such as those found in BGP implementations. The Passport 8000 Series Software Release 3.3 introduced a new mode, called M Mode, or 128K records mode, which requires the 8691SF module. When enabled, this mode allows M modules to use their full capabilities (128K records). When this mode is disabled, the M modules work in 32K mode (case of non E and E modules). If the chassis includes one 8690SF module, the mode default backs to 32K mode. To be effective, this mode requires that all modules installed in the same chassis support 128K records (M modules) and that the SF/CPU are 8691SF. If one, or more modules installed in the chassis is a 32K records module (non M module), these modules will be disabled if the chassis is configured to operate in M Mode (For instructions on enabling MMode, see *Managing Platform Operations and Using Diagnostic Tools*).
- 9 ATM MDAs inserted into an 8672ATME module require Passport 8000 Switch Series Software Release 3.1.1 or higher. ATM MDAs inserted into a 8672ATMM module require Passport 8000 Switch Series Software Release 3.3.0 or higher.
- 10 POS MDAs inserted into a 8683POSE module require Passport 8000 Switch Series Software Release 3.1.1 or higher, and POS MDAs inserted into a 8683POSM module require Passport 8000 Switch Series Software Release 3.3.0 or higher.
- 11 Non-supported GBICs are displayed in the CLI and Device Manager as “GBIC-other.”

Appendix D

Tap and OctaPID assignment

The switch fabric in the Passport 8600 modules has nine switching taps, one for each of the eight I/O slots (1 to 4 and 7 to 10) and one for the CPU slots (5 and 6). Taps 0-7 map to the eight I/O slots and can support up to eight OctaPIDs. Each OctaPID can support up to eight ports.

In the Passport 8000 Series switch, a physical port number is 10 bits long and has the following format:

```
9   6 5   3 2   0
+---+---+---+
|   |   |   |
+---+---+---+
```

bits 9–6: Tap number (0–15)

bits 5–3: OctaPID number (0–7)

bits 2-0: MAC port number (0-7)

The Tap number bits and the OctaPID number bits combined (bits 9–3) are usually referred to as the OctaPID ID.

[Table 36](#) lists the module types that are currently available, along with the associated OctaPID ID assignments for each module.

Table 36 Available module types and OctapPID ID assignments

Module type	Port type	OctaPID ID assignment
8608GBE and 8608GBM Modules	1000BASE-SX (GBIC)	Table 37 next
	1000BASE-LX (GBIC)	
	1000BASE-ZX (GBIC)	
	1000BASE-XD (GBIC)	
	1000BASE-TX (GBIC)	
8608GTE and 8608GTM Modules	1000BASE-T	Table 37 next
8608SXE Module	1000BASE-SX	Table 37 next
8616SXE Module	1000BASE-SX	Table 38 on page 405
8624FXE Module	100BASE-FX	Table 39 on page 406
8632TXE and 8632TXM Modules	10BASE-T/100BASE-TX	Table 40 on page 406
	1000BASE-SX (GBIC)	
	1000BASE-LX (GBIC)	
	1000BASE-ZX (GBIC)	
	1000BASE-XD (GBIC)	
8648TXE and 8648TXM Modules	10/100 Mb/s	Table 41 on page 406
8672ATME and 8672ATMM Modules	OC-3c MDA	Table 42 on page 407
	OC-12c MDA	
	DS3	
8681XLR Module	10GBASE-LR	Table 43 on page 407
8681XLW Module	10GBASE-LW	Table 44 on page 408
8683POSM Module	OC-3c MDA	Table 45 on page 408
	OC-12c MDA	

Table 37 describes the OctaPID ID and port assignments for the 8608GBE, Passport 8608GBM, 8608GTE, 8608GTM, and 8608SXE modules.

