7450 Ethernet Service Switch
7750 Service Router
7950 Extensible Routing System

Advanced Configuration Guide - Part II
Releases Up To 15.0.R5

3HE 13718 AAAA TQZZA 01

Issue: 01

November 2017

The information presented is subject to change without notice. No responsibility is assumed for inaccuracies contained herein.

Table of Contents

Preface 21
About This Guide..........21

Services Overview 25

G.8032 Ethernet Ring Protection Multiple Ring Topology 27
Applicability 27
Overview 27
Configuration 35
Conclusion 70

G.8032 Ethernet Ring Protection Single Ring Topology 71
Applicability 71
Overview 71
Configuration 74
Conclusion 90

Layer 2 Services and EVPN 91

Auto-Learn MAC Protect in EVPN..........93
Applicability 93
Overview 93
Configuration 96
Conclusion 124

BGP Multi-Homing for VPLS Networks 125
Applicability 125
Summary 125
Overview 127
Configuration 129
Conclusion 158

BGP Virtual Private Wire Services..........159
Applicability 159
Overview 160
Configuration 163
Conclusion 187

BGP VPLS 189
Applicability 189
Summary 189
Overview 190
Configuration 191

Conclusion 221

Black-hole MAC for EVPN Loop Protection 223
Applicability 223
Overview 223
Configuration 227
Conclusion 237

Conditional Static Black-Hole MAC in EVPN 239
Applicability 239
Overview 239
Configuration 242
Conclusion 268

EVPN for MPLS Tunnels 269
Applicability 269
Overview 269
Configuration 271
Conclusion 321

EVPN for MPLS Tunnels in Epipe Services (EVPN-VPWS) 323
Applicability 323
Overview 323
Configuration 325
Conclusion 352

EVPN for MPLS Tunnels in Routed VPLS 353
Applicability 353
Overview 353
Configuration 356
Conclusion 377

EVPN for PBB over MPLS (PBB-EVPN)..379
Applicability 379
Overview 379
Configuration 382
Conclusion 417

EVPN for VXLAN Tunnels (Layer 2)..419
Applicability 419
Overview 419
Configuration 421
Conclusion 446

EVPN for VXLAN Tunnels (Layer 3)..447
Applicability 447
Overview 447
Configuration 448

Conclusion 486

EVPN Interconnect Ethernet Segments 487
Applicability 487
Overview 487
Configuration 488
Conclusion 512

EVPN-MPLS Interconnect for EVPN-VXLAN VPLS Services....513
Applicability 513
Overview 513
Configuration 515
Conclusion 538

Fully Dynamic VSD Integration Model....539
Applicability 539
Overview 540
Configuration 547
Conclusion 590

Inter-AS Model C for VLL....591
Applicability 591
Overview 591
Configuration 593
Conclusion 608

Layer 2 Multicast Optimization for EVPN-VXLAN - Assisted Replication....609
Applicability 609
Overview 609
Configuration 618
Conclusion 633

LDP VPLS Using BGP Auto-Discovery 635
Applicability 635
Summary 635
Overview 636
Configuration 638
Conclusion 657

LDP VPLS Using BGP Auto-Discovery - Prefer Provisioned SDP....659
Applicability 659
Overview 659
Configuration 661
Conclusion 670

Multi-Chassis Endpoint for VPLS Active/Standby Pseudowire....671
Applicability 671
Overview 671

Configuration 675
Conclusion 695

Multi-Segment Pseudowire Routing 697
Applicability 697
Summary 697
Overview 698
Configuration 701
Conclusion 745

P2MP mLDP Tunnels for BUM Traffic in EVPN-MPLS Services 747
Applicability 747
Overview 747
Configuration 752
Conclusion 771

PBB-Epipe 773
Applicability 773
Overview 773
Configuration 775
Conclusion 792

PBB-EVPN ISID-based CMAC Flush 793
Applicability 793
Overview 793
Configuration 802
Conclusion 818

PBB-VPLS 819
Applicability 819
Summary 819
Overview 820
Configuration 821
Conclusion 854

Preference-based and Non-revertive EVPN DF Election 855
Applicability 855
Overview 855
Configuration 858
Conclusion 878

Shortest Path Bridging for MAC 879
Applicability 879
Overview 879
Configuration 881
Conclusion 909

Virtual Ethernet Segments 911
Applicability 911
Overview 911
Configuration 915
Conclusion 923

VLAN Range SAPs for VPLS and Epipe Services 925
Applicability 925
Overview 925
Configuration 935
Conclusion 943

Layer 3 Services 945

BGP Best External in a VPRN 947
Applicability 947
Overview 947
Configuration 954
Conclusion 973

Carrier Supporting Carrier IP VPNs 975
Applicability 975
Overview 975
Configuration 978
Conclusion 998

Layer 3 VPN: VPRN Type Spoke 999
Applicability 999
Summary 999
Overview 999
Configuration 1002
Conclusion 1012

NG-MVPN Configuration with MPLS 1013
Applicability 1013
Summary 1013
Overview 1015
Configuration 1017
Conclusion 1069

NG-MVPN Configuration with PIM 1071
Applicability 1071
Summary 1071
Overview 1074
Configuration 1076
Conclusion 1117

NG-MVPN Sender-Only, Receiver-Only 1119
Applicability 1119
Overview 1119
Configuration 1121
Conclusion 1170

NG-MVPN Source Redundancy 1171
Applicability 1171
Summary 1171
Overview 1172
Configuration 1174
Conclusion 1202

NG-MVPN Wildcard S-PMSI 1203
Applicability 1203
Overview 1203
Configuration 1207
Conclusion 1230

Rosen MVPN Core Diversity 1231
Applicability 1231
Overview 1231
Configuration 1237
Conclusion 1256

Rosen MVPN Inter-AS Option B 1257
Applicability 1257
Overview 1257
Configuration 1265
Conclusion 1282

Spoke Termination for IPv6-6VPE 1285
Applicability 1285
Overview 1285
Configuration 1287
Conclusion 1307

Traffic Leaking from VPRN to GRT 1309
Applicability 1309
Overview 1309
Configuration 1310
Conclusion 1333

VPRN Inter-AS VPN Model C 1335
Applicability 1335
Overview 1335
Configuration 1339
Conclusion 1352

Quality of Service 1353

Class Fair Hierarchical Policing for SAPs 1355
Applicability 1355
Summary 1355
Overview 1356
Configuration 1365
Conclusion 1392

FP and Port Queue Groups 1393
Applicability 1393
Overview 1393
Configuration 1395
Conclusion 1433

High Scale QoS IOM: QoS, Service, and Network Configuration 1435
Applicability 1435
Overview 1435
Configuration 1451
Conclusion 1511

Pseudowire QoS..1513
Applicability 1513
Overview 1513
Configuration 1518
Conclusion 1534

QoS Architecture and Basic Operation...1535
Applicability 1535
Overview 1535
Configuration 1536
Conclusion 1575

List of tables

G.8032 Ethernet Ring Protection Multiple Ring Topology 27
Table 1 Terminology Comparison 30

BGP VPLS 189
Table 2 VE-IDs and Labels 198
Table 3 VE-IDs and Number of Labels..199

EVPN for MPLS Tunnels 269
Table 4 Comparing EVPN Multi-homing and BGP Multi-homing 314

EVPN for PBB over MPLS (PBB-EVPN)..379
Table 5 EVPN and PBB-EVPN SR OS Feature Comparison 379
Table 6 PBB-EVPN Multi-Homing Supported Combinations in SR OS 400

Layer 2 Multicast Optimization for EVPN-VXLAN - Assisted Replication.....609
Table 7 Inclusive Multicast Route Information Sent By Different AR Roles 613

P2MP mLDP Tunnels for BUM Traffic in EVPN-MPLS Services 747
Table 8 IMET routes and Tunnel Types advertised based on the configuration 755

PBB-EVPN ISID-based CMAC Flush...793
Table 9 CMAC Flush Transmission Behavior 797
Table 10 CMAC Flush Reception Behavior 798

Virtual Ethernet Segments 911
Table 11 Supported Examples for Q-tag Values between 1 and 4094 914
Table 12 Supported Examples for Q-tag Values 0, *, and null 914

VLAN Range SAPs for VPLS and Epipe Services 925
Table 13 VLAN Manipulation in SAPs 926
Table 14 SAP Lookup Order for Dot1q Ports 929
Table 15 SAP Lookup Order for QinQ Ports 929

NG-MVPN Configuration with PIM 1071
Table 16 Next Generation MVPN Components...1072

NG-MVPN Wildcard S-PMSI 1203
Table 17 S-PMSI Auto-Discovery BGP NLRI 1206

Class Fair Hierarchical Policing for SAPs 1355
Table 18 Burst Levels..1357
Table 19 Policer stat-mode..1366

FP and Port Queue Groups 1393
Table 20 Default QoS and Queue Group Comparison 1394
Table 21 Queue Group Templates - Ingress 1397
Table 22 Queue Group Templates - Egress 1397
Table 23 Network Ingress FP Queue Group Policer Usage 1404

QoS Architecture and Basic Operation..1535
Table 24 SAP Ingress Classification Match Criteria..1539
Table 25 QinQ Dot1p Bit Classification 1541
Table 26 Forwarding Classes..1541
Table 27 Queue Priority vs. Profile Mode..1546
Table 28 Network QoS Policy DSCP Remarking 1552

List of figures

G.8032 Ethernet Ring Protection Multiple Ring Topology 27

Figure 1 G.8032 Major Ring and Sub-Ring 31
Figure 2 G.8032 Ring Components 33
Figure 3 G.8032 Sub-Ring Interconnection Components.......................................34
Figure 4 Ethernet Test Topology 39
Figure 5 ETH-CFM MEP Associations 41
Figure 6 Sub-Ring to VPLS Topology...63

G.8032 Ethernet Ring Protection Single Ring Topology 71

Figure 7 G.8032 Operation and Topologies 73
Figure 8 Example Topology...74
Figure 9 Ethernet CFM Configuration..78

Auto-Learn MAC Protect in EVPN...93

Figure 10 Example Topology - No LAG...97
Figure 11 MAC Address Learned Simultaneously on SAPs on PE-2 and PE-3...100
Figure 12 107
Figure 13 Default RPS-DF on SAPs - MAC Learned and Protected on SAP on PE-2...109
Figure 14 MAC Learned and Protected Simultaneously on PEs - RPS-DF on EVPN Endpoints...111
Figure 15 MAC Learned and Protected on SAP on PE-2 - RPS Enabled on SAP on PE-3 117
Figure 16 RPS Enabled on SAPs - RPS-DF on EVPN Endpoints, MACs Learned Simultaneously...119
Figure 17 ALMP in All-Active Multi-Homing SAPs...121
Figure 18 All-Active Multi-Homing - RPS-DF on SAPs and EVPN Endpoints.........123

BGP Multi-Homing for VPLS Networks 125

Figure 19 Example Topology..127
Figure 20 Nodes Involved in BGP MH...131
Figure 21 MAC Flush for BGP MH 143
Figure 22 Access PE/CE Signaling 145
Figure 23 Oper-Groups and BGP-MH 148

BGP Virtual Private Wire Services..159

Figure 24 Example Topology..160
Figure 25 Single Homed BGP VPWS using Auto-Provisioned SDPs.......................166
Figure 26 Single Homed BGP VPWS using Pre-Provisioned SDP 173
Figure 27 Dual Homed BGP VPWS with Single Pseudowire 177
Figure 28 Dual Homed BGP VPWS with Active/Standby Pseudowire 183

BGP VPLS 189
Figure 29 Network Topology...190
Figure 30 BGP VPLS Using Auto-Provisioned SDPs...196
Figure 31 BGP VPLS Using Pre-Provisioned SDP..212

Black-hole MAC for EVPN Loop Protection...223
Figure 32 Black-hole MAC for EVPN Loop Protection...224
Figure 33 Example Topology..227
Figure 34 Example Topology with All-active Multi-homing 235
Figure 35 Traffic Dropped when ALMP is Configured in All-active Multi-homing...236

Conditional Static Black-Hole MAC in EVPN 239
Figure 36 Proxy-ARP/ND and ARP Spoofing 241
Figure 37 Example Topology..242
Figure 38 Conditional Static Black-Hole MAC 244
Figure 39 VPLS 1 with Proxy-ARP and AS-MAC 254

EVPN for MPLS Tunnels 269
Figure 40 EVPN Route Types and NLRIs 270
Figure 41 EVPN-MPLS for VPLS Services...272
Figure 42 EVPN-MPLS All-Active Multi-Homing Concepts 287
Figure 43 EVPN-MPLS Single-Active Multi-Homing: Mass-Withdraw, Backup Path 302

EVPN for MPLS Tunnels in Epipe Services (EVPN-VPWS) 323
Figure 44 Route Types and NLRIs for EVPN-VPWS 324
Figure 45 EVPN-VPWS Example Topology 326
Figure 46 Example Topology for EVPN-VPWS without Multi-Homing 328
Figure 47 Example Topology EVPN-VPWS with Multi-Homing...............................334

EVPN for MPLS Tunnels in Routed VPLS 353
Figure 48 Passive VRRP - vMAC/vIP Advertised By GARP 355
Figure 49 R-VPLS with EVPN Tunnel, without Multi-homing 357
Figure 50 EVPN-MPLS R-VPLS with All-Active MH ES 363
Figure 51 EVPN-MPLS R-VPLS with Single-active Multi-Homing...........................373

EVPN for PBB over MPLS (PBB-EVPN)..379
Figure 52 EVPN Route Types 381
Figure 53 PBB-EVPN Network without Multi-Homing...382
Figure 54 PBB-EVPN — Flooding Lists..386
Figure 55 PBB-EVPN Multi-homing..398
Figure 56 The Use of es-bmac to Minimize CMAC Flush...400
Figure 57 PBB-EVPN Single-Active Support for Epipes...415

EVPN for VXLAN Tunnels (Layer 2)..419
Figure 58 EVPN-VXLAN Example Topology 422
Figure 59 BGP Adjacencies and Enabled Families 425

Figure 60 EVPN MAC Mobility....438

EVPN for VXLAN Tunnels (Layer 3)....447

Figure 61 EVPN-VXLAN for R-VPLS Services....449
Figure 62 BGP adjacencies and enabled families....452
Figure 63 EVPN-VXLAN for IRB Backhaul R-VPLS Services....457
Figure 64 EVPN-VXLAN in EVPN-tunnel R-VPLS Services....466
Figure 65 Routing Policies for Egress EVPN Routes....474
Figure 66 Routing Policies for Ingress EVPN Routes....475
Figure 67 EVPN in Parallel R-VPLS Services....479

EVPN Interconnect Ethernet Segments....487

Figure 68 EVPN-MPLS Interconnect for EVPN-VXLAN - BGP Topology....488
Figure 69 VPLS service and association with I-ESs....492
Figure 70 All-active Multi-homing and Unknown Unicast Example 1....504
Figure 71 All-active Multi-homing and Unknown Unicast Example 2....505
Figure 72 All-active Multi-homing and Unknown Unicast Example 3....505
Figure 73 All-active Multi-homing and send-imet-ir-on-ndf....506
Figure 74 All-active Multi-homing and no send-imet-ir-on-ndf....509

EVPN-MPLS Interconnect for EVPN-VXLAN VPLS Services....513

Figure 75 EVPN-MPLS Interconnect for EVPN-VXLAN - Example Topology....516
Figure 76 EVPN Destinations Created on Multi-Homed Anycast DC GWs....524
Figure 77 Use of Provider-tunnels Between Anycast DC GWs Create Packet Duplication....536

Fully Dynamic VSD Integration Model....539

Figure 78 Nuage VSP overview....540
Figure 79 DC Gateway fully dynamic provisioning workflow....542
Figure 80 F-D XMPP provisioning setup....548

Inter-AS Model C for VLL....591

Figure 81 Example Topology- Inter-AS Model C for VLL....592
Figure 82 Inter-AS Model C for VLL....592
Figure 83 Network Setup Configuration....593

Layer 2 Multicast Optimization for EVPN-VXLAN - Assisted Replication....609

Figure 84 PMSI Tunnel Attribute - Flags....611
Figure 85 EVPN Assisted Replication for VXLAN....612
Figure 86 Example Topology....619

LDP VPLS Using BGP Auto-Discovery....635

Figure 87 Example Topology....636
Figure 88 VPLS Instance with Auto-Provisioned SDPs....642
Figure 89 VPLS Instance using Pre-Provisioned SDPs....653

LDP VPLS Using BGP Auto-Discovery - Prefer Provisioned SDP....659

Figure 90 LDP VPLS Using BGP-AD with use-provisioned-sdp Option....660

Figure 91 LDP VPLS Using BGP-AD with prefer-provisioned-sdp Option 661
Figure 92 Example Topology...662
Figure 93 SDP Bindings in VPLS 1 with use-provisioned-sdp Option 666
Figure 94 Auto-Created SDP Bindings in VPLS 2 666
Figure 95 SDP Bindings in VPLS 1 with prefer-provisioned-sdp Option 670

Multi-Chassis Endpoint for VPLS Active/Standby Pseudowire 671
Figure 96 H-VPLS with STP 672
Figure 97 VPLS Pseudowire Redundancy 672
Figure 98 Multi-Chassis Endpoint with Mesh Resiliency 673
Figure 99 Multi-Chassis Endpoint with Square Resiliency 673
Figure 100 Example Topology...674
Figure 101 Core Node Failure 689
Figure 102 Multi-Chassis Node Failure 690
Figure 103 Multi-Chassis Passive Mode 693

Multi-Segment Pseudowire Routing 697
Figure 104 FEC129 Structure...698
Figure 105 All Type 2 Format...699
Figure 106 Pseudowire Routing NLRI (the AC ID is always zero)............................699
Figure 107 Configuration Flow Chart...701
Figure 108 Intra-AS MS-PW Example Topology 718
Figure 109 Inter-AS MS-PW Example Topology 731

P2MP mLDP Tunnels for BUM Traffic in EVPN-MPLS Services 747
Figure 110 P2MP mLDP Tree with Root Node PE-1 and Leaf Nodes PE-5, PE-6, and PE-7..748
Figure 111 BGP-EVPN Route Type 3 with PTA 749
Figure 112 PTA for Composite Tunnel IMET-P2MP-IR...751
Figure 113 P2MP mLPD in PBB-EVPN...767

PBB-Epipe 773
Figure 114 Network Topology...774
Figure 115 Setup Detailed View 776
Figure 116 Virtual MEPs for Flooding Avoidance 784

PBB-EVPN ISID-based CMAC Flush 793
Figure 117 CMAC flush when SAP in BGP Multi-homing Site Fails..........................794
Figure 118 EVPN BMAC Route with ISID Indication 795
Figure 119 ISID-independent CMAC Flush when ES Fails 800
Figure 120 Example Topology...802
Figure 121 Example Topology with BGP Multi-homing 803
Figure 122 Example Topology with Single-Active ES..810

PBB-VPLS 819
Figure 123 Example Topology 820
Figure 124 MTU-4 and PE-1 Nodes as Configuration Examples 822
Figure 125 Black-hole...833

Figure 126 Send Flush on BVPLS Failure Example....837
Figure 127 Inter-Domain B-VPLS and MMRP Policies/ISID-Based Filters Example....843

Preference-based and Non-revertive EVPN DF Election 855

Figure 128 Virtual Ethernet Segments....856
Figure 129 BGP-EVPN Extended Community for DF Election....857
Figure 130 Example Topology with All-active and Single-active vESs....858
Figure 131 Calculation....861

Shortest Path Bridging for MAC 879

Figure 132 Basic SPBM Topology....881
Figure 133 Control and User B-VPLS Test Topology....893
Figure 134 Access Resiliency Test Topology....897
Figure 135 Access Resiliency Example Topology....901

Virtual Ethernet Segments 911

Figure 136 vESs for PWs 912
Figure 137 Example Topology....915

VLAN Range SAPs for VPLS and Epipe Services 925

Figure 138 Customer VID is Popped and Pushed by VLAN SAPs - VLAN Translation....926
Figure 139 Customer VID is Preserved between Dot1q CP SAPs - No VLAN Translation....927
Figure 140 Customer VID is Preserved between QinQ CP SAPs - No VLAN Translation....927
Figure 141 Example Topology....935
Figure 142 Example Topology for VLAN Ranges in VPLS 1....936
Figure 143 Customer VIDs are Popped and Pushed by Dot1q VLAN SAPs 938
Figure 144 Customer VID is Preserved between Two Dot1q CP SAPs....939
Figure 145 No Traffic Between Dot1q CP SAP and Dot1q VLAN SAP....939
Figure 146 Traffic between Two QinQ VLAN SAPs - VLAN Translation....940
Figure 147 No Traffic between Two QinQ CP SAPs - VLAN Translation Not Supported....941
Figure 148 Traffic between Two QinQ CP SAPs - No VLAN Translation....942
Figure 149 Example Topology for VLAN Ranges in Epipe 2....942

BGP Best External in a VPRN 947

Figure 150 CE-4 Advertises prefix 10.0.0.0/8 to its eBGP Peers PE-1 and PE-2....949
Figure 151 Default BGP Behavior: BGP Advertises Best Route Only....950
Figure 152 VPRN BGP Best External Enabled: BGP Advertises Active and Standby Routes 951
Figure 153 BGP Fast Reroute on PE-1 after Failure of Active Link to CE....952
Figure 154 PE-2 Re-Advertises VPN-IPv4 Route with Label Based on VRF 953
Figure 155 Traffic Destined for Prefix 10.0.0.0/8 after Control Plane Convergence 954

Figure 156 Example Topology ... 955
Figure 157 Loadsharing for Traffic from PE-3 Destined to 10.0.0.0/8 ... 972

Carrier Supporting Carrier IP VPNs ... 975
Figure 158 CSC Network Topology ... 976

Layer 3 VPN: VPRN Type Spoke ... 999
Figure 159 CE Hub and Spoke Data Path ... 1000
Figure 160 CE Hub and Spoke Control Plane Isolation ... 1001
Figure 161 Internal VPRN Logic on a PE Router ... 1001
Figure 162 CE Hub and Spoke Topology and Addressing Scheme ... 1003

NG-MVPN Configuration with MPLS ... 1013
Figure 163 Network Topology ... 1016
Figure 164 VPRN 1 Topology used for mLDP ... 1021
Figure 165 VPRN 2 Topology used for RSVP-TE P2MP ... 1041
Figure 166 VPRN 2 Topology used for MVPN Source Redundancy ... 1061
Figure 167 VPRN 3 Topology used for MDT SAFI ... 1066

NG-MVPN Configuration with PIM ... 1071
Figure 168 Network Topology ... 1074
Figure 169 Example Topology for Anycast RP ... 1096
Figure 170 IGMP and PIM Control Messaging Schematic ... 1102
Figure 171 PIM SSM in Customer Signaling Plane ... 1106

NG-MVPN Sender-Only, Receiver-Only ... 1119
Figure 172 Default PMSI Hierarchy ... 1120
Figure 173 Optimized PMSI Structure ... 1120
Figure 174 Example Topology ... 1121
Figure 175 RSVP-Based BGP Message Flow Between PE-1 and PE-2 ... 1130
Figure 176 RSVP-Based BGP Message Flow Between PE-1 and PE-3 ... 1133
Figure 177 mLDP-Based BGP Message Flow Between PE-1 and PE-2 ... 1150
Figure 178 mLDP-Based BGP Message Flow Between PE-1 and PE-3 ... 1157

NG-MVPN Source Redundancy ... 1171
Figure 179 Source Redundancy Example. ... 1172
Figure 180 Schematic Topology ... 1175

NG-MVPN Wildcard S-PMSI ... 1203
Figure 181 Multicast VPN ... 1205
Figure 182 Schematic Topology ... 1208
Figure 183 S-PMSI P2MP LSP Schematic ... 1227

Rosen MVPN Core Diversity ... 1231
Figure 184 Core Diversity Schematic ... 1233
Figure 185 Core Diversity Network ... 1234
Figure 186 Core Diversity Network — Base OSPF ... 1235
Figure 187 Core Diversity Network - OSPF Instance 1 ... 1236

Rosen MVPN Inter-AS Option B 1257
Figure 188 General Topology for Inter-AS MVPN 1258
Figure 189 Protocols Used for Inter-AS MVPN 1258
Figure 190 BGP Signaling Steps 1261
Figure 191 PIM-P Signaling Steps for Default MDT 1262
Figure 192 PIM-C Signaling 1263
Figure 193 PIM-P Signaling Steps for Data MDT 1264
Figure 194 Example Topology Details...1265
Figure 195 BGP Signaling Steps 1269
Figure 196 PIM-P Signaling Steps for Default MDT 1273
Figure 197 PIM-C Signaling 1277
Figure 198 PIM-P Signaling Steps for Data MDT 1279

Spoke Termination for IPv6-6VPE 1285
Figure 199 Spoke Termination for IPv6...1286
Figure 200 IPv6 Addressing and IPv6 Prefixes 1287
Figure 201 MP-BGP VPN IPv6..1288
Figure 202 Spoke Termination for IPv6 Addressing 1290
Figure 203 PE-4 VPRN with SAP to CE-5...1295

Traffic Leaking from VPRN to GRT 1309
Figure 204 VPRN to GRT leak process..1310
Figure 205 Example Topology with IPv4 Addresses 1311
Figure 206 IPv4 VPRN to GRT route leaking for ISIS 1312
Figure 207 Example Topology with IPv6 Addresses 1323

VPRN Inter-AS VPN Model C 1335
Figure 208 Inter-AS VPN Model C..1337
Figure 209 Protocol Overview 1338

Class Fair Hierarchical Policing for SAPs 1355
Figure 210 Policer Token Bucket Model...1357
Figure 211 Peak Information Rate (PIR) Bucket 1358
Figure 212 Committed Information Rate (CIR) Bucket..1359
Figure 213 Fair Information Rate (FIR) Bucket...1360
Figure 214 Policer and Arbiter Hierarchy 1361
Figure 215 Parent Policer and Root Arbiter..1362
Figure 216 Configuration Example 1364
Figure 217 Post Policing Queues 1365
Figure 218 Parent Policer Thresholds 1371

High Scale QoS IOM: QoS, Service, and Network Configuration 1435
Figure 219 Egress HSQ IOM Scheduling Hierarchy..1438
Figure 220 HSQ Queue Group and Secondary Shaper Aggregate Scheduler Bucket..1444
Figure 221 HSQ Buffer Pool Hierarchy...1445
Figure 222 Configured QoS Paths..1452

Pseudowire QoS **1513**

Figure 223 Ingress PW QoS..1514
Figure 224 Egress PW QoS 1515
Figure 225 Example Epipe Pseudowire Topology..1518

QoS Architecture and Basic Operation **1535**

Figure 226 Service and Network QoS Policies..1538
Figure 227 Visualization of Default Network Policies 1556
Figure 228 Default Buffer Pools..1558
Figure 229 WRED Slope Characteristics..1564
Figure 230 Ingress Buffer Pools and Queue Sizing..1566
Figure 231 Egress Buffer Pools and Queue Sizing 1567
Figure 232 Scheduling (Dequeuing Packets from the Queue) 1569
Figure 233 IOM QoS Overview..1570
Figure 234 Ingress and Egress SAP Queue Statistics for an IES Service 1571
Figure 235 Ingress and Egress Network Port Queue Statistics..1572

Preface

About This Guide

The Advanced Configuration Guide is divided into three volumes, the Part I Guide, the Part II Guide, and the Part III Guide.

Part I provides advanced configurations for basic systems, system management, interface configuration, router configuration, unicast routing protocols, and MPLS.

Part II provides advanced configurations for services overview, Layer 2 and EVPN services, Layer 3 services, and Quality of Service.

Part III provides advanced configurations for Multi-Service Integrated Service Adapter and Triple Play Service Delivery Architecture.

The Advanced Configuration Guide supplements the user configuration guides listed below.

The guide is organized alphabetically within each category and provides feature and configuration explanations, CLI descriptions and overall solutions. The chapters in the Advanced Configuration Guide are written for and based on several releases, up to 15.0.R5. The Applicability section in each chapter specifies on which release the configuration is based.

Audience

This manual is intended for network administrators who are responsible for configuring the routers. It is assumed that the network administrators have a detailed understanding of networking principles and configurations.

List of Technical Publications

The 7x50 series documentation set also includes the following guides:

- 7450 ESS, 7750 SR, 7950 XRS, and VSR Basic System Configuration Guide

Describes CLI usage, file system management, boot option file (BOF) configuration, configuring basic system management, node timing, and synchronization functions.

- 7450 ESS, 7750 SR, 7950 XRS, and VSR System Management Guide

  Describes system security features, SNMP and NETCONF features, and event and accounting logs. It covers basic tasks such as configuring management access filters, passwords, and user profiles.

- 7450 ESS, 7750 SR, 7950 XRS, and VSR Interface Configuration Guide

  Describes how to provision Input/Output Modules (IOMs), XMA Control Modules (XCMs), Media Dependent Adapters (MDAs), XRS Media Adapters (XMAs), and ports.

- 7450 ESS, 7750 SR, 7950 XRS, and VSR Router Configuration Guide

  Describes logical IP routing interfaces and associated attributes such as IP addresses, as well as IP and MAC-based filtering, Virtual Router Redundancy Protocol (VRRP), and Cflowd.

- 7450 ESS, 7750 SR, 7950 XRS, and VSR Unicast Routing Protocols Guide

  Provides an overview of unicast routing concepts and provides configuration examples for Routing Information Protocol (RIP), Open Shortest Path First (OSPF), Intermediate-system-to-intermediate-system (IS-IS), and Border Gateway Protocol (BGP) routing protocols, and for route policies.

- 7450 ESS, 7750 SR, 7950 XRS, and VSR Multicast Routing Protocols Guide

  Provides an overview of multicast routing concepts and provides configuration examples for Internet Group Management Protocol (IGMP), Multicast Listener Discovery (MLD), Protocol Independent Multicast (PIM), Multicast Source Discovery Protocol (MSDP), Multipoint LDP, multicast extensions to BGP, and Multicast Connection Admission Control (MCAC).

- 7450 ESS, 7750 SR, 7950 XRS, and VSR MPLS Guide

  Describes how to configure Multiprotocol Label Switching (MPLS), Resource Reservation Protocol (RSVP), Generalized Multiprotocol Label Switching (GMPLS), and Label Distribution Protocol (LDP).

- 7450 ESS, 7750 SR, 7950 XRS, and VSR Services Overview Guide

  Provides a general overview of functionality provided by the routers and describes how to configure service parameters such as Service Access Points (SAPs), Service Distribution Points (SDPs), customer information, and user services.

- 7450 ESS, 7750 SR, 7950 XRS, and VSR Layer 2 Services and EVPN Guide: VLL, VPLS, PBB, and EVPN

Describes Layer 2 service and Ethernet Virtual Private Network (EVPN) functionality and provides examples to configure and implement Virtual Leased Lines (VLLs), Virtual Private LAN Service (VPLS), Provider Backbone Bridging (PBB), and EVPN.

- 7450 ESS, 7750 SR, 7950 XRS, and VSR Layer 3 Services Guide: IES and VPRN

  Describes Layer 3 service functionality and provides examples to configure and implement Internet Enhanced Services (IES) and Virtual Private Routed Network (VPRN) services.

- 7450 ESS, 7750 SR, and 7950 XRS Versatile Service Module Guide

  This guide describes how to configure service parameters for the Versatile Service Module (VSM).

- 7450 ESS, 7750 SR, 7950 XRS, and VSR OAM and Diagnostics Guide

  Describes how to configure features such as service mirroring and Lawful Intercept (LI), and how to use the Operations, Administration and Management (OAM) and diagnostics tools.

- 7450 ESS, 7750 SR, 7950 XRS, and VSR Log Events Guide

  Provides general information about log events.

- 7450 ESS, 7750 SR, and VSR Triple Play Service Delivery Architecture Guide

  Describes the Triple Play Service Delivery Architecture (TPSDA) support and provides examples to configure and implement various protocols and services.

- 7450 ESS, 7750 SR, 7950 XRS, and VSR Quality of Service Guide

  Describes how to configure Quality of Service (QoS) policy management.

- 7750 SR and VSR RADIUS Attributes Reference Guide

  Describes all supported RADIUS Authentication, Authorization, and Accounting attributes.

- 7450 ESS, 7750 SR, and VSR Multiservice Integrated Service Adapter Guide

  Describes services provided by integrated service adapters, such as Application Assurance, IPSec, ad insertion (ADI), and Network Address Translation (NAT).

- 7750 SR Gx AVPs Reference Guide

  Describes Gx Attribute Value Pairs (AVPs).

- 7450 ESS, 7750 SR, 7950 XRS, and VSR Acronyms Reference Guide

  Provides expansions of acronyms found in the user guides.

Services Overview

In This Section

This section provides configuration information for the following topics:

- G.8032 Ethernet Ring Protection Multiple Ring Topology
- G.8032 Ethernet Ring Protection Single Ring Topology

G.8032 Ethernet Ring Protection Multiple Ring Topology

This chapter provides information about G.8032 Ethernet ring protection multiple ring topologies.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

Initially, this chapter was written for release SR OS 12.0.R5, but the CLI in this edition is based on release 14.0.R2.

Overview

G.8032 Ethernet ring protection is supported for data service SAPs within a regular VPLS service, a PBB VPLS (I/B-component) or a routed VPLS (R-VPLS). G.8032 is one of the fastest protection schemes for Ethernet networks. This chapter describes the advanced topic of Multiple Ring Control, sometimes referred to as multi-chassis protection, with access rings being the most common form of multiple ring topologies. Single Rings are covered in G.8032 Ethernet Ring Protection Single Ring Topology.This chapter will use a VPLS service to illustrate the configuration of G.8032. For very large ring topologies, Provider Backbone Bridging (PBB) can also be used, but it is not configured in this chapter.

ITU-T G.8032v2 specifies protection switching mechanisms and a protocol for Ethernet layer network (ETH) Ethernet rings. Ethernet rings can provide wide-area multipoint connectivity more economically due to their reduced number of links. The mechanisms and protocol defined in ITU-T G.8032v2 are highly reliable with stable protection and never form loops, which would negatively affect network operation and service availability. Each ring node is connected to adjacent nodes participating in the same ring using two independent paths, which use ring links (configured on ports or link aggregation groups (LAGs)). A ring link is bounded by two adjacent nodes and a port for a ring link is called a ring port. The minimum number of nodes on a ring is two.

The fundamentals of this ring protection switching architecture are:

- the principle of loop avoidance and
- the utilization of learning, forwarding, and address table mechanisms defined in the ITU-T G.8032v2 Ethernet flow Forwarding Function (ETH_FF) (Control plane).

Loop avoidance in the ring is achieved by guaranteeing that, at any time, traffic may flow on all but one of the ring links. This particular link is called the ring protection link (RPL) and under normal conditions this link is blocked, so it is not used for traffic. One designated node, the RPL owner, is responsible to block traffic over the one designated RPL. Under a ring failure condition, the RPL owner is responsible for unblocking the RPL, allowing the RPL to be used for traffic. The protocol ensures that even without an RPL owner defined, one link will be blocked and it operates as a **break before make** protocol, specifically the protocol guarantees that no link is restored until a different link in the ring is blocked. The other side of the RPL is configured as an RPL neighbor. An RPL neighbor blocks traffic on the link.

The event of a ring link or ring node failure results in protection switching of the traffic. This is achieved under the control of the ETH_FF functions on all ring nodes. A ring automatic protection switching (R-APS) protocol is used to coordinate the protection actions over the ring. The protection switching mechanisms and protocol supports a multi-ring/ladder network that consists of connected Ethernet rings.

Ring Protection Mechanism

The ring protection protocol is based on the following building blocks:

- Ring status change on failure
  - Idle -> Link failure -> Protection -> Recovery -> Idle
- Ring control state changes
  - Idle -> Protection -> Manual Switch -> Forced Switch -> Pending
- Re-use existing ETH OAM
  - Monitoring: ETH continuity check messages (CCM)
  - Failure notification: Y.1731 signal failure

- Forwarding database MAC flush on ring status change
- RPL (Ring Protection Link)
  - Defines blocked link in idle status

When sub-rings are used, they can either connect to a major ring (which is configured in the exact same way as a single ring) or another sub-ring, or to a VPLS service. When connected to a major/sub ring, there is the option to extend the sub-ring control service through the major ring or not. This gives the following three options for sub-ring connectivity:

a. **Sub-ring to a major/sub ring with a virtual channel** — In this case, a data service on the major/sub ring is created which is used to forward the R-APS messages for the sub-ring over the major/sub ring, between the interconnection points of the sub-ring to the major/sub ring. This allows the sub-ring to operate as a fully connected ring and is mandatory if the sub-ring connects two major/sub rings since the virtual channel is the only mechanism that the sub-rings can use to exchange control messages. It also could improve failover times if the sub-ring was large as it provides two paths on the sub-ring interconnection nodes to propagate the fault indication around the sub-ring, whereas without a virtual channel the fault indication may need to traverse the entire sub-ring. Each sub-ring requires its own data service on the major/sub ring for the virtual channel.

b. **Sub-ring to a major/sub ring without a virtual channel** — In this case the sub-ring is not fully connected and does not require any resources on the major/sub ring. This option requires that the R-APS messages are not blocked on the sub-ring over its RPL.

c. **Sub-ring to a VPLS service** — This is similar to the preceding option, but it uses a VPLS service instead of a major ring. In this option, sub-ring failures can initiate the sending of an LDP MAC flush message into the VPLS service when spoke or MPLS mesh SDPs are used in the VPLS service.

Eth-Ring Terminology

The implementation of Ethernet Ring (eth-ring) on SROS uses a VPLS as the construct for a ring flow function (one for ETH_FF (solely for control) and one for each service_FF) and SAPs (on ports or LAGs) as ring links. The control VPLS must be a regular VPLS, but the data VPLS can be a regular VPLS, a PBB (B/I-) VPLS or a routed VPLS. The state of the data service SAPs is inherited from the state of the control service SAPs. Table 1 displays a comparison between the ITU-T and SR/ESS terminologies.

Table 1 **Terminology Comparison**

ITU-T G.8032v2 Terminology	SROS Terminology
ETH_FF	control vpls
Service_FF	data vpls
East Ring Link	path a
West Ring Link	path b
RPL owner	rpl-node owner
RPL Link	path {a|b} rpl-end
MEP	control-mep
ERP control process	eth-ring instance or ring-id
Major Ring	eth-ring
Sub-ring	eth-ring sub-ring
Ring node	Ring Node PE
Ring-ID	Not used; fixed at 1 per G.8032v2

There are various ways that multiple rings can be interconnected and the possible topologies may be large. Customers typically have two forms of networks; access ring edge networks or larger multiple ring networks. Both topologies require ring interconnection.

Figure 1 **G.8032 Major Ring and Sub-Ring**

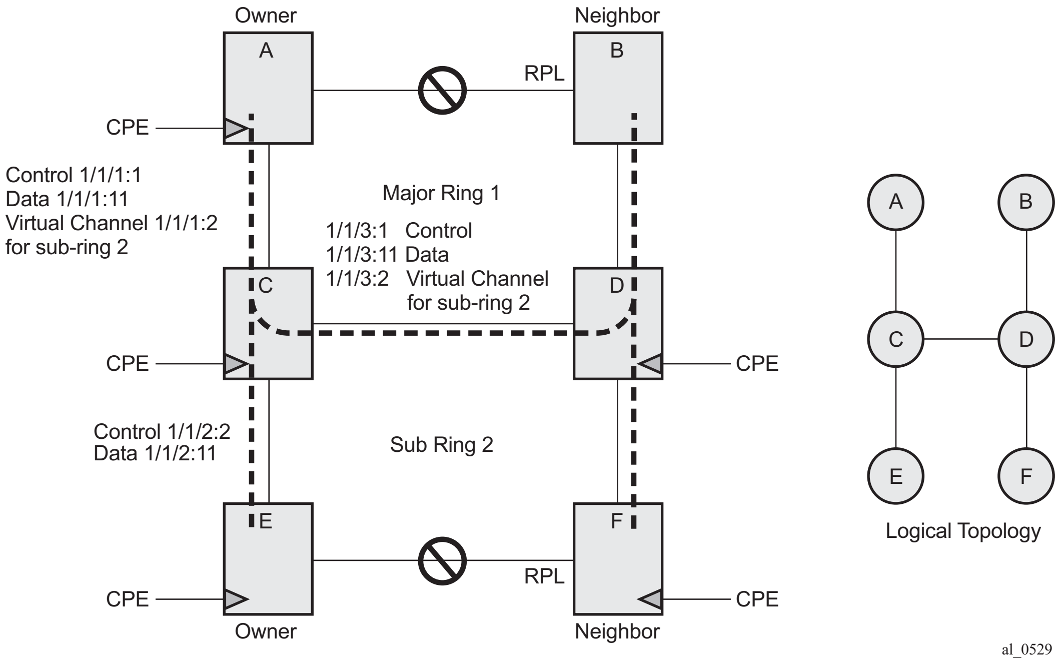


Figure 1 shows a ring of six nodes, with a major ring (regular Ethernet ring) on the top four nodes and a sub-ring on the bottom. A major ring is a fully connected ring. A sub-ring is a partial ring that depends on a major ring or a VPLS topology for part of the ring interconnect. Two major rings can be connected by a single sub-ring or a sub-ring can support other sub-rings.

In the major ring (on nodes A, B, C and D), one path of the RPL owner is designated to be the RPL and the respective SAPs will be blocked in order to prevent a loop. The choice of where to put the RPL is up to the network administrator and can be different for different control instances of the ring allowing an RPL to be used for some other ring's traffic. In the sub-ring, one path is designated as the RPL and will be blocked. Both the major ring and the sub-ring have their own RPL. The sub-ring interconnects to the major ring on nodes C and D and has a virtual channel on the major ring. SROS supports both virtual channel and non-virtual channel rings. Schematics of the physical and logical topologies are also shown in Figure 1.

The G.8032 protocol defines a Ring-ID value (1-255). The SR OS implementation only uses a Ring-ID value of 1, which complies with G.8032v2. The configuration on a node uses a ring instance with a number but all rings use a Ring-ID of 1. This ring instance number is purely local and does not have to match on other ring nodes. Only the VLAN ID must match between SROS ring nodes. For consistency in this example, VPLS instances and Ethernet ring instances are shown as matching for the same ring.

An RPL owner and RPL neighbor are configured for both the major ring and sub-ring. The path and associated link will be the RPL when the ring is fully operational and will be blocked by the RPL owner whenever there is no fault on other ring links. Each ring RPL is independent. If a different ring link fails, then the RPL will be unblocked by the RPL owner. The link shared between a sub-ring and the major ring is completely controlled by the major ring as if the sub-ring were not there. Each ring can completely protect one fault within its ring. When the failed link recovers, it will initially be blocked by one of its adjacent nodes. The adjacent node sends an R-APS message across the ring to indicate the error is cleared and after a configurable time, if reversion is enabled, the RPL will revert to being blocked with all other links unblocked. This ensures that the ring topology when fully operational is predictable.

If a specific RPL owner is not configured (not recommended by G.8032 specification), then the last link to become active will be blocked and the ring will remain in this state until another link fails. This operation makes the selection of the blocked link non-deterministic and is not recommended.

The protection protocol uses a specific control VLAN, with the associated data VLANs taking their forwarding state from the control VLAN. The control VLAN cannot carry data.

Load Balancing with Multiple Ring Instances

Each control ring is independent of the other control rings on the same topology. Therefore since the RPL is used by one control ring, it is often desirable to set up a second control ring that uses a different link as RPL. This spreads out traffic in the topology, but if there is a link failure in the ring, all traffic will be on the remaining links. In the following examples, only a single control ring instance is configured. Other control and data rings could be configured if desired.

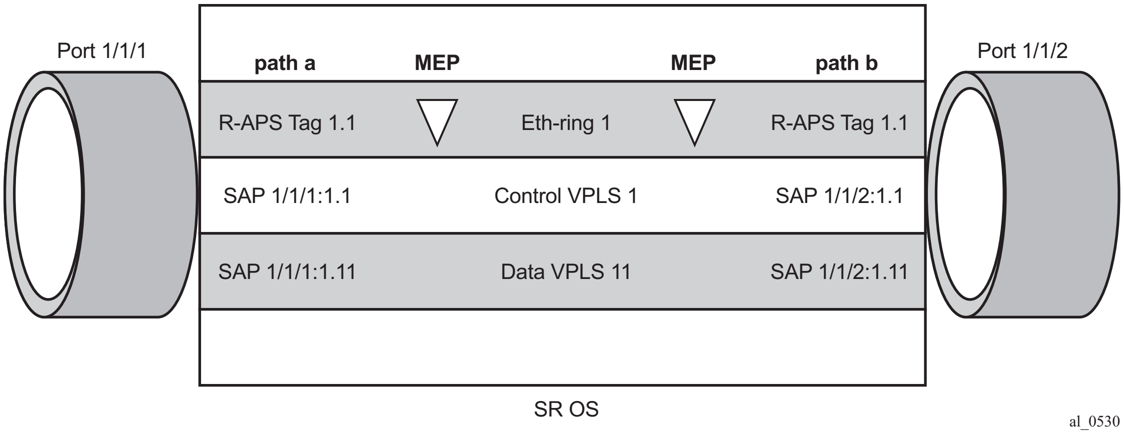
Provider Backbone Bridging (PBB) Support

PBB services also support G.8032 as data services (the services used for the control VPLS must be a regular VPLS). B/I-VPLS rings support both major rings and sub-rings. B-VPLS rings support multi-chassis link aggregation group (MC-LAG) as a dual homing option when aggregating I-VPLS traffic onto a B-VPLS ring. In other words, I-VPLS rings should not be dual-homed into two backbone edge bridge (BEB) nodes where the B-VPLS uses G.8032 to get connected to the rest of the B-VPLS network as the only mechanism which can propagate MAC flushes between an I-VPLS and B-VPLS is an LDP MAC flush.

SROS Implementation

G.8032 is built from VPLS components and each ring consists of the configuration components illustrated in Figure 2.

Figure 2 **G.8032 Ring Components**

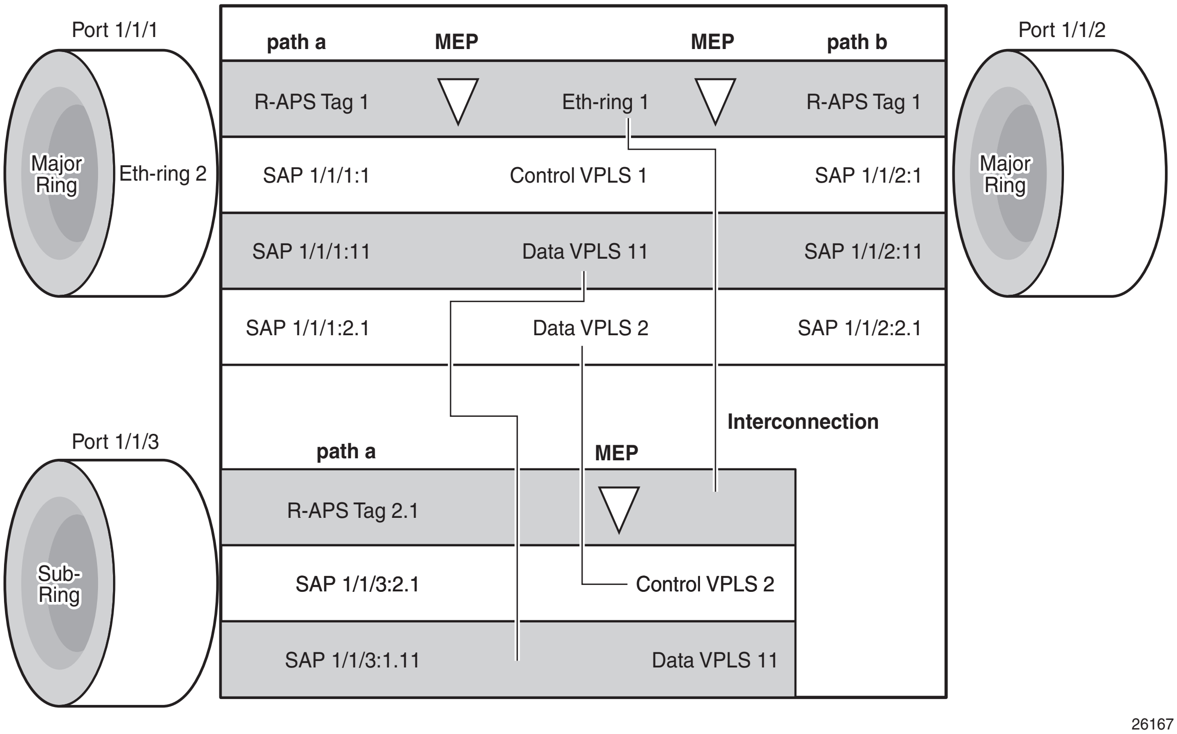


These components consist of:

- The Eth-ring instance which defines the R-APS tags, the MEPs and the ring behavior.
- The control VPLS which has the SAPs that match the R-APS.
- The data VPLS which is linked to the ring. All of the data VPLS SAPs follow the operational state of the control VPLS SAPs in that each blocked SAP controlled by the ring is blocked for all control and data instances.