Table 37 8608GBE/8608GBM/8608GTE/8608GTM, and 8608SXE modules

OctaPID ID assignment	Port assignment
OctaPID ID: 0	Port 1
OctaPID ID: 1	Port 2
OctaPID ID: 2	Port 3
OctaPID ID: 3	Port 4
OctaPID ID: 4	Port 5
OctaPID ID: 5	Port 6
OctaPID ID: 6	Port 7
OctaPID ID: 7	Port 8

Table 38 describes the OctaPID ID and port assignments for the 8616SXE Module.

Table 38 8616SXE module

OctaPID ID assignment	Port assignment
OctaPID ID: 0	Ports 1 and 2
OctaPID ID: 1	Ports 3 and 4
OctaPID ID: 2	Ports 5 and 6
OctaPID ID: 3	Ports 7 and 8
OctaPID ID: 4	Ports 9 and 10
OctaPID ID: 5	Ports 11 and 12
OctaPID ID: 6	Ports 13 and 14
OctaPID ID: 7	Ports 15 and 16

[Table 39](#) describes the OctaPID ID and port assignments for the 8624FXE Module.

Table 39 8624FXE module

OctaPID ID assignment	Port assignment
OctaPID ID: 0	Ports 1 through 8
OctaPID ID: 1	Ports 9 through 16
OctaPID ID: 2	Ports 17 through 24

[Table 40](#) describes the OctaPID ID and port assignments for the 8632TXE and 8632TXM Modules.

Table 40 8632TXE and 8632TZX modules

OctaPID ID assignment	Port assignment
OctaPID ID: 0	Ports 1 through 8
OctaPID ID: 1	Ports 9 through 16
OctaPID ID: 2	Ports 17 through 24
-	-
-	-
OctaPID ID: 5	Ports 25 through 32
OctaPID ID: 6	Port 33 (GBIC port)
OctaPID ID: 7	Port 34 (GBIC port)

[Table 41](#) describes the OctaPID ID and port assignments for the 8648TXE and 8648TXM Modules.

Table 41 8648TXE and 8648TXM modules

OctaPID ID assignment	Port assignment
OctaPID ID: 0	Ports 1 through 8
OctaPID ID: 1	Ports 9 through 16
OctaPID ID: 2	Ports 17 through 24
-	-
-	-

Table 41 8648TXE and 8648TXM modules

OctaPID ID assignment	Port assignment
OctaPID ID: 5	Ports 25 through 32
OctaPID ID: 6	Port 33 through 40
OctaPID ID: 7	Port 41 through 48

[Table 42](#) describes the OctaPID ID and port assignments for the 8672ATME and 8672ATMM Modules.

Table 42 8672ATME and 8672ATMM modules

OctaPID ID assignment	Port assignment
OctaPID ID: 0	<ul style="list-style-type: none"> • Ports 1 through 4 (with OC-3c MDA) • Port 1 (with OC-12c MDA) • Ports 1 through 2 (with DS-3 MDA)
OctaPID ID: 1	<ul style="list-style-type: none"> • Ports 5 through 8 (with OC-3c MDA) • Port 5 (with OC-12c MDA) • Ports 5 through 6 (with DS-3 MDA)
OctaPID ID: 2	Not used

[Table 43](#) describes the OctaPID ID and port assignments for the 8681XLR Module.

Table 43 8681XLR module

OctaPID ID assignment	Port assignment
OctaPID ID: 0	Port 1
OctaPID ID: 1	
OctaPID ID: 2	
OctaPID ID: 3	
OctaPID ID: 4	
OctaPID ID: 5	
OctaPID ID: 6	
OctaPID ID: 7	

Table 44 describes the OctaPID ID and port assignments for the 8681XLW Module.

Table 44 8681XLW module

OctaPID ID assignment	Port assignment
OctaPID ID: 0	Port 1
OctaPID ID: 1	
OctaPID ID: 2	
OctaPID ID: 3	
OctaPID ID: 4	
OctaPID ID: 5	
OctaPID ID: 6	
OctaPID ID: 7	

Table 45 describes the OctaPID ID and port assignments for the 8683POSM Module.