Figure 3 shows the major ring and sub-ring interconnection components:

Figure 3 **G.8032 Sub-Ring Interconnection Components**



For a sub-ring, the configuration is the same as a single ring except at the junction of the major ring and the sub-ring. The interconnection of a sub-ring and a major ring links the control VPLS of the sub-ring to a data VPLS of the major ring when a virtual link is used. Similarly the data VPLS of the sub-ring is linked to a data VPLS of the major ring. G.8032 Sub-Ring Interconnection Components illustrates the relationship of a sub-ring and a major ring. Since this sub-ring has a virtual channel, the data VPLS 2 has both data SAPs from the sub-ring and data SAPs from the major ring. The virtual channel is also optional and in non-virtual-link cases, no VPLS instance is required (see non-virtual-link in the section Configuration of a Sub-Ring to a VPLS Service (with a Non-Virtual Link)).

In Figure 3, the inner tag values are kept the same for clarity, but in fact any encapsulation that is consistent with the next ring link will work. In other words, ring SAPs can perform VLAN ID translation and even when connecting a sub-ring to a major ring. This also means that other ports may reuse the same tags when connecting independent services.

The R-APS tags (ring automatic protection switching tags) and SAPs on the rings can either be dot1Q or QinQ encapsulated. It is also possible to have the control VPLS using single tagged frames with the data VPLSs using double tagged framed; this requires the system to be configured with the new-qinq-untagged-sap parameter (configure system ethernet new-qinq-untagged-sap), with the ring path raps-tags and control VPLS SAPs configured as qtag.0, and the data VPLSs configured as QinQ SAPs. STP cannot be enabled on SAPs connected to eth-rings.

R-APS messages received from other nodes are normally blocked on the RPL interface but the sub-ring case with non-virtual channel recommends that R-APS messages be propagated over the RPL. Configuring sub-ring non-virtual-link on all nodes on the sub-ring propagation of R-APS messages is mandatory in order to achieve this.

R-APS messages are forwarded out of the egress using forwarding class NC, this should be prioritized accordingly in the SAP egress QoS policy to ensure that congestion does not cause R-APS messages to be dropped which could cause the ring to switch to another path.

Configuration

This section describes the configuration of multiple rings. The eth-ring configuration commands are as follows.

```
configure eth-ring <ring-index>
              ccm-hold-time { [down <down-timeout>] [up <up-timeout>] }
              compatible-version {1|2}
              description <description-string>
              guard-time <time>
              node-id <xx:xx:xx:xx:xx:xx or xx-xx-xx-xx-xx-xx>
              path {a|b} [ {<port-id>|<lag-id>]} raps-tag <qtag1>[.<qtag2>] ]
                    description <description-string>
                    eth-cfm
                          mep <mep-id> domain <md-index> association <ma-index>
                          ...
                    rpl-end
                    shutdown
              revert-time <time>
              rpl-node {owner|nbr}
              shutdown
              sub-ring {virtual-link|non-virtual-link}
                  interconnect [ring-id <ring-index>|vpls]
                      propagate-topology-change
```

Parameters:

- <ring-index> — This is the number by which the ring is referenced, values: 1 to128.
- ccm-hold-time { [down <down-timeout>] [up <up-timeout>] }
  - **down** — This command specifies the timer which controls the delay between detecting that ring path is down and reporting it to the G.8032 protection module. If a non-zero value is configured, the system will wait for the time specified in the value parameter before reporting it to the G.8032 protection module. This parameter applies only to ring path CCM. It does not apply to the ring port link state. To dampen ring port link state transitions, use the hold-time parameter from the physical member port. This is useful if the underlying path between two nodes is going across an optical system which implements its own protection.
  - **up** — This command specifies the timer which controls the delay between detecting that ring path is up and reporting it to the G.8032 protection module. If a non-zero value is configured, the system will wait for the time specified in the value parameter before reporting it to the G.8032 protection module. This parameter applies only to ring path CCM. It does not apply to the member port link state. To dampen member port link state transitions, use the hold-time parameter from the physical member port.

    Values:
    <down-timeout>   : [0..5000] in centiseconds - Default: 0
    <up-timeout>   : [0..5000] in deciseconds - Default: 20
    1 centisecond = 10ms
    1 decisecond = 100ms
- **compatible-version** — This command configures eth-ring compatibility version for the G.8032 state machine and messages. The default is version 2 (ITU G.8032v2) and all 7x50 systems use version 2. If there is a need to interwork with third party devices that only support version 1, this can be set to version 1 allowing the reception of version 1 PDUs. Version 2 is encoded as 1 in the R-APS messages. Compatibility allows the reception of version 1 (encoded as 0) R-APS PDUs but, as per the G.8032 specification, higher versions are ignored on reception. For SR OS, messages are always originated with version 2. Therefore if a third party switch supported version 3 (encoded as 2) or higher, interworking is also supported provided the other switch is compatible with version 2.
- **description** <*description-string*> — This configures a text string, up to 80 characters, which can be used to describe the use of the eth-ring.

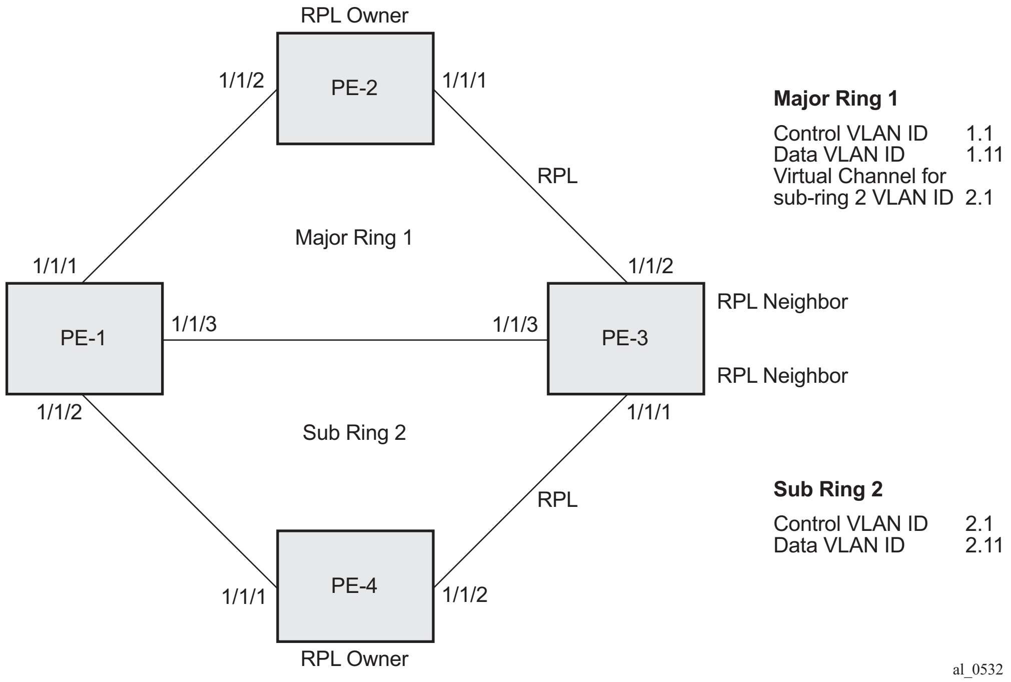
- **guard-time** <*time*> — The forwarding method, in which R-APS messages are copied and forwarded at every Ethernet ring node, can result in a message corresponding to an old request, that is no longer relevant, being received by Ethernet ring nodes. Reception of an old R-APS message may result in erroneous ring state interpretation by some Ethernet ring nodes. The guard timer is used to prevent Ethernet ring nodes from acting upon outdated R-APS messages and prevents the possibility of forming a closed loop. Messages are not forwarded when the guard-timer is running.

  Values: [1..20] in deciseconds - Default: 5
  1 decisecond = 100ms
- **node-id** <xx:xx:xx:xx:xx:xx> or <xx-xx-xx-xx-xx-xx> — This allows the node identifier to be explicitly configured. By default the chassis MAC is used. Not required in typical configurations.
- **path** {**a**|**b**} [{<*port-id*>|*lag-id*} **raps-tag** <*qtag*[.<*qtag*>]] — This parameter defines the paths around the ring, of which there are two in different directions on the ring: an “a” path and a “b” path, except on the interconnection node where a sub-ring connects to another major/sub ring in which case there is one path (either a or b) configured together with the **sub-ring** command. The paths are configured on a dot1Q or QinQ encapsulated access or hybrid port or a LAG with the encapsulation used for the R-APS messages on the ring. These can be either single or double tagged.
  - **description** <*description-string*> — This configures a text string, up to 80 characters, which can be used to describe the use of the path.
  - **eth-cfm** — Configures the associated Ethernet CFM parameters.
    - **mep** <*mep-id*> **domain** <*md-index*> **association** <*ma-index*> — The MEP defined under the path is used for the G.8032 protocol messages, which are based on IEEE 802.1ag/Y.1731 CFM frames.
  - **rpl-end** — When configured, this path is expected to be one end of the RPL. This parameter must be configured in conjunction with the *rpl-node* parameter.
  - **shutdown** — This command shuts down the path.
- **revert-time** <*time*> — This command configures the revert time for an Eth-Ring. Revert time is the time that the RPL will wait before returning to the blocked state. Configuring **no revert-time** disables reversion, effectively setting the revert-time to zero.
  Values: [60..720] in seconds - Default: 300

- **rpl-node** {**owner**|**nbr**} — A node can be designated as either the **owner** of the RPL, in which case this node is responsible for the RPL, or the **nbr** (neighbor), in which case this node is expected to be the neighbor to the RPL owner across the RPL. The **nbr** is optional and is included to be compliant with the specification. This parameter must be configured in conjunction with the **rpl-end** command. On a sub-ring without virtual channel it is mandatory to configure **sub-ring non-virtual-link** on all nodes on the sub-ring to ensure propagation of the R-APS messages around the sub-ring.
- **shutdown** — This command shuts down the ring.
- **sub-ring** {**virtual-link**|**non-virtual-link**} — This command is configured on the interconnection node between the sub-ring and its major/sub ring to indicate that this ring is a sub-ring. The parameter specifies whether it uses a virtual link through the major/sub ring for the R-APS messages or not. A ring configured as a sub-ring can only be configured with a single path.
  - **interconnect** [**ring-id** <*ring-index*>|**vpls**] — A sub-ring connects to either another ring or a VPLS service. If it connects to another ring (either a major ring or another sub-ring), the ring identifier must be specified and the ring to which it connects must be configured with both a path “a” and a path “b”, meaning that it is not possible to connect a sub-ring to another sub-ring on an interconnection node.
    Alternatively, the **vpls** parameter is used to indicate the sub-ring connects to a VPLS service. Interconnection using a VPLS service requires the sub-ring to be configured with **non-virtual-link**.
    - **propagate-topology-change** — If a topology change event happens in the sub-ring, it can be optionally propagated with the use of this parameter to either the major/sub ring it is connected to, using R-APS messages, or to the LDP VPLS SDP peers using an LDP “flush-all-from-me” message if the sub-ring is connected to a VPLS service.

The example topology is shown in Figure 4.

Figure 4 **Ethernet Test Topology**



The configuration is divided into the following sections:

- A sub-ring connected to a major ring using a virtual link through the major ring.
- A sub-ring connected to a major ring without a virtual link.
- A sub-ring connected to a VPLS service (without a virtual link).

Configure a Sub-Ring to a Major Ring with a Virtual Link

To configure an Ethernet ring using R-APS, there will be at least 2 VPLS services required for one Eth-Ring instance, one for the control channel and the other (or more) for data channel(s). The control channel is used for R-APS signaling while the data channel is for user data traffic. The state of the data channels is inherited from the state of the control channel.

The following needs to be configured:

- Encapsulation type for each ring port
- ETH-CFM
- Eth-ring for major ring 1

- Eth-ring for sub-ring 2
- Control channel service and add Eth-ring SAPs
- User data channels

Configure the Encapsulation for the Ring Ports.

Eth-Ring needs an R-APS tag to send/receive G.8032 signaling messages. To configure a control channel, an access SAP configuration is required on each path a/b port. The SAP configuration follows that of the port and must be either dot1Q or QinQ, consequently the control and data packets are either single tagged or double tagged. Single tagged control frames are supported on a QinQ port by configuring the system with the new-qinq-untagged-sap parameter (configure system ethernet new-qinq-untagged-sap), and the ring path raps-tags and control VPLS SAPs configured as qtag.0.

In this example, QinQ tags are used. The commands for the major and sub rings ring, on PE-1 for example, are:

```
*A:PE-1# configure port 1/1/1 ethernet mode access
*A:PE-1# configure port 1/1/2 ethernet mode access
*A:PE-1# configure port 1/1/3 ethernet mode access
*A:PE-1# configure port 1/1/1 ethernet encap-type qinq
*A:PE-1# configure port 1/1/2 ethernet encap-type qinq
*A:PE-1# configure port 1/1/3 ethernet encap-type qinq
```

Configuring ETH-CFM.

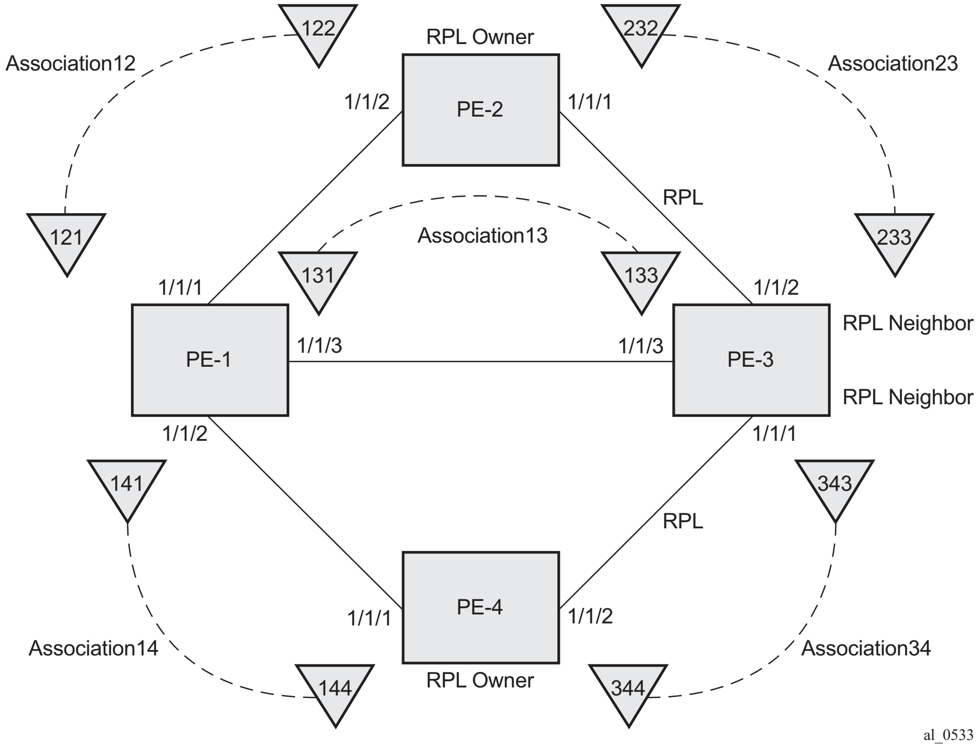
Configuring ETH-CFM domain, association and MEP is required before configuring Ethernet ring. The standard domain format is *none* and the association name should beITU Carrier Code-based (ICC-based - Y.1731), however, the SROS implementation is flexible in that it supports both IEEE and ICC formats. The *eth-ring* MEP requires a CCM interval with values such as 1s, 100ms or 10ms to be configured.

The MEPs used for R-APS control normally will have CCM configured on the control channel path MEPs for failure detection. Alternatively, detecting a failure of the ring may be achieved by running Ethernet in the First Mile (EFM) at the port level if CCM is not possible at 1s, 100ms or 10ms. Also rings can be run without CFM although the ETH-CFM association must be configured for R-APS messages to be exchanged. To omit the failure detecting CCMs, it would be necessary to remove the *ccm-enable* from under the path MEPs and to remove the *remote-mepid* on the corresponding ETH CFM configuration.

Loss-of-signal, in conjunction with other OAM mechanisms, is applicable only when the nodes are directly connected.

Figure 5 shows the details of the MEPs and their associations configured when both the major and sub rings are used. The associations only need to be pair wise unique but for clarity five unique associations are used. Any name format can be used, but it must be consistent on both adjacent nodes.

Figure 5 **ETH-CFM MEP Associations**



The configuration of ETH-CFM for the major and sub rings on each node is as follows. The CCMs for failure detection are configured for 1 second intervals.

Ring node PE-1: Association 12 and 13 are used for the major ring and association 14 is used for the sub-ring.

```
*A:PE-1# configure
    eth-cfm
        domain 1 format none level 2
            association 12 format icc-based name "Association12"
                ccm-interval 1
                remote-mepid 122
            exit
            association 13 format icc-based name "Association13"
                ccm-interval 1
                remote-mepid 133
            exit
            association 14 format icc-based name "Association14"
                ccm-interval 1
                remote-mepid 144
            exit
        exit
    exit
```

Ring node PE-2: Association 12 and 23 are used for the major ring.

```
*A:PE-2# configure
    eth-cfm
        domain 1 format none level 2
            association 12 format icc-based name "Association12"
                ccm-interval 1
                remote-mepid 121
            exit
            association 23 format icc-based name "Association23"
                ccm-interval 1
                remote-mepid 233
            exit
        exit
    exit
```

Ring node PE-3: Association 23 and 13 are used for the major ring and association 34 is used for the sub-ring.

```
*A:PE-3# configure
 eth-cfm
        domain 1 format none level 2
            association 13 format icc-based name "Association13"
                ccm-interval 1
                remote-mepid 131
            exit
            association 23 format icc-based name "Association23"
                ccm-interval 1
                remote-mepid 232
            exit
            association 34 format icc-based name "Association34"
                ccm-interval 1
                remote-mepid 344
            exit
        exit
    exit
```

Ring node PE-4: Association 14 and 34 are used for the sub-ring.

```
*A:PE-4# configure
    eth-cfm
        domain 1 format none level 2
            association 14 format icc-based name "Association14"
                ccm-interval 1
                remote-mepid 141
            exit
            association 34 format icc-based name "Association34"
                ccm-interval 1
                remote-mepid 343
            exit
        exit
    exit
```

Configuring Eth-Ring – Major Ring 1

Two paths must be configured to form a ring. In this example, VLAN tag 1.1 is used as control channel for R-APS signaling for the major ring (ring 1) on the ports shown in Figure 4 using the ETH CFM information shown in Figure 5. The revert-time is set to its minimum value and CCM messages are enabled on the MEP. The **control-mep** parameter is required to indicate that this MEP is used for ring R-APS messages.

Ring node PE-1:

```
*A:PE-1# configure
    eth-ring 1
        description "Ethernet Ring 1"
        revert-time 60
        path a 1/1/1 raps-tag 1.1
            description "Ethernet Ring 1 - PathA"
            eth-cfm
                mep 121 domain 1 association 12
                    ccm-enable
                    control-mep
                    no shutdown
                exit
            exit
            no shutdown
        exit
        path b 1/1/3 raps-tag 1.1
            description "Ethernet Ring 1 - PathB"
            eth-cfm
                mep 131 domain 1 association 13
                    ccm-enable
                    control-mep
                    no shutdown
                exit
            exit
            no shutdown
        exit
        no shutdown
    exit
```

It is mandatory to configure a MEP in the path context, otherwise this error will be displayed.

```
*A:PE-1>config>eth-ring>path# no shutdown
INFO: ERMGR #1001 Not permitted - must configure eth-cfm MEP first
```

While MEPs are mandatory, enabling CCMs on the MEPs under the paths as a failure detection mechanism is optional as explained earlier.

Ring node PE-2: This is configured as the RPL owner with the RPL being on path “a” as indicated by the **rpl-end** parameter.

```
*A:PE-2# configure
    eth-ring 1
        description "Ethernet Ring 1"
        revert-time 60
        rpl-node owner
        path a 1/1/1 raps-tag 1.1
            description "Ethernet Ring 1 - PathA"
            rpl-end
            eth-cfm
                mep 232 domain 1 association 23
                    ccm-enable
                    control-mep
                    no shutdown
                exit
            exit
            no shutdown
        exit
        path b 1/1/2 raps-tag 1.1
            description "Ethernet Ring 1 - PathB"
            eth-cfm
                mep 122 domain 1 association 12
                    ccm-enable
                    control-mep
                    no shutdown
                exit
            exit
            no shutdown
        exit
        no shutdown
    exit
```

It is not permitted to configure a path as an RPL end without having configured the node on this ring to be either the RPL *owner* or *nbr* otherwise the following error message is reported.

```
*A:PE-2>config>eth-ring>path# rpl-end
INFO: ERMGR #1001 Not permitted - path-type rpl-end is not consistent with eth-ring
'rpl-node' type
```

Ring node PE-3: This is configured as the RPL neighbor with the RPL being on path “b” as indicated by the **rpl-end** parameter.

```
*A:PE-3# configure
```

```
eth-ring 1
    description "Ethernet Ring 1"
    revert-time 60
    rpl-node nbr
    path a 1/1/3 raps-tag 1.1
        description "Ethernet Ring 1 - PathA"
        eth-cfm
            mep 133 domain 1 association 13
                ccm-enable
                control-mep
                no shutdown
            exit
        exit
        no shutdown
    exit
    path b 1/1/2 raps-tag 1.1
        description "Ethernet Ring 1 - PathB"
        rpl-end
        eth-cfm
            mep 233 domain 1 association 23
                ccm-enable
                control-mep
                no shutdown
            exit
        exit
        no shutdown
    exit
    no shutdown
exit
```

The link between PE-2 and PE-3 will be the RPL with PE-2 and PE-3 blocking that link when the ring is fully operational. In this example, the RPL is using path “a” on PE-2 and path “b” on PE-3.

Configuring Eth-Ring – Sub-Ring 2

Ring nodes PE-1, PE-3, and PE-4 form a sub-ring. The sub-ring attaches to the major ring (ring 1). The sub-ring in this case will use a virtual-link. The interconnection ring instance identifier (*ring-id*) is specified and *propagate-topology-change* indicates that sub-ring flushing will be propagated to the major ring. Only one path is specified since the other path is not required at an interconnection node. Sub-rings are almost identical to major rings in operation except that sub-rings send MAC flushes towards their connected ring (either a major or sub ring). Major or sub rings never send MAC flushes to their sub-rings. Therefore a couple of sub-rings connected to a major ring can cause MACs to flush on the major ring but the major ring will not propagate a sub-ring MAC flush to other sub-rings.