Table 45 8683POSM module

OctaPID ID assignment	Port assignment
OctaPID ID: 0	<ul style="list-style-type: none"> • Ports 1 and 2 (with OC-3c MDA) • Port 1 (with OC-12c MDA)
OctaPID ID: 1	<ul style="list-style-type: none"> • Ports 3 and 4 (with OC-3c MDA) • Port 3 (with OC-12c MDA)
OctaPID ID: 2	<ul style="list-style-type: none"> • Ports 5 and 6 (with OC-3c MDA) • Port 5 (with OC-12c MDA)

Index

Numbers

100BaseFX drivers without FEFI support 60

802.1d Spanning Tree 130

incomplete connectivity issues 130

using multiple STP instances 130

802.1q tags 329

A

ATM

802.1q tags 329

and voice applications 339

ATM latency testing results 340

design recommendations 340

DiffServ core port 330

IGMP fast leave 332

IP multicast guidelines 330

mapping QoS to class of service 330

MLT guidelines 329

network configuration guidelines 325

point to multipoint mode 331

resiliency 327

F5 OAM loopback request/reply 328

MLT and SMLT 327

scalability 325

maximum ELANs, PVCs and VLANs 326

maximum supported modules 325

throughput 326

video over DSL 338

WAN connectivity 333

OE/ATM interworking 335

point to point 333

service provider solutions 334

Auto-Negotiation 59

Parallel Detection function 59

recommended settings (10/100BASE-TX
ports) 60

unsupported products 62

B

BGP

definition of EBGP 165

definition of IBGP 165

design scenarios 166

edge aggregation 167

internet peering 166

ISP segmentation 168

mergers and acquisitions 167

hardware and software dependencies 164

multi-homed, non-transit AS 170

scaling considerations 165

BGP peering 165

BGP route management 165

BPDUs 130

C

control packet rate limit 95

CP-Limit 95

CPU

control packet rate limit 95

failover 69

functions 68

HA mode 69

SMLT recommendations 68

customer support 34

D

designated router (DR) 274

- designing redundant networks
 - SMLT
 - IST link- IST VLAN and peer IP configuration 96
- designing Ethernet networks
 - applications considerations 332
 - ATM and voice applications 339
 - ATM WAN connectivity 333
 - TLS 337
 - video over DSL over ATM 338
 - engineering considerations 325
 - ATM resiliency 327
 - ATM scalability 325
 - feature considerations 329
 - ATM and 802.1q tags 329
 - ATM and DiffServ 330
 - ATM and IP multicast 330
 - ATM and MLT 329
 - shaping 332
- designing L3 switched networks
 - BGP 163
 - ICMP redirect messages 152
 - IP routed interface scaling 182
 - IPX 180
 - OSPF 172
 - subnet-based VLANs 155
 - VRRP 150
- designing multicast networks
 - DVMRP design rules 240
 - general IP multicast rules 226
 - general PIM guidelines 248
 - IGMP and routing protocol interactions 237
 - IP multicast
 - scaling 221
 - IP multicast and MLT 216
 - IP multicast and SMLT 266
 - multicast handling 215
- designing redundant networks
 - 802.1d Spanning Tree 130
 - incomplete connectivity issues 130
 - using multiple STP instances 130
 - general considerations 53
 - reliability and availability 54
- isolated VLANs 132
 - configuration guidelines 132
 - STP blocking 132
- link redundancy 71
- MLT 71
- network design examples 115
 - Layer 1 examples 115
 - Layer 2 examples 118
 - Layer 3 examples 121
- network redundancy 86
 - basic layouts (physical structure) 86
 - recommended/not recommended network edge design 91
 - redundant network edge 90
- physical layer 55, 112
 - Auto-Negotiation 59
 - configuring SFFD using the CLI 64
 - Ethernet cable distances 56
 - FEFI 60
 - Gigabit and remote fault indication 61
 - RSMLT- designing and configuring 115
 - RSMLT- failure scenarios 113
 - RSMLT- operation in L3 environments 112
 - SFFD configuration rules 64
 - using SFFD for remote fault indication 62
 - VLACP 64
 - VLACP- with SMLT 66
- platform redundancy 67
 - configuration and image redundancy 68
 - CPU redundancy 68
 - I/O port redundancy 67
 - redundant power supplies 67
 - switch fabric redundancy 67
- PVST+ 129
 - BPDU's 130
- single port SMLT 98
- SMLT 91
 - and STP 110
 - configuration 93
 - designs 106
 - designs- full mesh configuration 109
 - designs- scaling 106
 - designs- square configuration 108
 - designs- triangle configuration 107

- failure scenarios 104
- ID configuration 98
- IST link 94
- IST link- dual purpose comm channel 94
- IST link- failure avoidance 95
- IST link- supported links 96
- IST link- using CP-Limit with 95
- Layer 2 traffic load sharing 103
- Layer 3 traffic load sharing 103
- scalability 110
- scalability- IST/SMLT scalability 111
- scalability- MAC address scalability 111
- scalability- SMLT and multicast
 - scalability 111
- scalability- VLAN scalability on MLT and
 - SMLT links 110
- SMLT-SMLT links 96
- SMLT-SMLT links- examples 97
- supported SMLT links 98
- terminology 92
- split VLANs
 - failure modes 132
 - guidelines 132
 - MLT backup path 132
 - single point of failure links 132
 - STG configuration 132
 - using MLT to protect against 132
- STP 124
 - multiple STG interoperability 125
 - STGs and BPDU forwarding 124
- VRRP
 - backup master 104
- designing secure networks
 - DoS attacks (See also DOS attacks) 288
 - implementing security in the Passport 8600 299
 - preventing malicious code 288
 - resiliency and availability attacks 289
- detecting single fiber faults 63
- DiffServ
 - access port in drop mode 346
 - ATM core port 330
 - layer 2 switches 345
 - per-hop behavior (PHB) 344
- DoS attacks 288
 - Passport 8600 security against 290
 - directed broadcast suppression 295
 - stopping spoofed IP packets 294
 - smurf attack 295
 - vs. DDoS attacks 288
- DSL
 - ATM 338
 - bandwidth limitations 338
 - IGMP 338
 - security 338
- DVMRP
 - configuration guidelines 221
 - design rules 240
 - general network design 240
 - sender and receiver placement 241
 - timer tuning 241
 - IGMPv2 and IGMPv1 230
 - MBR path considerations 255
 - multicast route capacity 222
 - scaling 221
 - interface scaling 221
 - route scaling 222
 - stream scaling 222
 - source/group pairs 222
- E**
 - enabling Layer 4- 7 services
 - introduction 183
 - Layer 4- 7 switching 184
 - enabling Layer 4-7 services
 - applications and services 192
 - available network architectures 202
 - WSM architectural details and limitations 206
 - WSM architecture 187
 - external firewalls
 - additional Nortel equipment 320
 - automatic load balancing 321
- F**
 - failover, HA CPU 69