Ring node PE-1 provides an interconnection between the major ring (1) and the sub-ring (2). Ring 2 is configured to be a sub-ring which interconnects to ring 1. It will use a virtual link on ring 1 to send R-APS messages to the other interconnection node and topology changes will be propagated from sub-ring 2 to the major ring 1.

```
*A:PE-1# configure
    eth-ring 2
        description "Ethernet Sub-ring 2 on Major Ring 1"
        revert-time 60
        sub-ring virtual-link
            interconnect ring-id 1
                propagate-topology-change
            exit
        exit
        path a 1/1/2 raps-tag 2.1
            description "Ethernet Ring 2 - PathA"
            eth-cfm
                mep 141 domain 1 association 14
                    ccm-enable
                    control-mep
                    no shutdown
                exit
            exit
            no shutdown
        exit
        no shutdown
```

The configuration of PE-3 is similar to PE-1, but PE-3 is the RPL neighbor, with the RPL end on path “a”, for the RPL between PE-3 and PE-4.

```
*A:PE-3# configure
    eth-ring 2
        description "Ethernet Sub-ring 2 on Major Ring 1"
        revert-time 60
        rpl-node nbr
        sub-ring virtual-link
            interconnect ring-id 1
                propagate-topology-change
            exit
        exit
        path a 1/1/1 raps-tag 2.1
            description "Ethernet Ring 2 - PathA"
            rpl-end
            eth-cfm
                mep 343 domain 1 association 34
                    ccm-enable
                    control-mep
                    no shutdown
                exit
            exit
            no shutdown
        exit
        no shutdown
```

Ring node PE-4: This node only has configuration for the sub-ring, ring 2. It is also the RPL owner, with path “b” being the RPL end, for the RPL between PE-3 and PE-4.

```
*A:PE-4# configure
    eth-ring 2
        description "Ethernet Sub-ring 2"
        revert-time 60
        rpl-node owner
        exit
        path a 1/1/1 raps-tag 2.1
            description "Ethernet Ring 2 - PathA"
            eth-cfm
                mep 144 domain 1 association 14
                    ccm-enable
                    control-mep
                    no shutdown
                exit
            exit
            no shutdown
        exit
        path b 1/1/2 raps-tag 2.1
            description "Ethernet Ring 2 - PathB"
            rpl-end
            eth-cfm
                mep 344 domain 1 association 34
                    ccm-enable
                    control-mep
                    no shutdown
                exit
            exit
            no shutdown
        exit
        no shutdown
```

Until the Ethernet Ring instance is attached to a VPLS service, the ring operational status is down and the forwarding status of each port is blocked. This prevents the operator from creating a loop by mis-configuration. This state can be seen on ring node PE-1 as follows:

```
*A:PE-1# show eth-ring 1
===============================================================================
Ethernet Ring 1 Information
===============================================================================
Description        : Ethernet Ring 1
Admin State        : Up                 Oper State       : Down
Node ID            : 4a:c4:ff:00:00:00
Guard Time         :    5 deciseconds   RPL Node         : rplNone
Max Revert Time    :   60 seconds       Time to Revert   : N/A
CCM Hold Down Time :    0 centiseconds  CCM Hold Up Time :   20 deciseconds
Compatible Version : 2
APS Tx PDU         : Request State: 0xB
                     Sub-Code     : 0x0
                     Status       : 0x20  ( BPR )
                     Node ID      : 4a:c4:ff:00:00:00
```

```
Defect Status      :
Sub-Ring Type      : none
-------------------------------------------------------------------------------
Ethernet Ring Path Summary
-------------------------------------------------------------------------------
Path Port     Raps-Tag     Admin/Oper      Type            Fwd State
-------------------------------------------------------------------------------
  a  1/1/1    1.1             Up/Down      normal          blocked
  b  1/1/3    1.1             Up/Down      normal          blocked
===============================================================================
*A:PE-1#
```

Configure the Control Channel VPLS Service

Path “a” and “b” configured in the eth-ring must be added as SAPs into a VPLS service (standard VPLS) using the **eth-ring** parameter. The SAP encapsulation values must match the values of the *raps-tag* configured for the associated path.

G.8032 uses the same raps-tag value on all nodes on the ring, as configured in this example. However, the SR OS implementation relaxes this constraint by requiring the tag to match only on adjacent nodes.

A VPLS service (identifier 1) is configured on PE-1, PE-2 and PE-3 for the control channel for the major ring (ring1), and another VPLS service (identifier 2) is used on PE-1, PE-3 and PE-4 for the sub-ring (ring 2).

Ring node PE-1: Control service for the major ring.

```
*A:PE-1# configure
    service
        vpls 1 customer 1 create
            description "Control VID 1.1 for Ring 1 - Major Ring"
            sap 1/1/1:1.1 eth-ring 1 create
            exit
            sap 1/1/3:1.1 eth-ring 1 create
            exit
            no shutdown
```

Ring node PE-2: Control service for the major ring.

```
*A:PE-2# configure
    service
        vpls 1 customer 1 create
            description "Control VID 1.1 for Ring 1 - Major Ring"
            sap 1/1/1:1.1 eth-ring 1 create
            exit
            sap 1/1/2:1.1 eth-ring 1 create
            exit
            no shutdown
```

Ring node PE-3: Control service for the major ring.

```
*A:PE-3# configure
    service
        vpls 1 customer 1 create
            description "Control VID 1.1 for Ring 1 - Major Ring"
            sap 1/1/2:1.1 eth-ring 1 create
            exit
            sap 1/1/3:1.1 eth-ring 1 create
            exit
            no shutdown
```

SAPs or SDPs can be added to a control channel VPLS on condition the **eth-ring** parameter is present. Trying to add a SAP without this parameter to a control channel VPLS will result in the following message being displayed.

```
*A:PE-1# configure service vpls 1 sap 1/2/1:1 create
MINOR: SVCMGR #1321 Service contains an Ethernet ring control SAP
```

For the sub-ring, the configuration of a split horizon group for the virtual channel on the major ring on the interconnection nodes is recommended. This avoids the looping of control R-APS messages in the case there is a mis-configuration in the major ring.

Ring node PE-1: Control service for the sub-ring. Notice that two SAPs connect to the major ring (ring 1), these being for the virtual channel, and the third SAP connects to the sub-ring (ring 2).

```
*A:PE-1# configure
    service
        vpls 2 customer 1 create
            description "Control/Virtual Channel VID 2.1 for Ring 2"
            split-horizon-group "shg-ring2" create
            exit
            sap 1/1/1:2.1 split-horizon-group "shg-ring2" eth-ring 1 create
                description "Ring 2 Interconnection using Ring 1"
            exit
            sap 1/1/2:2.1 eth-ring 2 create
            exit
            sap 1/1/3:2.1 split-horizon-group "shg-ring2" eth-ring 1 create
                description "Ring 2 Interconnection using Ring 1"
            exit
            no shutdown
```

Ring node PE-2: Control service for the sub-ring. Sub-ring 2 is not present on PE-2, however, its virtual channel on major ring 1 needs to exist throughout ring 1.

```
*A:PE-2# configure
    service
```

```
        vpls 2 customer 1 create
            description "Virtual Channel VID 2.1 for Ring 2"
            sap 1/1/1:2.1 eth-ring 1 create
            exit
            sap 1/1/2:2.1 eth-ring 1 create
            exit
            no shutdown
```

If multiple virtual channels are used (due to the aggregation of multiple sub-rings into the same major ring), their configuration could be simplified on non-interconnection nodes on the major ring. To achieve this on a ring node such as PE-2, a default SAP could be used rather than configuring a VPLS per virtual channel. If QinQ SAPs are used then a default SAP of 1/1/[1,2]:qtag.* could be used but requires all control channels for sub-rings to be using qtag as the outer VLAN ID, or 1/1/[1,2]:* if dot1Q SAPs were used. This is because the SAPs match explicit SAPs definitions first and the default SAP will handle any other traffic.

Ring node PE-3: Control service for the sub-ring. This is similar to the configuration of PE-1.

```
*A:PE-3# configure
    service
        vpls 2 customer 1 create
            description "Control/Virtual Channel VID 2.1 for Ring 2"
            split-horizon-group "shg-ring2" create
            exit
            sap 1/1/1:2.1 eth-ring 2 create
            exit
            sap 1/1/2:2.1 split-horizon-group "shg-ring2" eth-ring 1 create
                description "Ring 2 Interconnection using Ring 1"
            exit
            sap 1/1/3:2.1 split-horizon-group "shg-ring2" eth-ring 1 create
                description "Ring 2 Interconnection using Ring 1"
            exit
            no shutdown
```

Ring node PE-4: Control service for the sub-ring. Both SAPs are configured on the sub-ring
(ring 2).

```
*A:PE-4# configure
    service
        vpls 2 customer 1 create
            description "Control VID 2.1 for Ring 2 Sub-ring"
            sap 1/1/1:2.1 eth-ring 2 create
            exit
            sap 1/1/2:2.1 eth-ring 2 create
            exit
            no shutdown
```

At this point, the Eth-Ring 1 is operationally up and the RPL is blocking successfully on ring node PE-2 port 1/1/1, as expected for the RPL owner/end configuration and on port 1/1/2 on PE-3 as the RPL neighbor.

Show Output

An overview of all of the rings can be shown using the following commands, in this case on
PE-1.

First, the ETH ring status is shown.

```
*A:PE-1# show eth-ring status
===============================================================================
Ethernet Ring (Status information)
===============================================================================
Ring   Admin  Oper       Path Information                 MEP Information
ID     State  State  Path         Tag       State     Ctrl-MEP CC-Intvl Defects
-------------------------------------------------------------------------------
1      Up     Up     a - 1/1/1        1.1   Up        Yes      1         -----
                     b - 1/1/3        1.1   Up        Yes      1         -----
2      Up     Up     a - 1/1/2        2.1   Up        Yes      1         -----
                     b - N/A                 -        -        -         -----
===============================================================================
Ethernet Tunnel MEP Defect Legend:
R = Rdi, M = MacStatus, C = RemoteCCM, E = ErrorCCM, X = XconCCM
*A:PE-1#
```

It is expected that the state is “up”, even on ring paths which are blocked. The “Defects” column refers to the CFM defects of the MEPs. If there is a problem, these will be flagged.

The following output shows the ring and path forwarding states.

```
*A:PE-1# show eth-ring
===============================================================================
Ethernet Rings (summary)
===============================================================================
Ring Int  Admin Oper            Paths Summary                       Path States
ID   ID   State State                                               a     b
-------------------------------------------------------------------------------
1    -    Up    Up    a - 1/1/1        1.1   b - 1/1/3        1.1   U     U
2    1    Up    Up    a - 1/1/2        2.1   b - Not configured     U     -
===============================================================================
Ethernet Ring Summary Legend:   B - Blocked    U - Unblocked
*A:PE-1#
```

The specific ring information can be shown as follows.

Ring node PE-1:

```
*A:PE-1# show eth-ring 1
===============================================================================
Ethernet Ring 1 Information
===============================================================================
Description        : Ethernet Ring 1
Admin State        : Up                   Oper State       : Up
Node ID            : 4a:c4:ff:00:00:00
Guard Time         :    5 deciseconds   RPL Node         : rplNone
Max Revert Time    :   60 seconds       Time to Revert   : N/A
CCM Hold Down Time :    0 centiseconds  CCM Hold Up Time :   20 deciseconds
Compatible Version : 2
APS Tx PDU         : N/A
Defect Status      :
Sub-Ring Type      : none
-------------------------------------------------------------------------------
Ethernet Ring Path Summary
-------------------------------------------------------------------------------
Path Port     Raps-Tag     Admin/Oper      Type            Fwd State
-------------------------------------------------------------------------------
  a  1/1/1    1.1             Up/Up        normal          unblocked
  b  1/1/3    1.1             Up/Up        normal          unblocked
===============================================================================
*A:PE-1#
```

The status around the major ring can also be checked.

Ring node PE-2: Major ring.

```
*A:PE-2# show eth-ring 1
===============================================================================
Ethernet Ring 1 Information
===============================================================================
Description        : Ethernet Ring 1
Admin State        : Up                   Oper State       : Up
Node ID            : 4a:c5:ff:00:00:00
Guard Time         :    5 deciseconds   RPL Node         : rplOwner
Max Revert Time    :   60 seconds       Time to Revert   : N/A
CCM Hold Down Time :    0 centiseconds  CCM Hold Up Time :   20 deciseconds
Compatible Version : 2
APS Tx PDU         : Request State: 0x0
                     Sub-Code     : 0x0
                     Status       : 0x80  ( RB )
                     Node ID      : 4a:c5:ff:00:00:00
Defect Status      :

Sub-Ring Type      : none

-------------------------------------------------------------------------------
Ethernet Ring Path Summary
-------------------------------------------------------------------------------
Path Port     Raps-Tag     Admin/Oper      Type            Fwd State
-------------------------------------------------------------------------------
  a  1/1/1    1.1             Up/Up        rplEnd          blocked
  b  1/1/2    1.1             Up/Up        normal          unblocked
```

```
===============================================================================
*A:PE-2#
```

PE-2 is the RPL owner with port 1/1/1 as an RPL end, which is blocked as expected. The *revert-time* is also shown to be the configured value. Detailed information is shown relating to the R-APS PDUs being transmitted on this ring as this node is the RPL owner.

When a revert is pending, the “Time to Revert” will show the number of seconds remaining before the revert occurs.

Ring node PE-3: Major ring.

```
*A:PE-3# show eth-ring 1
===============================================================================
Ethernet Ring 1 Information
===============================================================================
Description        : Ethernet Ring 1
Admin State        : Up                 Oper State       : Up
Node ID            : 4a:c6:ff:00:00:00
Guard Time         :    5 deciseconds   RPL Node         : rplNeighbor
Max Revert Time    :   60 seconds       Time to Revert   : N/A
CCM Hold Down Time :    0 centiseconds  CCM Hold Up Time :   20 deciseconds
Compatible Version : 2
APS Tx PDU         : N/A
Defect Status      :

Sub-Ring Type      : none

-------------------------------------------------------------------------------
Ethernet Ring Path Summary
-------------------------------------------------------------------------------
Path Port     Raps-Tag     Admin/Oper      Type            Fwd State
-------------------------------------------------------------------------------
  a  1/1/3    1.1             Up/Up        normal          unblocked
  b  1/1/2    1.1             Up/Up        rplEnd          blocked
===============================================================================
*A:PE-3#
```

PE-3 is the RPL neighbor with port 1/1/2 as an RPL end which is blocked as expected.

The information for the sub-ring can also be shown using a similar command.

Ring node PE-1: Sub-ring.

```
*A:PE-1# show eth-ring 2
===============================================================================
Ethernet Ring 2 Information
===============================================================================
Description        : Ethernet Sub-ring 2 on Major Ring 1
Admin State        : Up                 Oper State       : Up
```

```
Node ID            : 4a:c4:ff:00:00:00
Guard Time         :    5 deciseconds   RPL Node         : rplNone
Max Revert Time    :   60 seconds       Time to Revert   : N/A
CCM Hold Down Time :    0 centiseconds  CCM Hold Up Time :   20 deciseconds
Compatible Version : 2
APS Tx PDU         : N/A
Defect Status      :

Sub-Ring Type      : virtualLink        Interconnect-ID  : 1
Topology Change    : Propagate

-------------------------------------------------------------------------------
Ethernet Ring Path Summary
-------------------------------------------------------------------------------
Path Port     Raps-Tag     Admin/Oper      Type            Fwd State
-------------------------------------------------------------------------------
  a  1/1/2    2.1             Up/Up        normal          unblocked
  b  -        -                 -/-         -               -
===============================================================================
*A:PE-1#
```

Only path “a” is active and unblocked. The second path, path “b” is not configured as only one path is required on an interconnection node. The “Sub-Ring Type” is shown to be a virtual link interconnecting to ring 1, with topology propagation enabled.

Ring node PE-3: Sub-ring.

```
*A:PE-3# show eth-ring 2
===============================================================================
Ethernet Ring 2 Information
===============================================================================
Description        : Ethernet Sub-ring 2 on Major Ring 1
Admin State        : Up                 Oper State       : Up
Node ID            : 4a:c6:ff:00:00:00
Guard Time         :    5 deciseconds   RPL Node         : rplNeighbor
Max Revert Time    :   60 seconds       Time to Revert   : N/A
CCM Hold Down Time :    0 centiseconds  CCM Hold Up Time :   20 deciseconds
Compatible Version : 2
APS Tx PDU         : N/A
Defect Status      :
Sub-Ring Type      : virtualLink        Interconnect-ID  : 1
Topology Change    : Propagate
-------------------------------------------------------------------------------
Ethernet Ring Path Summary
-------------------------------------------------------------------------------
Path Port     Raps-Tag     Admin/Oper      Type            Fwd State
-------------------------------------------------------------------------------
  a  1/1/1    2.1             Up/Up        rplEnd          blocked
  b  -        -                 -/-         -               -
===============================================================================
*A:PE-3#
```

PE-3 is the RPL neighbor with port 1/1/1 as an RPL end, which is blocked as expected.

Ring Node PE-4: Sub-ring.

```
*A:PE-4# show eth-ring 2

===============================================================================
Ethernet Ring 2 Information
===============================================================================
Description        : Ethernet Sub-ring 2
Admin State        : Up                 Oper State       : Up
Node ID            : 4a:c7:ff:00:00:00
Guard Time         :    5 deciseconds   RPL Node         : rplOwner
Max Revert Time    :   60 seconds       Time to Revert   : N/A
CCM Hold Down Time :    0 centiseconds  CCM Hold Up Time :   20 deciseconds
Compatible Version : 2
APS Tx PDU         : Request State: 0x0
                     Sub-Code     : 0x0
                     Status       : 0xE0  ( RB DNF BPR )
                     Node ID      : 4a:c7:ff:00:00:00
Defect Status      :

Sub-Ring Type      : none

-------------------------------------------------------------------------------
Ethernet Ring Path Summary
-------------------------------------------------------------------------------
Path Port     Raps-Tag     Admin/Oper      Type            Fwd State
-------------------------------------------------------------------------------
  a  1/1/1    2.1            Up/Up        normal          unblocked
  b  1/1/2    2.1            Up/Up        rplEnd          blocked
===============================================================================
*A:PE-4#
```

PE-4 is the RPL owner with port 1/1/2 as an RPL end, which is blocked as expected.

The details of an individual path can be shown.

```
*A:PE-1# show eth-ring 1 path a

===============================================================================
Ethernet Ring 1 Path Information
===============================================================================
Description        : Ethernet Ring 1 - PathA
Port               : 1/1/1              Raps-Tag         : 1.1
Admin State        : Up                 Oper State       : Up
Path Type          : normal             Fwd State        : unblocked
                                        Fwd State Change : 05/09/2016 08:35:29
Last Switch Command: noCmd
APS Rx PDU         : Request State: 0x0
                     Sub-Code     : 0x0
                     Status       : 0x80  ( RB )
                     Node ID      : 4a:c5:ff:00:00:00

===============================================================================
*A:PE-1#
```

The ring hierarchy created can be shown, either for all rings, or as below for a specific ring.

```
*A:PE-1# show eth-ring 1 hierarchy

===============================================================================
Ethernet Ring 1 (hierarchy)
===============================================================================
Ring Int  Admin Oper           Paths Summary                       Path States
ID   ID   State State                                              a     b
-------------------------------------------------------------------------------
1    -    Up    Up    a - 1/1/1        1.1   b - 1/1/3        1.1   U     U
2    1    Up    Up    a - 1/1/2        2.1   b - Not configured     U     -
===============================================================================
Ethernet Ring Summary Legend:   B - Blocked    U - Unblocked
*A:PE-1#
```

Configuring the User Data Channel VPLS Service

The user data channels are created on a separate VPLS, VPLS 11 in this example, using VLAN tag 1.11. The ring data channels must be on the same ports as the corresponding control channels configured above. The access into the data services can use normal SAPs and/or SDPs, for example the SAP on port 1/2/1 in the following output. Customer data traverses the ring on a data SAP. Multiple parallel data SAPs in different data services can be controlled by one control ring instance (eth-ring 1 in the example).

Ring node PE-1: Two data SAPs correspond to the major ring 1, while the third SAP is the data SAP on the sub-ring 2.

```
*A:PE-1# configure
    service
        vpls 11 customer 1 create
            description "Data VPLS"
            sap 1/1/1:1.11 eth-ring 1 create
            exit
            sap 1/1/2:1.11 eth-ring 2 create
            exit
            sap 1/1/3:1.11 eth-ring 1 create
            exit
            sap 1/2/1:11 create
                description "Sample Customer Service SAP"
            exit
            no shutdown
```

Ring node PE-3 (not shown) would be similar to ring node 1.

Ring node PE-2 is also similar with a single ring data service using VPLS 11 and tag 1.11.

```
*A:PE-2# configure
    service
        vpls 11 customer 1 create
            description "Data VPLS"
            sap 1/1/1:1.11 eth-ring 1 create
            exit
            sap 1/1/2:1.11 eth-ring 1 create
            exit
            sap 1/2/1:11 create
                description "Sample Customer Service SAP"
            exit
            no shutdown
```

Ring node PE- 4: On ring node PE-4 the data VLAN ID is configured as a normal ring data VPLS on ring 2.

```
*A:PE-4# configure
    service
        vpls 11 customer 1 create
            description "Data VPLS"
            sap 1/1/1:1.11 eth-ring 2 create
            exit
            sap 1/1/2:1.11 eth-ring 2 create
            exit
            sap 1/2/1:11 create
                description "Sample Customer Service SAP"
            exit
            no shutdown
```

All of the SAPs which are configured to use Ethernet rings can be shown. The following output is taken from PE-1, where there are:

- two SAPs in VPLS 1 for the control channel of ring 1 (VLAN ID 1.1)
- two SAPs in VPLS 2 on ring 1 for the virtual channel for ring 2 (VLAN ID 2.1).
- one SAP in VPLS 2 on ring 2 for the control channel for ring 2 (VLAN ID 2.1)
- three SAPs in VPLS 11, two on ring 1 and one on ring 2, for the data service (VLAN ID 1.11). This matches the information in Figure 3.

```
*A:PE-1# show service sap-using eth-ring

===============================================================================
Service Access Points (Ethernet Ring)
===============================================================================
SapId              SvcId             Eth-Ring Path Admin Oper  Blocked Control/
                                                   State State         Data
-------------------------------------------------------------------------------
1/1/1:1.1          1                 1        a    Up    Up    No      Ctrl
1/1/3:1.1          1                 1        b    Up    Up    No      Ctrl
1/1/1:2.1          2                 1        a    Up    Up    No      Ctrl
1/1/2:2.1          2                 2        a    Up    Up    No      Ctrl
1/1/3:2.1          2                 1        b    Up    Up    No      Ctrl
1/1/1:1.11         11                1        a    Up    Up    No      Data
1/1/2:1.11         11                2        a    Up    Up    No      Data
```

```
1/1/3:1.11         11                1        b    Up    Up    No      Data
-------------------------------------------------------------------------------
Number of SAPs : 8
===============================================================================
*A:PE-1#
```

Statistics are available showing both the CCM and R-APS messages sent and received on a node. An associated **clear** command is available.

```
*A:PE-1# show eth-cfm statistics
===============================================================================
ETH-CFM System Statistics
===============================================================================
Rx Count           : 1201               Tx Count           : 1066
Dropped Congestion : 0                  Discarded Error    : 0
AIS Currently Act  : 0                  AIS Currently Fail : 0
===============================================================================
=================================
ETH-CFM System Op-code Statistics
=================================
Op-code      Rx Count   Tx Count
---------------------------------
ccm              1018       1018
...
raps              183         48
...
---------------------------------
Total            1201       1066
=================================
*A:PE-1#
```

To see an example of the console messages on a ring failure, when the unblocked port (1/1/2) on PE-2 is shut down, the following messages are displayed.

```
*A:PE-2# configure port 1/1/2 shutdown

26 2016/05/09 12:38:16.65 UTC WARNING: SNMP #2004 Base 1/1/2
"Interface 1/1/2 is not operational"

27 2016/05/09 12:38:16.65 UTC MINOR: ERING #2001 Base eth-ring-1
"Eth-Ring 1 path b changed fwd state to blocked"

28 2016/05/09 12:38:16.65 UTC MINOR: ERING #2001 Base eth-ring-1
"Eth-Ring 1 path a changed fwd state to unblocked"
*A:PE-2#
29 2016/05/09 12:38:16.67 UTC MAJOR: SVCMGR #2210 Base
"Processing of an access port state change event is finished and the status of a

ll affected SAPs on port 1/1/2 has been updated."
30 2016/05/09 12:38:19.73 UTC MINOR: ETH_CFM #2001 Base
"MEP 1/12/122 highest defect is now defRemoteCCM"
```

For troubleshooting, the **tools dump eth-ring** <*ring-index*> command displays path information, the internal state of the control protocol, related statistics information and up to the last 16 protocol events (including messages sent and received, and the expiration of timers). An associated **clear** parameter exists, which clears the event information in this output when the command is entered. The following is an example of the output on PE-1.

```
*A:PE-1# tools dump eth-ring 1

ringId 1 (Up/Up): numPaths 2 nodeId 4a:c4:ff:00:00:00
 SubRing: none (interconnect ring 0, propagateTc  No), Cnt 1
  path-a, port 1/1/1 (Up), tag 1.1(Up) status (Up/Up/Fwd)
      cc (Dn/Up): Cnt 3/3 tm 000 04:26:03.850/000 04:31:25.920
      state: Cnt 6 B/F 000 04:26:03.850/000 04:31:28.910, flag: 0x0
  path-b, port 1/1/3 (Up), tag 1.1(Up) status (Up/Up/Fwd)
      cc (Dn/Up): Cnt 2/2 tm 497 02:26:01.820/000 00:25:29.040
      state: Cnt 4 B/F 497 02:26:01.820/000 00:25:31.620, flag: 0x0
  FsmState=  PEND, Rpl = None, revert = 60 s, guard = 5 ds
    Defects =
    Running Timers =
    lastTxPdu = 0x0000 Nr (stopped)
    path-a Normal, RxId(I)= 4a:c5:ff:00:00:00, rx= v1-0x0020 Nr, cmd= None
    path-b Normal, RxId(I)= 4a:c5:ff:00:00:00, rx= v1-0x0020 Nr, cmd= None
  DebugInfo:  aPathSts 5, bPathSts 3, pm (set/clr) 0/0, txFlush 0
    RxRaps: ok 12 nok 0 self 0, TmrExp - wtr 0(0), grd 3, wtb 0
    Flush: cnt 9 (4/5/0) tm 000 04:31:27.710-000 04:31:27.710 Out/Ack 0/1
    RxRawRaps: aPath 2940 bPath 94 vPath 0
    Now: 000 04:31:48.480 , softReset: No - noTx 0

  Seq Event  RxInfo(Path: NodeId-Bytes)
             state:TxInfo (Bytes)            Dir  pA  pB        Time
  === ===== ============================  ===== === === ================
  008   pdu A: 4a:c5:ff:00:00:00-0xb040 Sf(DNF)
              PROT   : 0xb060   Sf(DNF)      Rx<-- Fwd Blk 000 00:22:41.910
  009   pdu B: 4a:c6:ff:00:00:00-0xb020 Sf
              PROT   : 0xb060   Sf(DNF)      RxF<- Fwd Blk 000 00:25:28.930
  010   bUp
              PEND-G: 0x0020  Nr            Tx--> Fwd Blk 000 00:25:31.040
  011   pdu A: 4a:c5:ff:00:00:00-0x0000 Nr
              PEND   : 0x0020  Nr            Rx<-- Fwd Blk 000 00:25:31.620
  012   pdu
              PEND   :                       ----- Fwd Fwd 000 00:25:31.620
  013   pdu B: 4a:c6:ff:00:00:00-0xb060 Sf(DNF)
              PEND   :                       Rx<-- Fwd Fwd 000 00:25:31.930
  014   pdu
              PROT   :                       ----- Fwd Fwd 000 00:25:31.930
  015   pdu B: 4a:c6:ff:00:00:00-0x0020 Nr
              PROT   :                       Rx<-- Fwd Fwd 000 00:25:32.430
  016   pdu
              PEND   :                       ----- Fwd Fwd 000 00:25:32.430
  017   pdu A: 4a:c6:ff:00:00:00-0x0020 Nr
              PEND   :                       Rx<-- Fwd Fwd 000 00:25:32.430
  018   pdu A: 4a:c5:ff:00:00:00-0x0080 Nr(RB )
              PEND   :                       RxF<- Fwd Fwd 000 00:26:32.720
  019   pdu
              IDLE   :                       ----- Fwd Fwd 000 00:26:32.720
  000   pdu B: 4a:c5:ff:00:00:00-0xb020 Sf
```

```
            IDLE  :                       RxF<- Fwd Fwd 000 04:26:01.010
  001   pdu
            PROT  :                       ----- Fwd Fwd 000 04:26:01.010
  002   aDn
            PROT  : 0xb000  Sf            TxF-> Blk Fwd 000 04:26:03.850
  003   pdu A: 4a:c5:ff:00:00:00-0xb020 Sf
            PROT  : 0xb000  Sf            RxF<- Blk Fwd 000 04:31:27.710
  004   aUp
            PEND-G: 0x0000  Nr            Tx--> Blk Fwd 000 04:31:27.840
  005   pdu A: 4a:c5:ff:00:00:00-0x0020 Nr
            PEND  : 0x0000  Nr            Rx<-- Blk Fwd 000 04:31:28.910
  006   pdu
            PEND  :                       ----- Fwd Fwd 000 04:31:28.910
  007   pdu B: 4a:c5:ff:00:00:00-0x0020 Nr
            PEND  :                       Rx<-- Fwd Fwd 000 04:31:28.910

*A:PE-1#
```

Configuration of a Sub-Ring to a Major Ring with a Non-Virtual Link

The differences from the preceding virtual link configuration with a non-virtual link for the sub-ring are:

- The sub-ring configuration on the interconnection nodes, PE-1 and PE-3, is modified to indicate that the sub-ring is not using a virtual link, otherwise it remains the same.
- The sub-ring configuration on the sub-ring node, PE-4, is also modified to indicate that this is part of a sub-ring that is not using a virtual link. This is mandatory on all non-interconnection nodes on the sub-ring in order to ensure the propagation of R-APS messages around the sub-ring.
- The virtual link services and SAPs must be removed from PE-1, PE-2 and PE3, that is:
  - On PE-1 and PE-3, the SAPs in VPLS 2 around the major ring (configured with the parameter *eth-ring 1*) are removed.
  - The service VPLS 2 is removed completely from PE-2.

The new configuration of sub-ring 2 on PE-1 is sas follows, the configuration on PE-3 is similar.

```
*A:PE-1# configure eth-ring 2
*A:PE-1>config>eth-ring# info
----------------------------------------------
    description "Ethernet Sub-ring 2 on Major Ring 1"
        revert-time 60
        sub-ring non-virtual-link
            interconnect ring-id 1
```

```
                propagate-topology-change
            exit
        exit
        path a 1/1/2 raps-tag 2.1
            description "Ethernet Ring 2 - PathA"
            eth-cfm
                mep 141 domain 1 association 14
                    ccm-enable
                    control-mep
                    no shutdown
                exit
            exit
            no shutdown
        exit
        no shutdown
```

The configuration of sub-ring 2 on PE-4 is as follows; note the configuration of the sub-ring non-virtual-link.

```
*A:PE-4# configure eth-ring 2
*A:PE-4>config>eth-ring# info
----------------------------------------------
    description "Ethernet Sub-ring 2"
        revert-time 60
        rpl-node owner
        sub-ring non-virtual-link
        exit
        path a 1/1/1 raps-tag 2.1
            description "Ethernet Ring 2 - PathA"
            eth-cfm
                mep 144 domain 1 association 14
                    ccm-enable
                    control-mep
                    no shutdown
                exit
            exit
            no shutdown
        exit
        path b 1/1/2 raps-tag 2.1
            description "Ethernet Ring 2 - PathB"
            rpl-end
            eth-cfm
                mep 344 domain 1 association 34
                    ccm-enable
                    control-mep
                    no shutdown
                exit
            exit
            no shutdown
        exit
        no shutdown
```

The SAP usage on PE-1 is as follows with only the control and data SAPs to PE-4 now using sub-ring 2.

```
*A:PE-1# show service sap-using eth-ring
===============================================================================
Service Access Points (Ethernet Ring)
===============================================================================
SapId              SvcId             Eth-Ring Path Admin Oper  Blocked Control/
                                                   State State         Data
-------------------------------------------------------------------------------
1/1/1:1.1          1                 1        a    Up    Up    No      Ctrl
1/1/3:1.1          1                 1        b    Up    Up    No      Ctrl
1/1/2:2.1          2                 2        a    Up    Up    No      Ctrl
1/1/1:1.11         11                1        a    Up    Up    No      Data
1/1/2:1.11         11                2        a    Up    Up    No      Data
1/1/3:1.11         11                1        b    Up    Up    No      Data
-------------------------------------------------------------------------------
Number of SAPs : 6
===============================================================================
*A:PE-1#
```

The information relating to sub-ring 2 is as follows and it can be seen that this is now not using a virtual link, but sub-ring 2 is still connected to major ring 1 and propagation is still enabled from the sub-ring to the major ring. The single ring path (a) is unblocked as the RPL is configured between PE-3 and PE-4.

```
*A:PE-1# show eth-ring 2

===============================================================================
Ethernet Ring 2 Information
===============================================================================
Description        : Ethernet Sub-ring 2 on Major Ring 1
Admin State        : Up                 Oper State       : Up
Node ID            : 4a:c4:ff:00:00:00
Guard Time         :    5 deciseconds   RPL Node         : rplNone
Max Revert Time    :   60 seconds       Time to Revert   : N/A
CCM Hold Down Time :    0 centiseconds  CCM Hold Up Time :   20 deciseconds
Compatible Version : 2
APS Tx PDU         : N/A
Defect Status      :

Sub-Ring Type      : nonVirtualLink     Interconnect-ID  : 1
Topology Change    : Propagate

-------------------------------------------------------------------------------
Ethernet Ring Path Summary
-------------------------------------------------------------------------------
Path Port     Raps-Tag     Admin/Oper      Type            Fwd State
-------------------------------------------------------------------------------
  a  1/1/2    2.1             Up/Up        normal          unblocked
  b  -        -                 -/-         -               -
===============================================================================
*A:PE-1#
```

Configuration of a Sub-Ring to a VPLS Service (with a Non-Virtual Link)

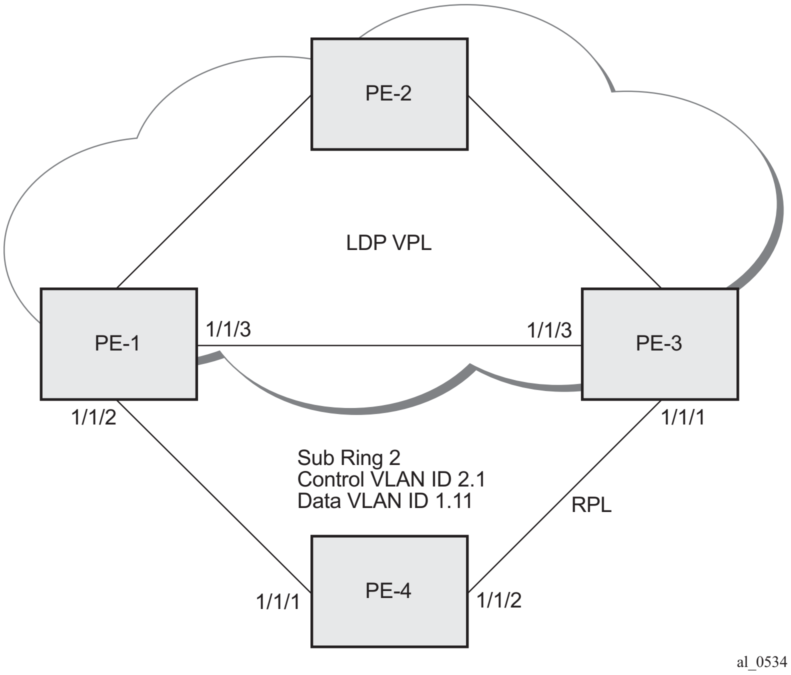
Sub-rings can be connected to VPLS services, in which case a virtual link is not used and is not configurable. While similar to the ring interconnect, there are a few differences.

Flush propagation is from the sub-ring to the VPLS, in the same way as it was for the sub-ring to the major ring. The same configuration parameter is used to propagate topology changes, note that in this case LDP flush messages (flush-all-from-me) are sent into the LDP portion of the network to account for ring changes without the need to configure anything in the VPLS service.

As with other rings, until an Ethernet ring instance is attached to the VPLS service, the ring operational status is down and the forwarding status of each port is blocked. This prevents operator from creating a loop by mis-configuration.

The topology for this case is shown in Figure 6. The configuration is very similar to the sub-ring with a non-virtual link described earlier, but ring 1 is replaced by a VPLS service using LDP signaled mesh SDPs between PE-1, PE-2 and PE-3 to create a fully meshed VPLS service. Both spoke and mesh SDPs using LDP could be used for the VPLS, however, only mesh SDPs have been used in this example.

Figure 6 **Sub-Ring to VPLS Topology**



The differences for the VPLS service connection to the configuration when the sub-ring is connected to a major ring without a virtual link are:

- The sub-ring configuration on the interconnection nodes, PE-1 and PE-3, is modified to indicate that the sub-ring is connected to a VPLS service.
- The sub-ring configuration on the sub-ring node, PE-4, is also modified to indicate that this is part of a sub-ring that is not using a virtual link. This is mandatory on all non-interconnection nodes on the sub-ring in order to ensure the propagation of R-APS messages around the sub-ring.
- The service (VPLS 1) and SAPs relating to the major ring 1 on PE-1, PE-2 and PE-3 are removed. These are replaced by routed IP interfaces configured with a routing protocol and LDP in order to signal the required MPLS labels, together with the necessary SDPs to provide interconnection at a service level.
- The data service (VPLS 11) is configured with mesh SDPs between PE-1, PE-2 and PE-3.

The configuration on PE-1 of the sub-ring 2 is as follows with the interconnect indicating a VPLS service. The configuration on PE-3 is similar.

```
*A:PE-1>config# eth-ring 2
*A:PE-1>config>eth-ring# info
----------------------------------------------
        description "Ethernet Sub-ring 2 on Major Ring 1"
        revert-time 60
        sub-ring non-virtual-link
            interconnect vpls
                propagate-topology-change
            exit
        exit
        path a 1/1/2 raps-tag 2.1
            description "Ethernet Ring 2 - PathA"
            eth-cfm
                mep 141 domain 1 association 14
                    ccm-enable
                    control-mep
                    no shutdown
                exit
            exit
            no shutdown
        exit
        no shutdown
```

The configuration of sub-ring 2 on PE-4 is as follows; note the configuration of the sub-ring non-virtual-link.

```
*A:PE-4# configure eth-ring 2
*A:PE-4>config>eth-ring# info
----------------------------------------------
        description "Ethernet Sub-ring 2"
        revert-time 60
        rpl-node owner
        sub-ring non-virtual-link
        exit
        path a 1/1/1 raps-tag 2.1
            description "Ethernet Ring 2 - PathA"
            eth-cfm
                mep 144 domain 1 association 14
                    ccm-enable
                    control-mep
                    no shutdown
                exit
            exit
            no shutdown
        exit
        path b 1/1/2 raps-tag 2.1
            description "Ethernet Ring 2 - PathB"
            rpl-end
            eth-cfm
                mep 344 domain 1 association 34
                    ccm-enable
                    control-mep
                    no shutdown
                exit
            exit
            no shutdown
        exit
        no shutdown
```

The data service on PE-1 is as follows. The configuration on PE-3 is similar.

```
*A:PE-1# configure
    service
        vpls 11 customer 1 create
            description "Data VPLS"
            sap 1/1/2:1.11 eth-ring 2 create
            exit
            sap 1/2/1:11 create
                description "Sample Customer Service SAP"
            exit
            mesh-sdp 12:11 create
            exit
            mesh-sdp 13:11 create
            exit
            no shutdown
```

The state of the sub-ring is as follows and shows the sub-ring is not using a virtual link, is connected to a VPLS service and has propagation of topology change events enabled. As earlier, the single ring path (a) is unblocked as the RPL is configured between PE-3 and PE-4.

```
*A:PE-1# show eth-ring 2

===============================================================================
Ethernet Ring 2 Information
```

```
===============================================================================
Description        : Ethernet Sub-ring 2 on Major Ring 1
Admin State        : Up                  Oper State       : Up
Node ID            : 4a:c4:ff:00:00:00
Guard Time         :    5 deciseconds   RPL Node         : rplNone
Max Revert Time    :   60 seconds       Time to Revert   : N/A
CCM Hold Down Time :    0 centiseconds  CCM Hold Up Time :   20 deciseconds
Compatible Version : 2
APS Tx PDU         : N/A
Defect Status      :

Sub-Ring Type      : nonVirtualLink     Interconnect-ID  : VPLS
Topology Change    : Propagate

-------------------------------------------------------------------------------
Ethernet Ring Path Summary
-------------------------------------------------------------------------------
Path Port     Raps-Tag     Admin/Oper      Type            Fwd State
-------------------------------------------------------------------------------
  a  1/1/2    2.1             Up/Up        normal          unblocked
  b  -        -                 -/-         -               -
===============================================================================
*A:PE-1#
```

In this case, if a topology change event occurs in the sub-ring, an LDP flush all-from-me message is sent by PE-1 and PE-3 to their LDP peers. This can be seen by enabling the following debugging for PE-1, where packets 1 and 2 are the flush messages.

```
*A:PE-1# debug router ldp peer 192.0.2.2 packet init
*A:PE-1# debug router ldp peer 192.0.2.3 packet init
*A:PE-1#
*A:PE-1# show debug
debug
    router "Base"
        ldp
            peer 192.0.2.2
                event
                exit
                packet
                    init
                exit
            exit
            peer 192.0.2.3
                event
                exit
                packet
                    init
                exit
            exit
        exit
    exit
exit
*A:PE-1#
```

```
*A:PE-1# configure port 1/1/2 shutdown

100 2016/05/10 07:16:59.16 UTC WARNING: SNMP #2004 Base 1/1/2
"Interface 1/1/2 is not operational"

101 2016/05/10 07:16:59.16 UTC MINOR: ERING #2001 Base eth-ring-2
"Eth-Ring 2 path a changed fwd state to blocked"
*A:PE-1#
1 2016/05/10 07:16:59.16 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Send Address Withdraw packet (msgId 95) to 192.0.2.2:0
 MAC Flush (All MACs learned from me)
Service FEC PWE3: ENET(5)/11 Group ID = 0 cBit = 0
"

2 2016/05/10 07:16:59.16 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Send Address Withdraw packet (msgId 82) to 192.0.2.3:0
 MAC Flush (All MACs learned from me)
Service FEC PWE3: ENET(5)/11 Group ID = 0 cBit = 0
"

102 2016/05/10 07:16:59.18 UTC MAJOR: SVCMGR #2210 Base
"Processing of an access port state change event is finished and the status of a
ll affected SAPs on port 1/1/2 has been updated."

3 2016/05/10 07:17:02.19 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Recv Address Withdraw packet (msgId 79) from 192.0.2.3:0
 "

4 2016/05/10 07:17:02.30 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Recv Address Withdraw packet (msgId 80) from 192.0.2.3:0
 "

103 2016/05/10 07:17:02.58 UTC MINOR: ETH_CFM #2001 Base
"MEP 1/14/141 highest defect is now defRemoteCCM"
```

Operational Procedures

Operators may wish to configure rings with or without control over reversion. Reversion can be controlled by timers or the ring can be run without reversion allowing the operator to choose when the ring reverts. To change a ring topology, the **manual** or **force** switch command may be used to block a specified ring path. A ring will still address failures when run without reversion but will not automatically revert to the RPL when resources are restored. A **clear** command can be used to clear the manual or force state of a ring.

The following **tools** commands are available to control the state of paths on a ring.

```
tools perform eth-ring clear <ring-index>
tools perform eth-ring force <ring-index> path {a|b}
tools perform eth-ring manual <ring-index> path {a|b}
```

In the following output , path “b” of eth-ring 1 is manually blocked and then cleared. Initially, both ports are unblocked.

```
*A:PE-1# show eth-ring 1

===============================================================================
Ethernet Ring 1 Information
===============================================================================
Description        : Ethernet Ring 1
Admin State        : Up                 Oper State       : Up
Node ID            : 4a:c4:ff:00:00:00
Guard Time         :    5 deciseconds   RPL Node         : rplNone
Max Revert Time    :   60 seconds       Time to Revert   : N/A
CCM Hold Down Time :    0 centiseconds  CCM Hold Up Time :   20 deciseconds
Compatible Version : 2
APS Tx PDU         : N/A
Defect Status      :

Sub-Ring Type      : none

-------------------------------------------------------------------------------
Ethernet Ring Path Summary
-------------------------------------------------------------------------------
Path Port     Raps-Tag     Admin/Oper      Type            Fwd State
-------------------------------------------------------------------------------
  a  1/1/1    1.1             Up/Up        normal          unblocked
  b  1/1/3    1.1             Up/Up        normal          unblocked
===============================================================================
*A:PE-1#
*A:PE-1#
*A:PE-1# tools perform eth-ring manual 1 path b
*A:PE-1# show eth-ring 1

===============================================================================
Ethernet Ring 1 Information
===============================================================================
Description        : Ethernet Ring 1
Admin State        : Up                 Oper State       : Up
Node ID            : 4a:c4:ff:00:00:00
Guard Time         :    5 deciseconds   RPL Node         : rplNone
Max Revert Time    :   60 seconds       Time to Revert   : N/A
CCM Hold Down Time :    0 centiseconds  CCM Hold Up Time :   20 deciseconds
Compatible Version : 2
APS Tx PDU         : Request State: 0x7
                     Sub-Code     : 0x0
                     Status       : 0x20  ( BPR )
                     Node ID      : 4a:c4:ff:00:00:00
Defect Status      :

Sub-Ring Type      : none

-------------------------------------------------------------------------------
Ethernet Ring Path Summary
-------------------------------------------------------------------------------
```

```
Path Port     Raps-Tag     Admin/Oper      Type            Fwd State
-------------------------------------------------------------------------------
  a  1/1/1    1.1             Up/Up        normal          unblocked
  b  1/1/3    1.1             Up/Up        normal          blocked
===============================================================================
*A:PE-1#
*A:PE-1#
*A:PE-1# tools perform eth-ring clear 1
*A:PE-1# show eth-ring 1

===============================================================================
Ethernet Ring 1 Information
===============================================================================
Description        : Ethernet Ring 1
Admin State        : Up                 Oper State       : Up
Node ID            : 4a:c4:ff:00:00:00
Guard Time         :    5 deciseconds   RPL Node         : rplNone
Max Revert Time    :   60 seconds       Time to Revert   : N/A
CCM Hold Down Time :    0 centiseconds  CCM Hold Up Time :   20 deciseconds
Compatible Version : 2
APS Tx PDU         : N/A
Defect Status      :

Sub-Ring Type      : none

-------------------------------------------------------------------------------
Ethernet Ring Path Summary
-------------------------------------------------------------------------------
Path Port     Raps-Tag     Admin/Oper      Type            Fwd State
-------------------------------------------------------------------------------
  a  1/1/1    1.1             Up/Up        normal          unblocked
  b  1/1/3    1.1             Up/Up        normal          unblocked
===============================================================================
*A:PE-1#
```

Both the **manual** and **force** command block the path specified, however, the **manual** command fails if there is an existing forced switch or signal fail event in the ring, as seen in the following output. The **force** command will block the port regardless of any existing ring state and there can be multiple force states simultaneously on a ring on different nodes.

```
*A:PE-1# tools perform eth-ring manual 1 path b
INFO: ERMGR #1001 Not permitted -
 The switch command is not compatible to the current state (FS), effective priority
(FS) or rpl-node type (None)
*A:PE-1#
```

Conclusion

Ethernet Ring APS provides an optimal solution for designing native Ethernet services with ring topology. With sub-rings, both multiple rings and access rings increase the versatility of G.8032. G.8032 has been expanded to more of the SR platforms by allowing R-APS with slower MEPs (including CCMs intervals of 1 second). This protocol provides simple configuration, operation and guaranteed fast protection time. The implementation also has a flexible encapsulation that allows dot1Q, QinQ or PBB for the ring traffic. It could be utilized on various services such as mobile backhaul, business VPN access, aggregation and core.