FEFI 60

G

general network design considerations

electrical considerations- power supply
matrix 48

hardware considerations 37

10GE applications for WAN and MAN
markets 39

10GE comparison with 1GE 40

10GE port mirroring 43

10GE port mirroring and dual switch fabric
use 44

10GE port mirroring and internal MLT and
load balancing 44

8692SF module 48

E- and M-modules 38

hardware record optimization 46

record reservation 46

software considerations and hardware
dependencies 49

generic network design rules 55

Gigabit Ethernet

10GE

applications for WAN and MAN markets 39

comparison with 1GE 40

design constraints 43

physical interfaces 39

over copper cabling specification 58

Gigabit IEEE 802.3z specification 67

GNS

request 180

response 180

H

HA mode

about 69

enable 71

failover 69

hardware and software reliability

drivers 54

interacting software (OSPF) 54

local software (MLT) 54

hardware considerations

10GE physical interfaces 39

High Availability mode, about 69

See also HA mode

I

ICMP redirect messages 152

options for avoiding 152

IDS 296

categories 296

connecting Passport 8600 to external
servers 296

products 296

SLB 298

IEEE 802.1ad 135

IEEE 802.3ab specification 58

IGAP 281

IGMP

ATM 332

DSL 338

snooping 234

IGMPv2 281

IP filtering

bridged traffic on DiffServ access ports 342

filter IDs 344

forwarding decisions 343

global filters capacity 343

global filters description 343

global filters for IP bridged traffic 343

IP multicast traffic 342, 343

routed traffic 343

routed traffic on DiffServ access ports 342

source/destination filter configuration 344

source/destination filter mask length 344

source/destination filters capacity 344

source/destination filters description 343

IP multicast

and IGAP 281

and MLT 216

- DVMRP probe 216
- DVMRP routing 217
- IGMP query 216
- ATM guidelines 330
- filtering guidelines 235
 - for IGMP vs DVMRP and PIM 236
- filtering IGMP snoop 234
- flow distribution over MLT 219
- general rules 226
 - DVMRP IGMPv2 back-down to IGMPv1 230
 - dynamic configuration changes 229
 - multicast filtering and multicast access control 234
 - split-subnet and IP multicast 236
 - TTL in IP multicast packets 230
- IGMP and DVMRP 237
- IGMP and PIM 238
- pruning multicast streams 231
- scaling 221, 222, 223, 224
- IP routing, disable per port for VLAN 157
- IPSEC 321
 - Nortel product support 322
- IPX
 - design guidelines 180
 - GNS 180
 - hop counts 180
 - LLC encapsulation and translation 181
 - Netware services 180
 - RIP route cost 180
 - RIP/SAP policies 181
 - SAP route cost 180
- isolated VLANs 132
 - configuration guidelines 132
 - STP blocking 132
- IST link- failure avoidance 95

L

- L2 and L3 multicast 268
- L2 IGMP snooping 267
- L2 traffic classification 345

- link, SMLT single 98
- loop detection
 - and sVLAN 142

M

- malicious code
 - Passport 8600 security against 296
 - viruses, worms, and Trojan horses (definitions) 288
- MLT 67, 71
 - ATM
 - guidelines 329
 - BPDUs 86
 - client/server configuration 85
 - configuration guidelines 71
 - description 219
 - E-module support 219
 - multicast flow distribution 73, 219
 - preventing bridging loops of BPDUs 72
 - routed links 72
 - STG configuration 73
 - switch-to-server configuration 84
 - switch-to-switch configuration 83
 - switch-to-switch links 72
 - traffic distribution algorithm 73
- M-modules
 - comparisons with E-modules 38
 - scaling numbers 39
 - supported modules 39
- multicast
 - E-module support for MLT 219
 - flow distribution over MLT 73, 219

N

- N + 1 power supply redundancy 67
- network management
 - stacked VLAN 147
- network problem tracking statistics 55
- NNI port, sVLAN 138
 - and SMLT 142
 - behavior 140

O

OctaPID

- and sVLAN 138

OctaPID ID

- description 403

onion architecture

- multi-level design, sVLAN 146
- one-level design, sVLAN 144
- two-level design, sVLAN 145

OSPF

- design guidelines 173
- formula for determining area numbers 172
- LSA limits calculation formula 172
- network design scenarios 174
 - one subnet in one area 174
 - two subnets in one area 176
 - two subnets in two areas 177
- route summarization and black hole routers 173
- scalability calculation formula 172
- scalability guidelines 172

P

Passport 8000 series

- scaling numbers (release 3.3) 395
- system management 387
 - backup configuration files 392
 - default management IP address 392
 - DNS client 392
 - offline switch configuration 387
 - pemboot.cfg 391
 - port mirroring- general considerations 388
 - port mirroring- identifying E-modules 388
 - port mirroring- local port mirroring 388
 - port mirroring- mirroring scalability 389
 - remote mirroring 390