G.8032 Ethernet Ring Protection Single Ring Topology

This chapter provides information about G.8032 Ethernet ring protection single ring topology.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

The chapter was initially written for SROS release 8.0.R7, There have been several updates and the CLI in this edition corresponds to release 14.0.R2. This chapter describes ring protection for a single ring topology. Protection for multiple ring topologies is covered in G.8032 Ethernet Ring Protection Multiple Ring Topology.

Overview

G.8032 Ethernet ring protection is supported for data service SAPs within a regular VPLS service, a provider backbone bridging (PBB) VPLS (I/B-component) or a routed VPLS (R-VPLS). G.8032 is one of the fastest protection schemes for Ethernet networks.

ITU-T G.8032v2 specifies protection switching mechanisms and a protocol for Ethernet layer network (ETH) Ethernet rings. Ethernet rings can provide wide-area multi-point connectivity more economically due to their reduced number of links. The mechanisms and protocol defined in ITU-T G.8032v2 achieve highly reliable and stable protection and never form loops, which would negatively affect network operation and service availability. Each ring node is connected to adjacent nodes participating in the same ring using two independent paths, which use ring links (configured on ports or link aggregation groups (LAGs)). A ring link is bounded by two adjacent nodes and a port for a ring link is called a ring port. The minimum number of nodes on a ring is two.

The fundamentals of this ring protection switching architecture are:

- the principle of loop avoidance and
- the utilization of learning, forwarding, and address table mechanisms defined in the ITU-T G.8032v2 Ethernet flow forwarding function (ETH_FF) (Control plane).

Loop avoidance in the ring is achieved by guaranteeing that, at any time, traffic may flow on all but one of the ring links. This particular link is called the ring protection link (RPL) and under normal conditions this link is blocked, so it is not used for traffic. One designated node, the RPL owner, is responsible to block traffic over the one designated RPL. Under a ring failure condition, the RPL owner is responsible for unblocking the RPL, allowing the RPL to be used for traffic. The protocol ensures that even without an RPL owner defined, one link will be blocked and it operates as a “break before make protocol”, specifically the protocol guarantees that no link is restored until a different link in the ring is blocked. The other side of the RPL is configured as an RPL neighbor. An RPL neighbor blocks traffic on the link.

The event of a ring link or ring node failure results in protection switching of the traffic. This is achieved under the control of the ETH_FF functions on all ring nodes. A ring automatic protection switching (R-APS) protocol is used to coordinate the protection actions over the ring. The protection switching mechanisms and protocol supports a multi-ring/ladder network that consists of connected Ethernet rings, however, that is not covered in this chapter.

Ring Protection Mechanism

The Ring Protection protocol is based on the following building blocks:

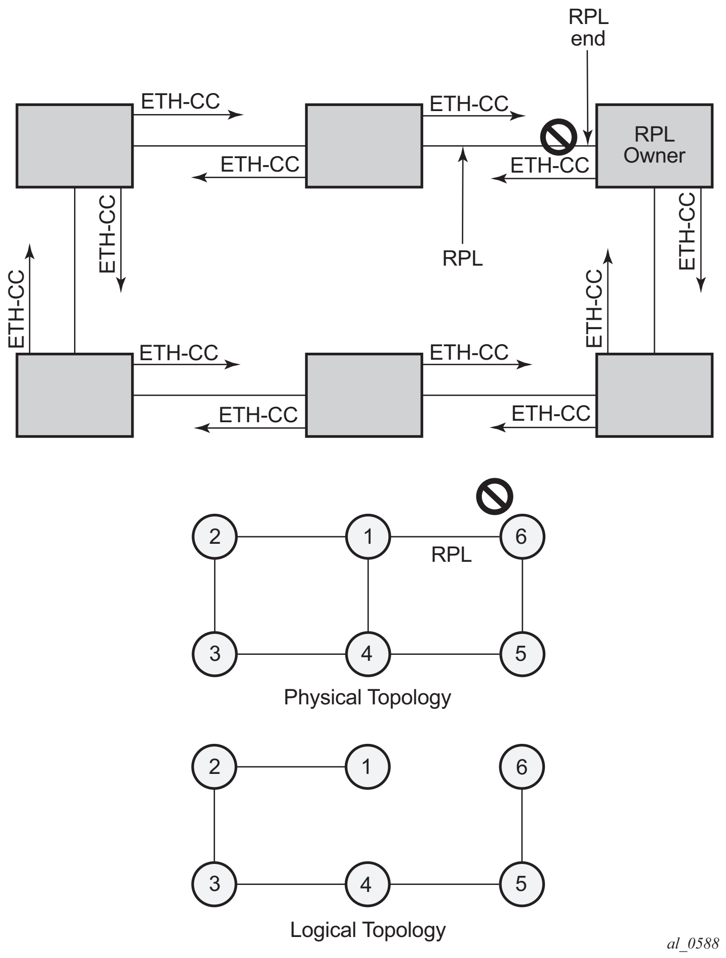
- Ring status change on failure
  - Idle -> Link failure -> Protection -> Recovery -> Idle
- Ring Control State changes
  - Idle -> Protection -> Manual Switch -> Forced Switch -> Pending
- Re-use existing ETH OAM
  - Monitoring: ETH Continuity Check messages
  - Failure Notification: Y.1731 Signal Failure
- Forwarding Database MAC Flush on ring status change
- RPL (Ring Protection Link)
  - Defines blocked link in idle status

Figure 7 shows a ring of six nodes, with the RPL owner on the top right. One link of the RPL owner is designated to be the RPL and will be blocked in order to prevent a loop. Schematics of the physical and logical topologies are also shown.

When an RPL owner and RPL end are configured, the associated link will be the RPL when the ring is fully operational and so be blocked by the RPL owner. If a different ring link fails then the RPL will be unblocked by the RPL owner. When the failed link recovers, it will initially be blocked by one of its adjacent nodes. The adjacent node sends an R-APS message across the ring to indicate the error is cleared and after a configurable time, if reversion is enabled, the RPL will revert to being blocked with all other links unblocked. This ensures that the ring topology is predictable when fully operational.

If a specific RPL owner is not configured, then the last link to become active will be blocked and the ring will remain in this state until another link fails. However, this operation makes the selection of the blocked link non-deterministic.

Figure 7 **G.8032 Operation and Topologies**

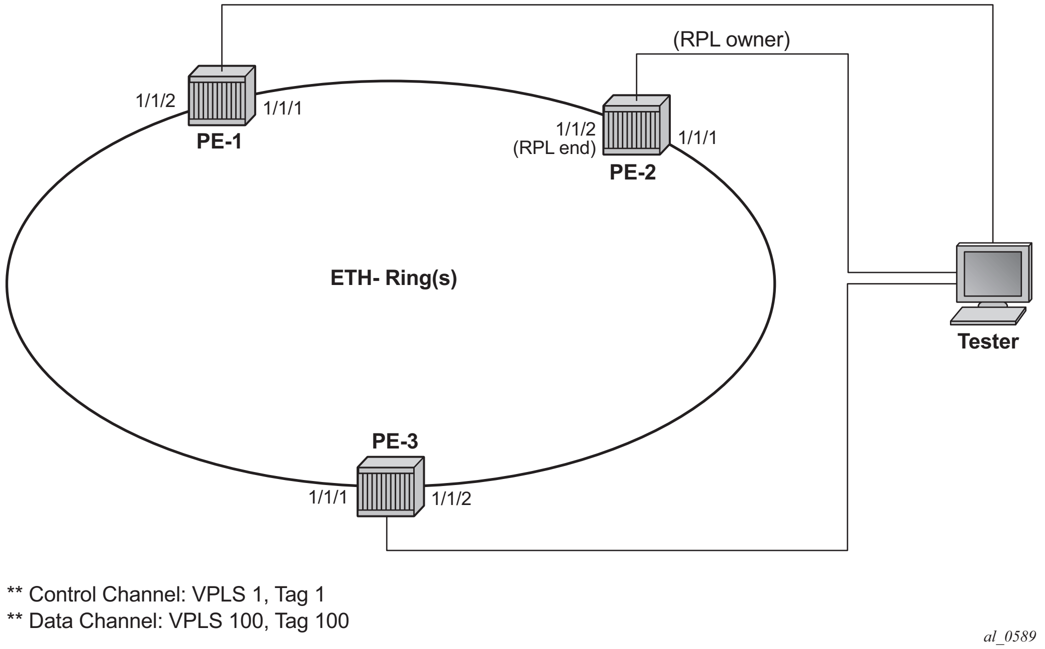


The protection protocol uses a specific control VLAN, with the associated data VLANs taking their forwarding state from the control VLAN.

Configuration

The example topology is shown in Figure 8.

Figure 8 **Example Topology**



** Control Channel: VPLS 1, Tag 1
** Data Channel: VPLS 100, Tag 100

The **eth-ring** configuration commands are as follows:

```
configure
    eth-ring <ring-index>
        ccm-hold-time { [down <down-timeout>] [up <up-timeout>] }
        compatible-version <version>
        description <description-string>
        guard-time <time>
        node-id <xx:xx:xx:xx:xx:xx or xx-xx-xx-xx-xx-xx>
        path {a|b} [ { <port-id>|<lag-id> } raps-tag <qtag>[.<qtag>] ]
            description <description-string>
            eth-cfm
                mep <mep-id> domain <md-index> association <ma-index>
                ...
            rpl-end
            shutdown
        revert-time <time>
        rpl-node {owner|nbr}
        shutdown
```

```
sub-ring {virtual-link|non-virtual-link}
```

Parameters:

- *ring-index* — This is the number by which the ring is referenced, values: 1 to128.
- **ccm-hold-time** {[down <*down-timeout*>] [up <*up-timeout*>]}
  - down — This command specifies the timer that controls the delay between detecting that ring path is down and reporting it to the G.8032 protection module. If a non-zero value is configured, the system will wait for the time specified in the value parameter before reporting it to the G.8032 protection module. This parameter applies only to the ring path continuity check message (CCM). It does *not* apply to the ring port link state. To dampen ring port link state transitions, use the hold-time parameter from the physical member port. This is useful if the underlying path between two nodes is going across an optical system which implements its own protection.
  - up — This command specifies the timer which controls the delay between detecting that the ring path is up and reporting it to the G.8032 protection module. If a non-zero value is configured, the system will wait for the time specified in the value parameter before reporting it to the G.8032 protection module. This parameter applies only to ring path CCM. It does *not* apply to the member port link state. To dampen member port link state transitions, use the hold-time parameter from the physical member port.
  - Values: <down-timeout> : [0..5000] in deciseconds - Default: 0
    <up-timeout> : [0..5000] in deciseconds - Default: 20
    1 centisecond = 10ms
    1 decisecond = 100ms
- compatible version — This command configures eth-ring compatibility version for the G.8032 state machine and messages. The default is version 2 (ITU G.8032v2) and all 7x50 systems use version 2. If there is a need to interwork with third party devices that only support version 1, this can be set to version 1 allowing the reception of version 1 PDUs. Version 2 is encoded as 1 in the R-APS messages. Compatibility allows the reception of version 1 (encoded as 0) R-APS PDUs but, as per the G.8032 specification, higher versions are ignored on reception. For the SR/ESS, messages are always originated with version 2. Therefore if a third party switch supported version 3 (encoded as 2) or higher, interworking is also supported provided the other switch is compatible with version 2.
- description <description-string> — This configures a text string, up to 80 characters, which can be used to describe the use of the eth-ring.

- guard-time <time> — The forwarding method, in which R-APS messages are copied and forwarded at every Ethernet ring node, can result in a message corresponding to an old request, that is no longer relevant, being received by Ethernet ring nodes. Reception of an old R-APS message may result in erroneous ring state interpretation by some Ethernet ring nodes. The guard timer is used to prevent Ethernet ring nodes from acting upon outdated R-APS messages and prevents the possibility of forming a closed loop. Messages are not forwarded when the guard-timer is running.

  Values: [1..20] in deciseconds - Default: 5
  1 decisecond = 100ms
- node-id <xx:xx:xx:xx:xx:xx or xx-xx-xx-xx-xx-xx> — This allows the node identifier to be explicitly configured. By default the chassis MAC is used. It is not required in typical configurations.
- path {a|b} [{<port-id>|<lag-id>} raps-tag <qtag>[.<qtag>]] — This parameter defines the paths around the ring, of which there are two in different directions on the ring: an “a” path and a “b” path. In addition, it configures the encapsulation used for the R-APS messages on the ring. These can be either single or double tagged.
  - description <description-string> — This configures a text string, up to 80 characters, which can be used to describe the use of the path.
  - eth-cfm — Configures the associated Ethernet connectivity fault management (CFM) parameters.
    - mep <mep-id> domain <md-index> association <ma-index> — The maintenance endpoint (MEP) defined under the path is used for the G.8032 protocol messages, which are based on IEEE 802.1ag/Y.1731 CFM frames.
  - rpl-end — When configured, this path is expected to be one end of the RPL. This parameter must be configured in conjunction with the **rpl-node**.
  - shutdown — This command shuts down the path.
- revert-time <time> — This command configures the revert time for an Eth-Ring. Revert time is the time that the RPL will wait before returning to the blocked state. Configuring “no revert-time” disables reversion, effectively setting the revert-time to zero.

  Values: [60..720] in seconds - Default: 300
- rpl-node {owner|nbr} — A node can be designated as either the owner of the RPL, in which case this node is responsible for the RPL, or the nbr, in which case this node is expected to be the neighbor to the RPL owner across the RPL. The nbr is optional and is included to be compliant with the specification. This parameter must be configured in conjunction with the **rpl-end** parameter.
- shutdown — This command shuts down the ring.

- sub-ring {virtual-link|non-virtual-link} — This is beyond the scope of this chapter, as it is only required for multiple ring topologies.

Prerequisites

Logging

Create following log-id on PE-2 to see major events logged to the console on PE-2. This is an optional step; alternatively, log 99 can be consulted.

```
configure
    log
        log-id 1
            from main
            to console
        exit
```

Configure Encapsulation for Ring Ports

To configure R-APS, there should be at least two VPLS services for one Eth-Ring instance, one for the control channel and the other (or more) for data channel(s). The control channel is used for R-APS signaling while data channel is for user data traffic. The state of the data channels is inherited from the state of the control channel.

- Eth-Ring needs R-APS tags to send and receive G.8032 signaling messages. To configure a control channel, an access SAP configuration is required on each path a/b port. The SAP configuration follows that of the port and must be either *dot1q* or *qinq*, consequently the control and data packets are either single tagged or double tagged. It is also possible to have the control VPLS using single tagged frames with the data VPLSs using double tagged framed; this requires the system to be configured with the **new-qinq-untagged-sap** parameter (**configure system ethernet new-qinq-untagged-sap**), with the ring path raps-tags and control VPLS SAPs configured as qtag.0, and the data VPLSs configured as QinQ SAPs.

  In this example, single tags are used so the commands for ring node PE-1 are:

```
*A:PE-1# configure port 1/1/[1..2] ethernet mode access
*A:PE-1# configure port 1/1/[1..2] ethernet encap-type dot1q
*A:PE-1# configure port 1/1/[1..2] no shutdown
```

Configure ETH-CFM

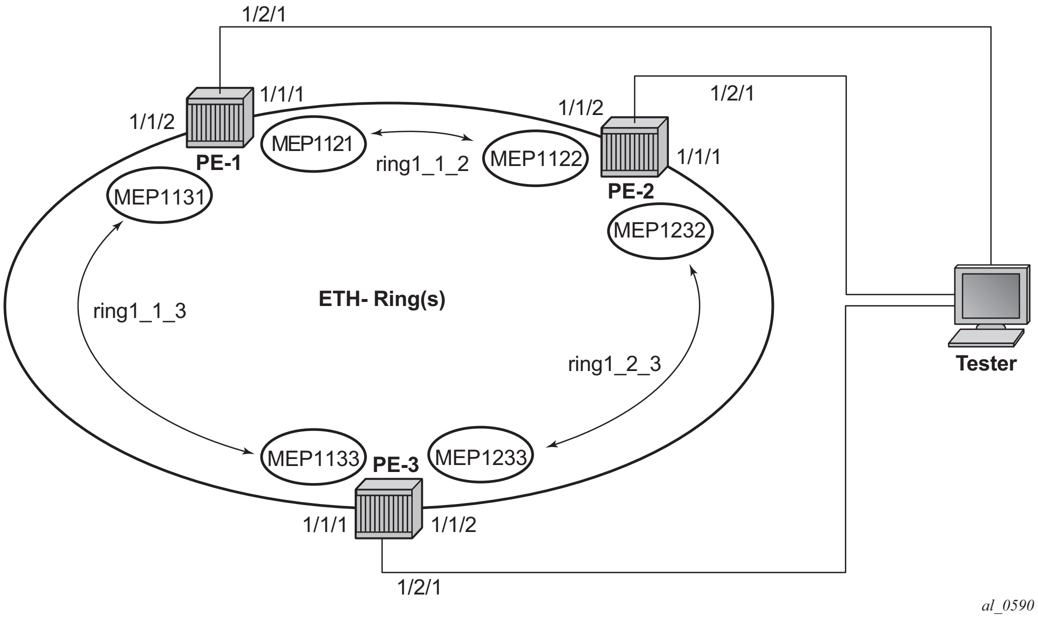
Ethernet Ring requires Eth-CFM domains, associations and MEPs being configured. The domain format should be none and association name should be ITU-T carrier code- based (ICC-based - Y.1731). The minimum CCM interval for the 7x50 is 10ms. The eth-ring MEP requires a CCM interval, such as 10ms, 100ms, or 1s, to be configured.

The MEPs used for R-APS control normally will have CCM configured on the control channel path MEPs for failure detection. Alternatively, detecting a failure of the ring may be achieved by running Ethernet in the First Mile (EFM) at the port level if CCM is not possible at 10ms, 100ms, or 1s. Loss-of-signal, in conjunction with other OAM, is applicable only when the nodes are directly connected.

To omit the failure detecting CCMs, it would be necessary to remove the *ccm-enable* from under the path MEPs and to remove the *remote-mepids* from under the *eth-cfm associations* on all nodes.

Figure 9 shows the Ethernet CFM configuration used here.

Figure 9 **Ethernet CFM Configuration**



The configuration of each node is as follows.

PE-1:

```
configure
```

```
    eth-cfm
        domain 1 format none level 3
            association 1 format icc-based name "ring1_1_2"
                ccm-interval 1
                remote-mepid 1122
            exit
            association 2 format icc-based name "ring1_1_3"
                ccm-interval 1
                remote-mepid 1133
            exit
        exit
    exit
```

PE-2:

```
configure
    eth-cfm
        domain 1 format none level 3
            association 1 format icc-based name "ring1_2_3"
                ccm-interval 1
                remote-mepid 1233
            exit
            association 2 format icc-based name "ring1_1_2"
                ccm-interval 1
                remote-mepid 1121
            exit
        exit
    exit
```

PE-3:

```
configure
    eth-cfm
        domain 1 format none level 3
            association 1 format icc-based name "ring1_1_3"
                ccm-interval 1
                remote-mepid 1131
            exit
            association 2 format icc-based name "ring1_2_3"
                ccm-interval 1
                remote-mepid 1232
            exit
        exit
    exit
```

Configure Eth-Ring

Two paths should be configured to form a ring. In this example, VLAN tag 1 is used as control channel for R-APS signaling in the ring.

PE-1:

```
configure
    eth-ring 1
        path a 1/1/1 raps-tag 1
            eth-cfm
                mep 1121 domain 1 association 1
                    ccm-enable
                    control-mep
                    no shutdown
                exit
            exit
            no shutdown
        exit
        path b 1/1/2 raps-tag 1
            eth-cfm
                mep 1131 domain 1 association 2
                    ccm-enable
                    control-mep
                    no shutdown
                exit
            exit
            no shutdown
        exit
        no shutdown
    exit
```

It is mandatory to configure a MEP in the path context, otherwise the following error will be displayed:

```
*A:PE-1>config>eth-ring>path# no shutdown
INFO: ERMGR #1001 Not permitted - must configure eth-cfm MEP first
*A:PE-1>config>eth-ring>path#
```

While MEPs are mandatory, enabling CCM on the MEP in the pathcontext as a failure detection mechanism is optional.

In order to define the RPL, node PE-2 has been configured as the RPL owner and path “b” as the RPL end. The link between nodes PE-1 and PE-2 will be the RPL with node PE-2 blocking that link when the ring is fully operational.

PE-2:

```
configure
    eth-ring 1
        revert-time 60
        rpl-node owner
        path a 1/1/1 raps-tag 1
            eth-cfm
                mep 1232 domain 1 association 1
                    ccm-enable
                    control-mep
                    no shutdown
                exit
            exit
            no shutdown
        exit
        path b 1/1/2 raps-tag 1
            rpl-end
            eth-cfm
                mep 1122 domain 1 association 2
                    ccm-enable
                    control-mep
                    no shutdown
                exit
            exit
            no shutdown
        exit
        no shutdown
    exit
```

It is not permitted to configure a path as an RPL end without having configured the node on this ring to be either the RPL *owner* or *nbr* otherwise the following error message is reported.

```
*A:PE-2>config>eth-ring# path b rpl-end
INFO: ERMGR #1001 Not permitted - path-type rpl-end is not consistent with eth-ring
'rpl-node' type
*A:PE-2>config>eth-ring#
```

PE-3:

```
configure
    eth-ring 1
        path a 1/1/1 raps-tag 1
            eth-cfm
                mep 1133 domain 1 association 1
                    ccm-enable
                    control-mep
                    no shutdown
                exit
            exit
            no shutdown
        exit
        path b 1/1/2 raps-tag 1
            eth-cfm
                mep 1233 domain 1 association 2
                    ccm-enable
                    control-mep
                    no shutdown
                exit
            exit
            no shutdown
        exit
        no shutdown
    exit
```

Until the Ethernet Ring instance is attached to the service (VPLS in this case), the ring operational status is down and the forwarding status of each port is blocked. This prevents operator from creating a loop by mis-configuration. This state can be seen on ring node PE-1 as follows:

```
*A:PE-1# show eth-ring 1
===============================================================================
Ethernet Ring 1 Information
===============================================================================
Description        : (Not Specified)
Admin State        : Up                  Oper State       : Down
Node ID            : 4a:c4:ff:00:00:00
Guard Time         :    5 deciseconds   RPL Node         : rplNone
Max Revert Time    :  300 seconds       Time to Revert   : N/A
CCM Hold Down Time :    0 centiseconds  CCM Hold Up Time :   20 deciseconds
Compatible Version : 2
APS Tx PDU         : Request State: 0xB
                     Sub-Code     : 0x0
                     Status       : 0x20  ( BPR )
                     Node ID      : 4a:c4:ff:00:00:00
Defect Status      :
Sub-Ring Type      : none
-------------------------------------------------------------------------------
Ethernet Ring Path Summary
-------------------------------------------------------------------------------
Path Port     Raps-Tag     Admin/Oper      Type            Fwd State
-------------------------------------------------------------------------------
  a  1/1/1    1               Up/Down      normal          blocked
  b  1/1/2    1               Up/Down      normal          blocked
===============================================================================
*A:PE-1#
```

Configure Control Channel VPLS Service

Paths a and b defined in the eth-ring must be added as SAPs into a VPLS service (standard VPLS in this example) using the *eth-ring* parameter. The SAP encapsulation values must match the values of the *raps-tag* configured for the associated path.

G.8032 uses the same raps-tag value on all nodes on the ring, as configured in this example. However, the SROS implementation relaxes this constraint by requiring the tag to match only on adjacent nodes.