Passport 8600

- hardware queues 353
 - queue weights and PTO 354

QoS

- 8 different levels 353
- access vs. core port network design 359

- accessing MAC level 367
- accessing port level 367
- algorithm 393
- and filtering 366
- and SLAs 358
- bridged vs. routed traffic network design 361
- bursty network congestion 376
- core port (IP bridged traffic) 369
- core port and access port 351
- determining internal level 367
- DiffServ access (IP routed traffic) 370
- DiffServ access mode flow charts 371
- DiffServ core (IP routed traffic) 371
- emission priority queuing 352
- enabling DiffServ on a port (See also DiffServ) 366
- filtering and decision-making 357
- highlights 351
- internal QoS levels 352
- mechanisms 350
- network design considerations 358
- network scenarios for bridged traffic 380
- network scenarios for routed traffic 384
- no network congestion 376
- non-IP traffic (bridged or L2) 370
- packet classification (ingress interface configuration) 355
- policy and rate metering 357
- severe network congestion 378
- summary 364
- tagged vs. untagged packet network design 362
- trusted vs. untrusted network design 359

security measures 299

- control plane 309
- control plane (6 management access levels) 311
- control plane (access policies) 314
- control plane (enforcing RADIUS) 315
- control plane (management) 309
- control plane (password rules) 313
- control plane (Secure Shell and Secure Copy) 318
- control plane (SNMP) 319

- control plane (using other Nortel equipment) 320
- data plane 300
- data plane (L2 and L3 filtering capabilities) 303
- data plane (routing policies) 307
- data plane (routing protocol protection) 308
- data plane (VLAN traffic isolation) 303
- PIM**
 - general guidelines 249
 - MBR and DVMRP path considerations 255
 - recommended MBR configuration 251
 - redundant MBR configuration 252
 - RP placement 260
 - PIM network with non-PIM interfaces 265
 - receivers on interconnected VLANs 264
 - RP and extended VLANs 264
 - scaling
 - interface scaling 223
 - route scaling 223
 - rules for improving scalability 224
 - stream scaling 224
- PIM-SM**
 - DR**
 - (designated router) 274
- Point-to-Point protocol over Ethernet, about 157
 - See also PPPoE
- port mirroring
 - OctaPID ID and port assignments 404
- PPPoE
 - about 157
- product support 34
- protocol-based VLAN, PPPoE 157
- provider bridges 135
- provisioning QoS networks
 - admin weights for traffic queues 344
 - combining IP filtering with DiffServ 342
 - DiffServ access ports in drop mode 346
 - DiffServ interoperability with L2 345
 - IP filter ID 344
 - IP filtering and ARP 342
 - IP filtering and forwarding decisions 343
 - Nortel's QoS strategy 348
 - class of service mapping 350
 - traffic classification 348
 - per-hop behaviors 344
 - QoS mechanisms for Passport 8600 350
 - QoS overview (See also QoS) 346
- publications
 - hard copy 34
- PVST+ 129, 130
- Q**
 - Q-inQ 135
 - QoS**
 - admin weights 344
 - algorithm 393
 - benefits 347
 - DSCP marking 346
 - fair servicing 345
 - focus 347
 - high priority queues 345
 - key parameters 347
 - low priority queues 345
 - mapping to ATM class of service 330
 - Nortel's strategy 348
 - overview 346
 - packet transmission opportunities 344
 - QoS to DSCP table 346
- R**
 - RADIUS**
 - configuring a client 316
 - customizable parameters 316
 - Passport 8000 as a client 315
 - supported servers 315
 - rate limit, control packet 95
 - resiliency and availability attacks 289
 - organizations to contact for information 289
 - Passport 8600 security against 299
 - RSMLT** 112
 - designing and configuring 115
 - failure scenarios 113