PE-1:

```
configure
    service
        vpls 1 customer 1 create
            sap 1/1/1:1 eth-ring 1 create
            exit
            sap 1/1/2:1 eth-ring 1 create
            exit
            no shutdown
        exit
```

PE-2:

```
configure
    service
        vpls 1 customer 1 create
            sap 1/1/1:1 eth-ring 1 create
            exit
            sap 1/1/2:1 eth-ring 1 create
            exit
            no shutdown
        exit
```

PE-3:

```
configure
    service
        vpls 1 customer 1 create
            sap 1/1/1:1 eth-ring 1 create
            exit
            sap 1/1/2:1 eth-ring 1 create
            exit
            no shutdown
        exit
```

A normal SAP or SDP can be added in a control channel VPLS on condition the *eth-ring* parameter is present. Trying to add a SAP or SDP without this parameter into a control channel VPLS will result in the following message being displayed.

```
*A:PE-1# configure service vpls 1 sap 1/2/1:1 create
MINOR: SVCMGR #1321 Service contains an Ethernet ring control SAP
*A:PE-1#
```

Now the Eth-Ring is operationally up and the RPL is blocking successfully on ring node PE-2 port 1/1/2, as expected from the RPL owner/end configuration.

An overview of all of the rings can be shown using the following commands, in this case on node PE-2.

First, the ETH ring status is shown.

```
*A:PE-2# show eth-ring status
===============================================================================
Ethernet Ring (Status information)
===============================================================================
Ring   Admin  Oper       Path Information                 MEP Information
ID     State  State  Path         Tag       State     Ctrl-MEP CC-Intvl Defects
```

```
-------------------------------------------------------------------------------
1      Up     Up     a - 1/1/1        1     Up        Yes      1         -----
                     b - 1/1/2        1     Up        Yes      1         -----
===============================================================================
Ethernet Tunnel MEP Defect Legend:
R = Rdi, M = MacStatus, C = RemoteCCM, E = ErrorCCM, X = XconCCM
*A:PE-2#
```

The ring and path forwarding states is shown with following command.

```
*A:PE-2# show eth-ring
===============================================================================
Ethernet Rings (summary)
===============================================================================
Ring Int  Admin Oper            Paths Summary                       Path States
ID   ID   State State                                               a     b
-------------------------------------------------------------------------------
1    -    Up    Up    a - 1/1/1        1     b - 1/1/2        1     U     B
===============================================================================
Ethernet Ring Summary Legend:   B - Blocked    U - Unblocked
*A:PE-2#
```

Specific ring information can be shown on each node separately, as follows.

PE-1:

```
*A:PE-1# show eth-ring 1
===============================================================================
Ethernet Ring 1 Information
===============================================================================
Description        : (Not Specified)
Admin State        : Up                 Oper State       : Up
Node ID            : 4a:c4:ff:00:00:00
Guard Time         :    5 deciseconds   RPL Node         : rplNone
Max Revert Time    :  300 seconds       Time to Revert   : N/A
CCM Hold Down Time :    0 centiseconds  CCM Hold Up Time :   20 deciseconds
Compatible Version : 2
APS Tx PDU         : N/A
Defect Status      :
Sub-Ring Type      : none
-------------------------------------------------------------------------------
Ethernet Ring Path Summary
-------------------------------------------------------------------------------
Path Port     Raps-Tag     Admin/Oper      Type            Fwd State
-------------------------------------------------------------------------------
  a  1/1/1    1               Up/Up        normal          unblocked
  b  1/1/2    1               Up/Up        normal          unblocked
===============================================================================
*A:PE-1#
```

PE-2:

```
*A:PE-2# show eth-ring 1
```

```
===============================================================================
Ethernet Ring 1 Information
===============================================================================
Description        : (Not Specified)
Admin State        : Up                 Oper State       : Up
Node ID            : 4a:c5:ff:00:00:00
Guard Time         :    5 deciseconds   RPL Node         : rplOwner
Max Revert Time    :   60 seconds       Time to Revert   : N/A
CCM Hold Down Time :    0 centiseconds  CCM Hold Up Time :   20 deciseconds
Compatible Version : 2
APS Tx PDU         : Request State: 0x0
                     Sub-Code     : 0x0
                     Status       : 0xA0  ( RB BPR )
                     Node ID      : 4a:c5:ff:00:00:00
Defect Status      :
Sub-Ring Type      : none
-------------------------------------------------------------------------------
Ethernet Ring Path Summary
-------------------------------------------------------------------------------
Path Port     Raps-Tag     Admin/Oper      Type            Fwd State
-------------------------------------------------------------------------------
  a  1/1/1    1               Up/Up        normal          unblocked
  b  1/1/2    1               Up/Up        rplEnd          blocked
===============================================================================
*A:PE-2#
```

Node PE-2 is the RPL owner and port 1/1/2 is the RPL end. The *revert-time* shows the configured value.

When a revert is pending, the “Time to Revert” will show the number of seconds remaining before the revert occurs, as follows:

```
*A:PE-2# show eth-ring 1
===============================================================================
Ethernet Ring 1 Information
===============================================================================
Description        : (Not Specified)
Admin State        : Up                 Oper State       : Up
Node ID            : 4a:c5:ff:00:00:00
Guard Time         :    5 deciseconds   RPL Node         : rplOwner
Max Revert Time    :   60 seconds       Time to Revert   :   43 seconds
CCM Hold Down Time :    0 centiseconds  CCM Hold Up Time :   20 deciseconds
Compatible Version : 2
APS Tx PDU         : N/A
Defect Status      :
Sub-Ring Type      : none
-------------------------------------------------------------------------------
Ethernet Ring Path Summary
-------------------------------------------------------------------------------
Path Port     Raps-Tag     Admin/Oper      Type            Fwd State
-------------------------------------------------------------------------------
  a  1/1/1    1               Up/Up        normal          unblocked
  b  1/1/2    1               Up/Up        rplEnd          unblocked
===============================================================================
*A:PE-2#
```

On reversion, the following console message is logged.

```
68 2016/05/02 11:22:50.87 UTC MINOR: ERING #2001 Base eth-ring-1
"Eth-Ring 1 path b changed fwd state to blocked"
```

PE-3:

```
*A:PE-3# show eth-ring 1
===============================================================================
Ethernet Ring 1 Information
===============================================================================
Description        : (Not Specified)
Admin State        : Up                 Oper State       : Up
Node ID            : 4a:c6:ff:00:00:00
Guard Time         :    5 deciseconds   RPL Node         : rplNone
Max Revert Time    :  300 seconds       Time to Revert   : N/A
CCM Hold Down Time :    0 centiseconds  CCM Hold Up Time :   20 deciseconds
Compatible Version : 2
APS Tx PDU         : N/A
Defect Status      :
Sub-Ring Type      : none
-------------------------------------------------------------------------------
Ethernet Ring Path Summary
-------------------------------------------------------------------------------
Path Port     Raps-Tag     Admin/Oper      Type            Fwd State
-------------------------------------------------------------------------------
  a  1/1/1    1               Up/Up        normal          unblocked
  b  1/1/2    1               Up/Up        normal          unblocked
===============================================================================
*A:PE-3#
```

Finally, the details of an individual path can be shown.

```
*A:PE-2# show eth-ring 1 path a
===============================================================================
Ethernet Ring 1 Path Information
===============================================================================
Description        : (Not Specified)
Port               : 1/1/1              Raps-Tag         : 1
Admin State        : Up                 Oper State       : Up
Path Type          : normal             Fwd State        : unblocked
                                        Fwd State Change : 05/02/2016 11:17:59
Last Switch Command: noCmd
APS Rx PDU         : Request State: 0x0
                     Sub-Code     : 0x0
                     Status       : 0x20  ( BPR )
                     Node ID      : 4a:c6:ff:00:00:00
===============================================================================
*A:PE-2#
```

Configure User Data Channel VPLS Service

The user data channels are created on a separate VPLS, VPLS 100 in the example. Tag 100 and VPLS 100 are used here. The ring data channels must be on the same ports as the corresponding control channels configured above. The access into the data services can use SAPs and/or SDPs.

PE-1:

```
configure
    service
        vpls 100 customer 1 create
            sap 1/1/1:100 eth-ring 1 create
            exit
            sap 1/1/2:100 eth-ring 1 create
            exit
            sap 1/2/1:100 create
            exit
            no shutdown
        exit
    exit
```

PE-2:

```
configure
    service
        vpls 100 customer 1 create
            sap 1/1/1:100 eth-ring 1 create
            exit
            sap 1/1/2:100 eth-ring 1 create
            exit
            sap 1/2/1:100 create
            exit
            no shutdown
        exit
    exit
```

PE-3:

```
configure
    service
        vpls 100 customer 1 create
            sap 1/1/1:100 eth-ring 1 create
            exit
            sap 1/1/2:100 eth-ring 1 create
            exit
            sap 1/2/1:100 create
            exit
            no shutdown
        exit
    exit
```

All of the SAPs which are configured to use ETH rings can be shown, using PE-1 as an example.

```
*A:PE-1# show service sap-using eth-ring
===============================================================================
Service Access Points (Ethernet Ring)
===============================================================================
SapId              SvcId             Eth-Ring Path Admin Oper  Blocked Control/
                                                   State State         Data
-------------------------------------------------------------------------------
1/1/1:1            10                1        a    Up    Up    No      Ctrl
1/1/2:1            10                1        b    Up    Up    No      Ctrl
1/1/1:100          100               1        a    Up    Up    No      Data
1/1/2:100          100               1        b    Up    Up    No      Data
-------------------------------------------------------------------------------
Number of SAPs : 4
===============================================================================
*A:PE-1#
```

Debug

To see an example of the console messages on a ring failure, when the unblocked port (1/1/1) on node PE-2 is shut down, the following messages are displayed.

```
*A:PE-2# configure port 1/1/1 shutdown
15 2016/05/02 09:21:11.15 UTC WARNING: SNMP #2004 Base 1/1/1
"Interface 1/1/1 is not operational"
16 2016/05/02 09:21:11.15 UTC MINOR: ERING #2001 Base eth-ring-1
"Eth-Ring 1 path a changed fwd state to blocked"
17 2016/05/02 09:21:11.15 UTC MINOR: ERING #2001 Base eth-ring-1
"Eth-Ring 1 path b changed fwd state to unblocked"
*A:PE-2#
18 2016/05/02 09:21:11.16 UTC MAJOR: SVCMGR #2210 Base
"Processing of an access port state change event is finished and the status of a
ll affected SAPs on port 1/1/1 has been updated."
19 2016/05/02 09:21:14.85 UTC MINOR: ETH_CFM #2001 Base
"MEP 1/1/1232 highest defect is now defRemoteCCM"
```

For troubleshooting, the **tools dump eth-ring** <*ring-index*> command displays path information, the internal state of the control protocol, related statistics information and up to the last 20 protocol events (including messages sent and received, and the expiration of timers). An associated parameter *clear* exists, clearing the event information in this output when the command is entered. The following is an example of the output on node PE-2 with port 1/1/1 active.

```
*A:PE-2# tools dump eth-ring 1
ringId 1 (Up/Up): numPaths 2 nodeId 4a:c5:ff:00:00:00
 SubRing: none (interconnect ring 0, propagateTc  No), Cnt 0
  path-a, port 1/1/1 (Up), tag 1.0(Up) status (Up/Up/Fwd)
      cc (Dn/Up): Cnt 5/5 tm 000 00:48:02.730/000 00:48:04.430
      state: Cnt 8 B/F 000 00:47:58.630/000 00:48:11.380, flag: 0x0
  path-b, port 1/1/2 (Up), tag 1.0(Up) status (Up/Up/Blk)
      cc (Dn/Up): Cnt 3/3 tm 000 00:41:33.740/000 00:41:33.960
      state: Cnt 11 B/F 000 00:49:11.680/000 00:47:58.630, flag: 0x0
  FsmState=  IDLE, Rpl = Owner, revert = 60 s, guard = 5 ds
    Defects =
    Running Timers = PduReTx
    lastTxPdu = 0x00a0 Nr(RB )
    path-a Normal, RxId(I)= 4a:c6:ff:00:00:00, rx= v1-0x0020 Nr, cmd= None
    path-b Rpl, RxId(I)= 4a:c6:ff:00:00:00, rx= v1-0x0020 Nr, cmd= None
  DebugInfo:  aPathSts 9, bPathSts 5, pm (set/clr) 0/0, txFlush 0
    RxRaps: ok 40 nok 0 self 1756, TmrExp - wtr 4(2), grd 6, wtb 0
    Flush: cnt 14 (9/5/0) tm 000 00:49:11.680-000 00:49:11.680 Out/Ack 0/1
    RxRawRaps: aPath 954 bPath 976 vPath 0
    Now: 000 01:13:50.160 , softReset: No - noTx 0
  Seq Event  RxInfo(Path: NodeId-Bytes)
             state:TxInfo (Bytes)            Dir  pA  pB        Time
  === ===== ==============================  ===== === === ================
  004   bDn
              PROT  : 0xb020  Sf            TxF-> Fwd Blk 000 00:41:33.740
  005   pdu B: 4a:c4:ff:00:00:00-0x0000 Nr
              PROT  : 0xb020  Sf            Rx<-- Fwd Blk 000 00:41:33.840
  006   pdu A: 4a:c4:ff:00:00:00-0x0000 Nr
              PROT  : 0xb020  Sf            Rx<-- Fwd Blk 000 00:41:33.840
  007   pdu B: 4a:c4:ff:00:00:00-0xb040 Sf(DNF)
              PROT  : 0xb020  Sf            Rx<-- Fwd Blk 000 00:41:33.960
  008   pdu A: 4a:c4:ff:00:00:00-0xb040 Sf(DNF)
              PROT  : 0xb020  Sf            Rx<-- Fwd Blk 000 00:41:33.960
  009   bUp
              PEND-G: 0x0020  Nr            Tx--> Fwd Blk 000 00:41:35.980
  010   pdu B: 4a:c4:ff:00:00:00-0x0000 Nr
              PEND  : 0x0020  Nr            Rx<-- Fwd Blk 000 00:41:36.800
  011   pdu A: 4a:c4:ff:00:00:00-0x0000 Nr
              PEND  : 0x0020  Nr            Rx<-- Fwd Blk 000 00:41:36.800
  012  xWtr
              IDLE  : 0x00e0  Nr(RB DNF)    Tx--> Fwd Blk 000 00:42:50.680
  013   aDn
              PROT  : 0xb000  Sf            TxF-> Blk Fwd 000 00:47:58.630
  014   pdu B: 4a:c6:ff:00:00:00-0xb020 Sf
              PROT  : 0xb000  Sf            RxF<- Blk Fwd 000 00:48:01.440
  015   pdu A: 4a:c6:ff:00:00:00-0x0020 Nr
              PROT  : 0xb000  Sf            Rx<-- Blk Fwd 000 00:48:05.780
  016   pdu B: 4a:c6:ff:00:00:00-0x0020 Nr
              PROT  : 0xb000  Sf            Rx<-- Blk Fwd 000 00:48:05.780
  017   aUp
              PEND-G: 0x0000  Nr            Tx--> Blk Fwd 000 00:48:06.380
  018   pdu A: 4a:c6:ff:00:00:00-0x0020 Nr
              PEND-G: 0x0000  Nr            Rx<-- Blk Fwd 000 00:48:06.380
  019   pdu B: 4a:c6:ff:00:00:00-0x0020 Nr
              PEND-G: 0x0000  Nr            Rx<-- Blk Fwd 000 00:48:06.380
  000   pdu A: 4a:c6:ff:00:00:00-0x0020 Nr
              PEND  : 0x0000  Nr            Rx<-- Blk Fwd 000 00:48:11.380
  001   pdu
              PEND  :                       ----- Fwd Fwd 000 00:48:11.380
  002   pdu B: 4a:c6:ff:00:00:00-0x0020 Nr
              PEND  :                       Rx<-- Fwd Fwd 000 00:48:11.390
  003  xWtr
              IDLE  : 0x00a0  Nr(RB )       TxF-> Fwd Blk 000 00:49:11.68

*A:PE-2#
```

Conclusion

Ethernet Ring APS provides optimal solution for designing native Ethernet services with ring topology. This protocol provides simple configuration, operation and guaranteed fast protection time. SROS also has a flexible encapsulation that allows dot1Q, qinq or PBB for the ring traffic. It could be utilized for various services such as mobile backhaul, business VPN access, aggregation and core.

Layer 2 Services and EVPN

In This Section

This section provides configuration information for the following topics:

- Auto-Learn MAC Protect in EVPN
- BGP Multi-Homing for VPLS Networks
- BGP Virtual Private Wire Services
- BGP VPLS
- Black-hole MAC for EVPN Loop Protection
- Conditional Static Black-Hole MAC in EVPN
- EVPN for MPLS Tunnels
- EVPN for MPLS Tunnels in Epipe Services (EVPN-VPWS)
- EVPN for MPLS Tunnels in Routed VPLS
- EVPN for PBB over MPLS (PBB-EVPN)
- EVPN for VXLAN Tunnels (Layer 2)
- EVPN for VXLAN Tunnels (Layer 3)
- EVPN Interconnect Ethernet Segments
- EVPN-MPLS Interconnect for EVPN-VXLAN VPLS Services
- Fully Dynamic VSD Integration Model
- Inter-AS Model C for VLL
- Layer 2 Multicast Optimization for EVPN-VXLAN - Assisted Replication
- LDP VPLS Using BGP Auto-Discovery
- LDP VPLS Using BGP Auto-Discovery - Prefer Provisioned SDP
- Multi-Chassis Endpoint for VPLS Active/Standby Pseudowire
- Multi-Segment Pseudowire Routing
- P2MP mLDP Tunnels for BUM Traffic in EVPN-MPLS Services
- PBB-Epipe
- PBB-EVPN ISID-based CMAC Flush
- PBB-VPLS
- Preference-based and Non-revertive EVPN DF Election

- Shortest Path Bridging for MAC
- Virtual Ethernet Segments
- VLAN Range SAPs for VPLS and Epipe Services

Auto-Learn MAC Protect in EVPN

This chapter provides information about Auto-Learn MAC Protect in EVPN.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SR OS Release 14.0.R5, but the CLI in the current edition is based on SR OS Release 15.0.R2. Auto-Learn MAC Protect (ALMP) is supported for EVPN in SR OS Release 14.0.R1, and later.

Overview

MAC protection is needed in Layer 2 services to safeguard business-critical MAC addresses against the possibility of being learned on the wrong SAP/SDP. When a MAC address is learned on the wrong SAP/SDP, traffic would be diverted from its intended destination. This could be caused by misconfiguration or by a malicious source launching a Denial of Service (DoS) attack. MAC protect can also be used to prevent loops in certain topologies.

Chapter EVPN for VXLAN Tunnels (Layer 2) describes MAC protection for static MAC addresses that are configured on SAPs or spoke-SDPs. The command to configure static MAC addresses in a VPLS service is as follows:

```
configure service vpls 1 static-mac mac
  - mac <ieee-address> [create] black-hole
  - mac <ieee-address> [create] sap <sap-id> monitor {fwd-status}
  - no mac <ieee-address>
  - mac <ieee-address> [create] spoke-sdp <sdp-id:vc-id> monitor {fwd-status}
```

Configuring static MAC addresses is not scalable if large numbers of MAC addresses need to be protected. Also, configuring static MAC addresses is not an option when the MAC addresses are unknown. Auto-Learn MAC Protect (ALMP) offers the same protection for learned MAC addresses in services such as EVPN VPLS and EVPN R-VPLS. However, ALMP is not supported for PBB-EVPN.

ALMP can be enabled with the **auto-learn-mac-protect** command in EVPN with VXLAN or MPLS bindings on the following:

- SAPs
- Mesh-SDPs
- Spoke-SDPs
- Pseudowire (PW) templates
- Split Horizon Groups (SHGs)
- SHGs in PW templates

When enabled, all MAC addresses learned on those objects become protected.

The following commands can be used to enable ALMP on objects in VPLS 1:

```
configure service vpls 1 sap 1/1/1:1 auto-learn-mac-protect
configure service vpls 1 spoke-sdp 46:1 auto-learn-mac-protect
configure service vpls 1 mesh-sdp 56:1 auto-learn-mac-protect
configure service vpls 1 split-horizon-group "SHG1" auto-learn-mac-protect

configure service pw-template 1 auto-learn-mac-protect
```

When enabled on an SHG, it is only applicable to the SAPs within the SHG, not to spoke-SDPs. If ALMP is required on spoke-SDPs in the SHG, the parameter must be configured on each spoke-SDP individually. All MAC Source Addresses (SAs) learned on these objects will be protected and advertised with the sticky bit set. The sticky bit indicates that these MAC addresses should be treated as protected on the remote PEs, where these protected MAC addresses are considered to have been learned on the EVPN MPLS/VXLAN destinations. The remote EVPN peers then use the MAC protection functionality in the same way as the local peer to protect the MAC address.

ALMP enables an implicit **restrict-protected-src discard-frame** (RPS-DF) by default on SAPs and spoke/mesh-SDPs. When enabled, frames with a protected MAC SA are discarded if received on objects where they were not learned and protected. This configuration is the default and cannot be configured on objects where MAC addresses are learned, such as SAPs, spoke/mesh-SDPs, and SHGs.

However, RPS-DF can optionally be configured on destinations in EVPN MPLS or EVPN VXLAN, where data plane MAC learning is never performed for incoming traffic. For EVPN MPLS, the RPS-DF configuration is in the BGP EVPN context, as follows:

```
configure service vpls 1 bgp-evpn mpls restrict-protected-src discard-frame
```

For EVPN VXLAN, the RPS-DF configuration is in the VXLAN context, as follows:

```
configure service vpls 1 vxlan vni 1 restrict-protected-src discard-frame
```

Instead of discarding the frame, the SAP or spoke/mesh-SDP can be brought operationally down when a frame is received with a protected MAC SA that has not been learned on the object, by configuring **restrict-protected-src** (RPS) without any parameter on the object in EVPN services. After the object has been brought down, an operator needs to disable (**shutdown**) and enable (**no shutdown**) the object in order to make it operational again.

RPS can be enabled without any parameter on SAPs, spoke/mesh-SDPs, SHGs, and PW templates, but not on EVPN MPLS/VXLAN destinations, using following commands:

```
configure service vpls 1 sap 1/1/1:1 restrict-protected-src
configure service vpls 1 spoke-sdp 46:1 restrict-protected-src
configure service vpls 1 mesh-sdp 56:1 restrict-protected-src
configure service vpls 1 split-horizon-group "SHG1" restrict-protected-src

configure service pw-template 1 restrict-protected-src
configure service pw-template 1 split-horizon-group restrict-protected-src
```

The operation for an object is reverted to **restrict-protected-src discard-frame** after configuring the **no restrict-protected-src** command.

RPS cannot be configured without any parameter on EVPN MPLS destinations; if attempted, the following error will be raised:

```
*A:PE-2# configure service vpls 1 bgp-evpn mpls restrict-protected-src
                                                                       ^
Error: Missing parameter
```

Likewise, RPS cannot be configured without any parameter on EVPN VXLAN destinations; if attempted, the following error will be raised:

```
*A:PE-2# configure service vpls 1 vxlan vni 1 restrict-protected-src
                                                                     ^
Error: Missing parameter
```

Note: The configuration of restrict-protected-src alarm-only and restrict-unprotected-dst are not allowed in EVPN.

Protection is provided at the point where a MAC address first enters the EVPN part of the network. Therefore, the preference for an auto-learned protected MAC address is higher than that of a MAC address received in a BGP update with the sticky bit set.