operation in L3 environments 112

S

scaling numbers for E-modules 39

SFFD 63

configuration rulesI 64

configuring using the CLI 64

remote fault indication 62

single fiber fault detection 63

SMLT 91, 95, 266

and stacked VLAN 141

and STP 110

configuration 93

design that avoids duplicate traffic 270

designs 106

designs- full mesh configuration 109

designs- square configuration 108

designs- triangle configuration 107

failure scenarios 104

ID configuration 98

IST link 94

IST link- dual purpose comm channel 94

IST link- IST VLAN and peer IP

configuration 96

IST link- supported links 96

IST link- using CP-Limit with 95

Layer 2 traffic load sharing 103

Layer 3 traffic load sharing 103

scalability- MAC address scalability 111

scalability- SMLT and multicast scalability 111

scalabilty- IST/SMLT scalability 111

scalablilty 110

single port 98

SMLT-SMLT links 96

SMLT-SMLT links- examples 97

square design 269

supported SMLT links 98

terminology 92

triangle design 267, 268

SNMP

security holes 319

split VLANs

failure modes 132

guidelines 132

MLT backup path 132

single point of failure links 132

STG configuration 132

using MLT to protect against 132

spoofed IP packets 294

configuring generic filters 294

denying invalid source IP addresses 294

source addresses to be filtered 294

SSH protocol 318

Passport 8000 support 318

security aspects 318

SSL 322

Nortel product support 322

stacked VLAN, about 135

See also sVLAN

STGs

and BPDU forwarding 124

configuring roots 128

configuring VLANs 128

layer 2 switches 125

sample problem 125

sample solution 126

specifying source MAC addresses 127

subnet-based VLANs 155

and DHCP 156

and IP routing 155

and multinetting 156

and VRRP 155

scalability 156

support, Nortel Networks 34

sVLAN

about 135

and loop detection 142

enabling 143

and SMLT 141

components 138

design

multi-level 146

one-level 144

two-level 145

- features 136
- independent VLAN learning 146
- management 147
- operation 137
- restrictions 147
- switch level 139

switch level, sVLAN 138, 139

T

- tags, stacked 135
- Tap and OctaPID assignment 403
- technical publications 34
- technical support 34
- TLS 337
- types of records 38

U

- UNI port, sVLAN 138
 - and SMLT 141
 - behavior 140

V

- video over DSL over ATM 338
 - point-to-multipoint configuration 339
 - point-to-point configuration 339
- VLACP 64
 - and SMLT 66
- VLANs
 - designing stacked VLAN networks 135
 - port, disable IP routing 157
 - protocol-based, PPPoE 157
- VPNs
 - elements 321
 - IPSEC and its benefits 321
 - SSL and its benefits 322
 - TLS and its benefits 322
- VRRP 150
 - backup master 104
 - configuration guidelines 150

- optimal convergence 151
- slow convergence 151
- STG configuration 151
- virtual IP addresses 150

W

WSM

- applications and services 192
 - application abuse protection 199
 - application redirection 197
 - GSLB 196
 - health checking metrics 194
 - Layer 7 deny filters 200
 - local server load balancing 192
 - network problems addressed by WSM 201
 - VLAN filtering 198
- architecture 187
 - console and management support 212
 - image management 210
 - Passport unknown MAC discard 209
 - SNMP and MIB management 211
 - syslog 210
 - user and password management 207
- available L2- L7 processing designs 202
 - L3 routing in the Passport 8600 203
 - L4-7 services with a single Passport 8600 204
 - L4-7 services with dual Passport 8600s 205
 - pure Passport L2 environment 202
- components 186
- default architecture 191
- default parameters 190
- default Passport parameters and settings 188
 - VLAN 4093 and STG 64 189
- detailed data path architecture 190
- front-facing ports 188
- introduction 183
- Layer 4-7 switching 184
- Layer 4-7 switching in the Passport 8600 185
- location 185
- password mapping for the Passport 8600 208
- rear-facing ports 188
- simplified data path architecture 187