The following list shows the MAC learning priority, with the highest priority first:

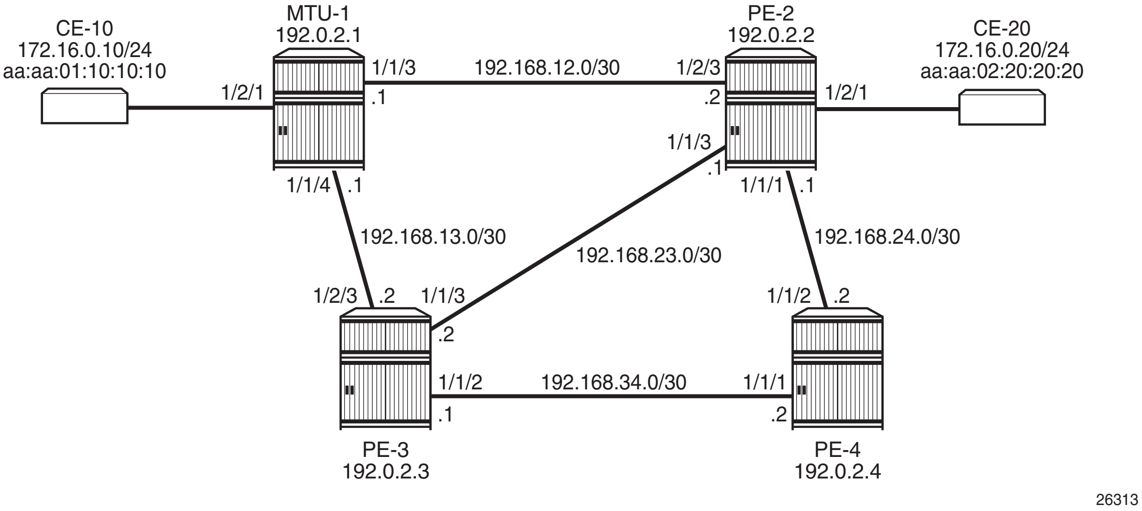
1. Local MAC address (including AS-MAC without static-black-hole, es-bmac, src-bmac, OAM, and so on)
2. Conditional static MAC address (including AS-MAC with static-black-hole)
3. **Auto-Learn Protected MAC address**
4. EVPN MAC address with sticky/static bit set
5. Data plane learned MAC address (regular learning on SAP/SDP-binding)
6. EVPN MAC address without sticky/static bit set

Note: ALMP MAC addresses have a higher priority but do not overwrite EVPN static MAC addresses.

Configuration

Figure 10 shows the example topology with one MTU and three PEs.

Figure 10 **Example Topology - No LAG**



- Cards, MDAs
- The ports between the PEs are configured as network ports; the other ports are access ports. No LAG is configured initially.
- IGP (IS-IS is used in this example) between the PEs
- LDP between the PEs
- BGP with address family EVPN on the PEs

PE-2 is the BGP route reflector. The BGP configuration on the PEs is similar. BGP is configured on PE-3 as follows:

```
configure
    router
        autonomous-system 64500
        bgp
            vpn-apply-import
            vpn-apply-export
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            split-horizon
            rapid-update evpn
            group "internal"
                family evpn
                peer-as 64500
                neighbor 192.0.2.2
                exit
            exit
        exit
```

VPLS 1 is configured on all nodes. Initially, ALMP is disabled. On MTU-1, the VPLS 1 contains three SAPs: one toward CE-10, one toward PE-2, and one toward PE-3.

On PE-2, VPLS 1 is configured with EVPN MPLS and contains a SAP toward CE-20 and a SAP toward MTU-1, as follows:

```
configure
    service
        vpls 1 customer 1 create
            bgp
            exit
            bgp-evpn
                evi 1
                mpls
                    ingress-replication-bum-label
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            sap 1/2/1:1 create
            exit
            sap 1/2/3:1 create
            exit
            no shutdown
        exit
```

On PE-3, VPLS 1 is configured with EVPN MPLS and contains a SAP toward MTU-1, as follows:

```
configure
    service
        vpls 1 customer 1 create
            bgp
            exit
            bgp-evpn
                evi 1
                mpls
                    ingress-replication-bum-label
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            sap 1/2/3:1 create
            exit
            no shutdown
        exit
```

The following use cases will be described in this section:

- EVPN MPLS without multi-homing.
  - Default behavior: no ALMP on SAPs, no protected MAC addresses
  - No ALMP on SAPs, RPS-DF on EVPN MPLS destinations
  - ALMP and implicit RPS-DF on SAPs.

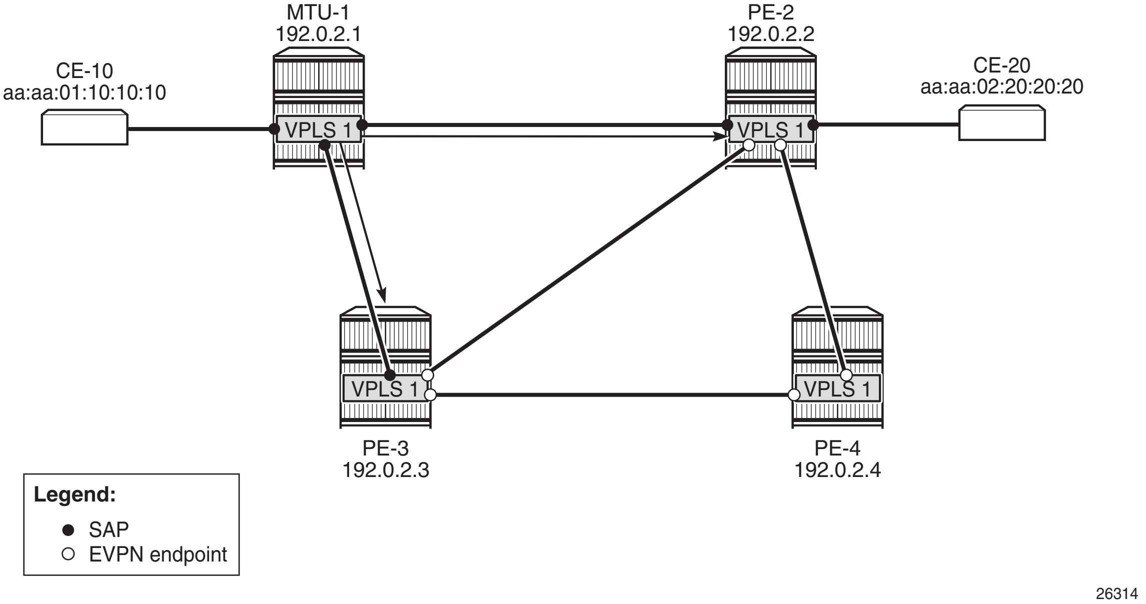
        - RPS-DF on EVPN MPLS destinations, MAC first learned on PE-2
        - RPS-DF on EVPN MPLS destinations, MAC simultaneously learned on PE-2 and PE-3
        - No RPS-DF on EVPN MPLS destinations, MAC simultaneously learned on PE-2 and PE-3
    - ALMP and RPS on SAPs.
        - RPS-DF on EVPN MPLS destinations, MAC first learned on PE-2
        - RPS-DF on EVPN MPLS destinations, MAC simultaneously learned on PE-2 and PE-3
        - No RPS-DF on EVPN MPLS destinations, MAC simultaneously learned on PE-2 and PE-3
- EVPN MPLS with ALMP in all-active multi-homing.
    - RPS-DF on SAPs, RPS-DF on EVPN MPLS destinations

Default Behavior: No Protected MAC Addresses

The following example is not a recommended configuration because it causes a loop. By default, ALMP is disabled and no static MAC addresses are configured. As described in chapter EVPN for VXLAN Tunnels (Layer 2), duplicate MAC addresses are detected in BGP EVPN and the MAC address will be put in a hold-down state on the EVPN destinations after a configurable threshold is reached. This applies to EVPN-MPLS as well as to EVPN-VXLAN. By default, the maximum number of MAC address moves is five in a time window of 3 minutes.

Figure 11 shows that the MAC address from CE-10 is learned simultaneously on the SAPs in VPLS 1 on PE-2 and PE-3.

Figure 11 **MAC Address Learned Simultaneously on SAPs on PE-2 and PE-3**



CE-10 sends frames to CE-20 with MAC Destination Address (DA) aa:aa:02:20:20:20. MTU-1 has not learned that MAC DA, so the frames are flooded to PE-2 and PE-3, where they enter the SAPs simultaneously. PE-2 and PE-3 have not learned the MAC DA either, so the frames are flooded to all potential destinations. The frames received on PE-2 will be sent (among others) to PE-3, and vice versa. These frames are forwarded back out of the SAP toward MTU-1. This causes a loop.

Both PEs send a BGP update for the MAC SA aa:aa:01:10:10:10 to the other PEs with no sticky bit set. That MAC SA is learned, but not protected on the destination to the other PE. The stream of frames will cause the learned MAC SA to oscillate between the SAP and EVPN destinations on PE-2 and PE-3, and between the EVPN destinations on PE-4.

After a configurable number of BGP EVPN MAC address moves in a time span (by default, after five MAC address moves in a period of 3 minutes), the MAC address is put in a hold-down state on the EVPN destinations for a specific duration (until the next MAC address duplication detection retry; by default, after 9 minutes).

The following message in log 99 on PE-2 (and also on PE-3) indicates that duplicate MAC addresses have been detected:

```
59 2017/05/11 10:12:45.46 UTC MINOR: SVCMGR #2331 Base
"VPLS Service 1 has MAC(s) detected as duplicates by EVPN mac-duplication
detection."
```

The following shows the settings for EVPN MAC address duplication detection, which are the default. It also lists the detected duplicate MAC addresses of CE-10 and CE-20:

```
*A:PE-3# show service id 1 bgp-evpn

===============================================================================
BGP EVPN Table
===============================================================================
MAC Advertisement  : Enabled            Unknown MAC Route  : Disabled
CFM MAC Advertise  : Disabled
VXLAN Admin Status : Disabled           Creation Origin    : manual
MAC Dup Detn Moves : 5                  MAC Dup Detn Window: 3
MAC Dup Detn Retry : 9                  Number of Dup MACs : 2
MAC Dup Detn BH    : Disabled
IP Route Advert    : Disabled

EVI                : 1
Ing Rep Inc McastAd: Enabled
Accept IVPLS Flush : Disabled
Send EVPN Encap    : Enabled

-------------------------------------------------------------------------------
Detected Duplicate MAC Addresses             Time Detected
-------------------------------------------------------------------------------
aa:aa:01:10:10:10                            05/11/2017 15:04:24
aa:aa:02:20:20:20                            05/11/2017 15:04:24
-------------------------------------------------------------------------------
===============================================================================

===============================================================================
BGP EVPN MPLS Information
===============================================================================
Admin Status       : Enabled
Force Vlan Fwding  : Disabled           Control Word       : Disabled
Split Horizon Group: (Not Specified)
Ingress Rep BUM Lbl: Enabled            Max Ecmp Routes    : 0
Ingress Ucast Lbl  : 262140             Ingress Mcast Lbl  : 262139
Entropy Label      : Disabled
RestProtSrcMacAct  : none
Evpn Mpls Encap    : Enabled            Evpn MplsoUdp      : Disabled
===============================================================================

===============================================================================
BGP EVPN MPLS Auto Bind Tunnel Information
===============================================================================
Resolution         : any
Filter Tunnel Types: (Not Specified)
===============================================================================
*A:PE-3#
```

RPS is disabled (by default) on the EVPN destinations (**RestProtSrcMacAct : none**).

The MAC addresses are in a hold-down state on the EVPN destinations and no MAC address moves take place until the next MAC address duplication detection retry after 9 minutes. After 9 minutes, the EVPN MAC address duplication alarm is cleared, but after the next five MAC address moves within a time span of 3 minutes, the alarm is raised again and this threshold is reached soon after the alarm has been cleared.

The MAC addresses of both CEs are learned on the SAP of PE-3 (CE-20's MAC address is also learned on the SAP toward MTU-1), not on the EVPN destinations, because of the MAC address duplication detection and hold-down state in EVPN, as follows:

```
*A:PE-3# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         aa:aa:01:10:10:10 sap:1/2/3:1              L/0      05/11/17 15:04:24
1         aa:aa:02:20:20:20 sap:1/2/3:1              L/0      05/11/17 15:04:24
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-3#
```

A similar output can be shown for PE-2.

Both PE-2 and PE-3 learn the MAC addresses locally and send BGP EVPN MAC address route updates to their BGP peers. PE-3 received the following BGP EVPN MAC address routes from PE-2, with the MAC address mobility sequence number representing the number of MAC address moves:

```
*A:PE-3# show router bgp routes evpn mac
===============================================================================
 BGP Router ID:192.0.2.3        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN MAC Routes
===============================================================================
Flag  Route Dist.         MacAddr           ESI
      Tag                 Mac Mobility      Label1
                          Ip Address
                          NextHop
-------------------------------------------------------------------------------
u*>i  192.0.2.2:1         aa:aa:01:10:10:10 ESI-0
```

```
      0                   Seq:4            LABEL 262140
                          N/A
                          192.0.2.2

u*>i  192.0.2.2:1         aa:aa:02:20:20:20 ESI-0
      0                   Seq:4            LABEL 262140
                          N/A
                          192.0.2.2

-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-3#
```

PE-3 does not use these BGP EVPN MAC address routes in its FDB, because locally learned MAC addresses are preferred.

The remote PE (PE-4) received the following BGP EVPN MAC routes from PE-2 and PE-3:

```
*A:PE-4# show router bgp routes evpn mac
===============================================================================
 BGP Router ID:192.0.2.4        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN MAC Routes
===============================================================================
Flag  Route Dist.         MacAddr          ESI
      Tag                 Mac Mobility     Label1
                          Ip Address
                          NextHop
-------------------------------------------------------------------------------
u*>i  192.0.2.2:1         aa:aa:01:10:10:10 ESI-0
      0                   Seq:4            LABEL 262140
                          N/A
                          192.0.2.2

u*>i  192.0.2.2:1         aa:aa:02:20:20:20 ESI-0
      0                   Seq:4            LABEL 262140
                          N/A
                          192.0.2.2

u*>i  192.0.2.3:1         aa:aa:01:10:10:10 ESI-0
      0                   Seq:3            LABEL 262140
                          N/A
                          192.0.2.3

u*>i  192.0.2.3:1         aa:aa:02:20:20:20 ESI-0
      0                   Seq:5            LABEL 262140
                          N/A
                          192.0.2.3
```

```
-------------------------------------------------------------------------------
Routes : 4
===============================================================================
*A:PE-4#
```

In the preceding output, MAC aa:aa:01:10:10:10 is learned from BGP peer 192.0.2.3 with MAC mobility sequence number 3, and from BGP peer 192.0.2.2 with sequence number 4. MAC aa:aa:02:20:20:20 is learned from BGP peer 192.0.2.2 with sequence number 4 and from BGP peer 192.0.2.3 with sequence number 5. The FDB for VPLS 1 on PE-4 contains the MAC addresses learned from BGP EVPN MAC updates with the highest MAC mobility sequence number, as follows:

```
*A:PE-4# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         aa:aa:01:10:10:10 eMpls:                   Evpn     05/11/17 15:04:23
                            192.0.2.2:262140
1         aa:aa:02:20:20:20 eMpls:                   Evpn     05/11/17 15:04:23
                            192.0.2.3:262140
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-4#
```

VPLS 1 on MTU-1 does not have EVPN configured and no MAC address duplication detection mechanism implemented. The MAC address from CE-10 is last learned on the SAP toward PE-2 (it might equally have been the SAP toward PE-3) instead of the SAP toward CE-10, resulting from the loop, as follows:

```
*A:MTU-1# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         aa:aa:01:10:10:10 sap:1/1/3:1              L/0      05/11/17 15:04:53
1         aa:aa:02:20:20:20 sap:1/1/3:1              L/0      05/11/17 15:04:45
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:MTU-1#
```

No ALMP on SAPs, RPS-DF on EVPN Destinations

When there are no protected MAC addresses (ALMP is disabled and no static MAC addresses are configured), the behavior is as described earlier. RPS-DF discards frames with protected MAC addresses that were not learned on the object, but there are no protected MAC addresses, because ALMP is not configured. RPS-DF does not discard frames with MAC SAs that are not protected.

RPS-DF is enabled on EVPN destinations on all PEs, as follows:

```
configure service vpls 1 bgp-evpn mpls restrict-protected-src discard-frame
```

The state of RPS is now "discard-frame" instead of "none", as follows:

```
*A:PE-3# show service id 1 bgp-evpn | match RestProtSrcMacAct
RestProtSrcMacAct  : Discard-frame
```

It is also allowed to configure RPS without parameters on the SAPs, but that does not change the behavior when ALMP is disabled and there are no protected MAC addresses. RPS will not bring down a SAP after receiving a frame with an unprotected MAC SA.

ALMP and Implicit RPS-DF on SAPs

ALMP is enabled on the SAPs in PE-2 as follows:

```
configure service vpls 1 sap 1/2/1:1 auto-learn-mac-protect     # SAP toward CE-20
configure service vpls 1 sap 1/2/3:1 auto-learn-mac-protect     # SAP toward MTU-1
```

The configuration is similar on PE-3.

The following shows that ALMP is enabled on the SAP and that the default RPS-DF is used:

```
*A:PE-2# show service id 1 sap 1/2/3:1 detail

===============================================================================
Service Access Points(SAP)
===============================================================================
Service Id         : 1
SAP                : 1/2/3:1                  Encap             : q-tag
Description        : (Not Specified)
Admin State        : Up                       Oper State        : Up
Flags              : None
---snip---
Restr MacUnpr Dst  : Disabled
Auto Learn Mac Prot: Enabled
```

```
RestMacProtSrc Act : none (oper: Discard-frame)
---snip---
```

ALMP and RPS-DF on SAPs, RPS-DF on EVPN MPLS Destinations, MAC First Learned on PE-2

Initially, the SAP on PE-3 is shut down to ensure that the MAC address will first be learned on PE-2, then on PE-3, as follows:

```
*A:PE-3# configure service vpls 1 sap 1/2/3:1 shutdown
```

Each learned MAC address on the SAPs on PE-2 will be protected; therefore, a BGP update with the static/sticky bit set will be sent to the BGP EVPN peers. In this example, the MAC aa:aa:01:10:10:10 of CE-10 is learned first on SAP 1/2/3:1 on PE-2, and MAC aa:aa:02:20:20:20 is learned on SAP 1/2/1:1 on PE-2. Consequently, PE-2 sends BGP updates with the static/sticky bit set to PE-3 for both MAC aa:aa:01:10:10:10 and MAC aa:aa:02:20:20:20, as follows:

```
59 2017/05/11 15:06:37.07 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 96
    Flag: 0x90 Type: 14 Len: 44 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.2
        Type: EVPN-MAC Len: 33 RD: 192.0.2.2:1 ESI: ESI-0, tag: 0, mac len: 48
                       mac: aa:aa:01:10:10:10, IP len: 0, IP: NULL, label1: 4194240
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:1
        bgp-tunnel-encap:MPLS
        mac-mobility:Seq:0/Static
"


61 2017/05/11 15:06:37.08 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.5
"Peer 1: 192.0.2.5: UPDATE
Peer 1: 192.0.2.5 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 96
    Flag: 0x90 Type: 14 Len: 44 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.2
        Type: EVPN-MAC Len: 33 RD: 192.0.2.2:1 ESI: ESI-0, tag: 0, mac len: 48
                       mac: aa:aa:02:20:20:20, IP len: 0, IP: NULL, label1: 4194240
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:1
        bgp-tunnel-encap:MPLS
        mac-mobility:Seq:0/Static
"
```

Note: The MPLS label is label1 in the BGP update divided by 16 (2^4), as follows:

Figure 12

$$\frac{4194240}{16} = 262140$$

PE-2 sends similar BGP EVPN updates to peer PE-4.

After these BGP EVPN updates have been sent to PE-3 (and PE-4), the SAP on PE-3 is enabled again, as follows:

```
*A:PE-3# configure service vpls 1 sap 1/2/3:1 no shutdown
```

The MAC addresses in the FDB on PE-2, where these MAC addresses are learned, get the indication "L" for learned and "P" for protected MAC address, as follows:

```
*A:PE-2# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         aa:aa:01:10:10:10 sap:1/2/3:1              LP/60    05/11/17 15:06:37
1         aa:aa:02:20:20:20 sap:1/2/1:1              LP/60    05/11/17 15:06:37
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-2#
```

The MAC addresses in the FDB on PE-3 are learned from the BGP EVPN updates and get the indication "S" for static (sticky bit) and "P" for protected MAC address, as follows

```
*A:PE-3# show service id 1 fdb detail
```

```
===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         aa:aa:01:10:10:10 eMpls:                   EvpnS    05/11/17 15:06:35
                                                     P
                            192.0.2.2:262140
1         aa:aa:02:20:20:20 eMpls:                   EvpnS    05/11/17 15:06:35
                                                     P
                            192.0.2.2:262140
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-3#
```

The FDB on the remote PE (PE-4) looks similar, as follows:

```
*A:PE-4# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         aa:aa:01:10:10:10 eMpls:                   EvpnS    05/11/17 15:06:34
                                                     P
                            192.0.2.2:262140
1         aa:aa:02:20:20:20 eMpls:                   EvpnS    05/11/17 15:06:34
                                                     P
                            192.0.2.2:262140
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-4#
```

The BGP EVPN MAC address routes on PE-3 have MAC address mobility equal to "static", as follows:

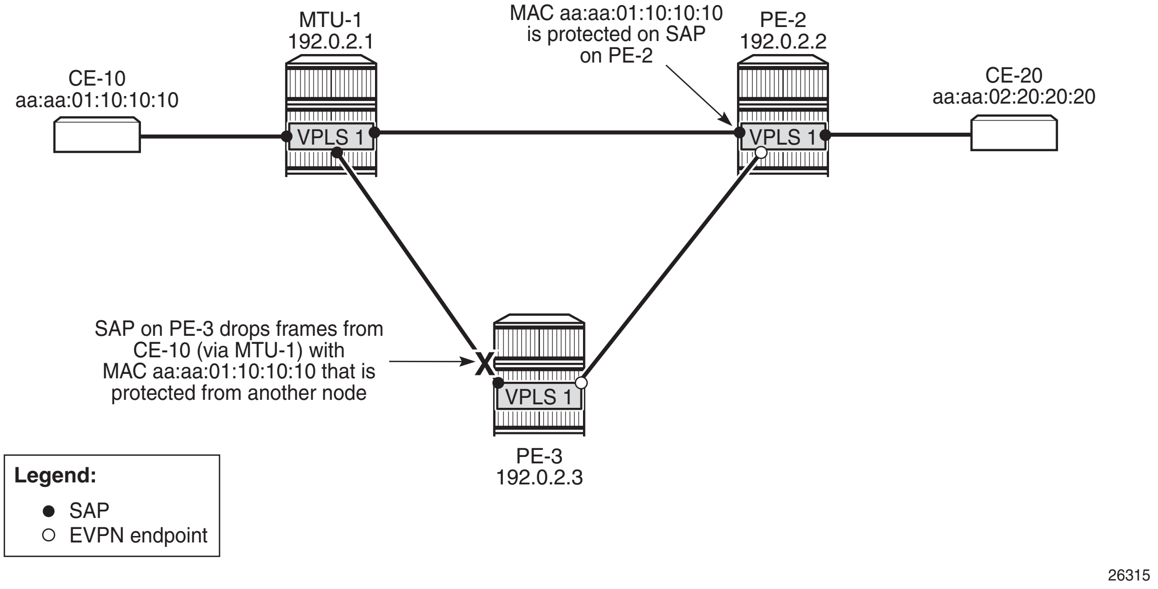
```
*A:PE-3# show router bgp routes evpn mac
===============================================================================
 BGP Router ID:192.0.2.3        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN MAC Routes
```

```
===============================================================================
Flag  Route Dist.         MacAddr           ESI
      Tag                 Mac Mobility      Label1
                          Ip Address
                          NextHop
-------------------------------------------------------------------------------
u*>i  192.0.2.2:1         aa:aa:01:10:10:10 ESI-0
      0                   Static            LABEL 262140
                          N/A
                          192.0.2.2

u*>i  192.0.2.2:1         aa:aa:02:20:20:20 ESI-0
      0                   Static            LABEL 262140
                          N/A
                          192.0.2.2

-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-3#
```

The BGP EVPN MAC routes on PE-4 are similar.

When a stream of frames with MAC SA aa:aa:01:10:10:10 enters the SAP on PE-3, these frames will be dropped by this SAP because of the implicit RPS-DF behavior in the SAP for protected MAC addresses, as shown in Figure 13.

Figure 13 Default RPS-DF on SAPs - MAC Learned and Protected on SAP on PE-2



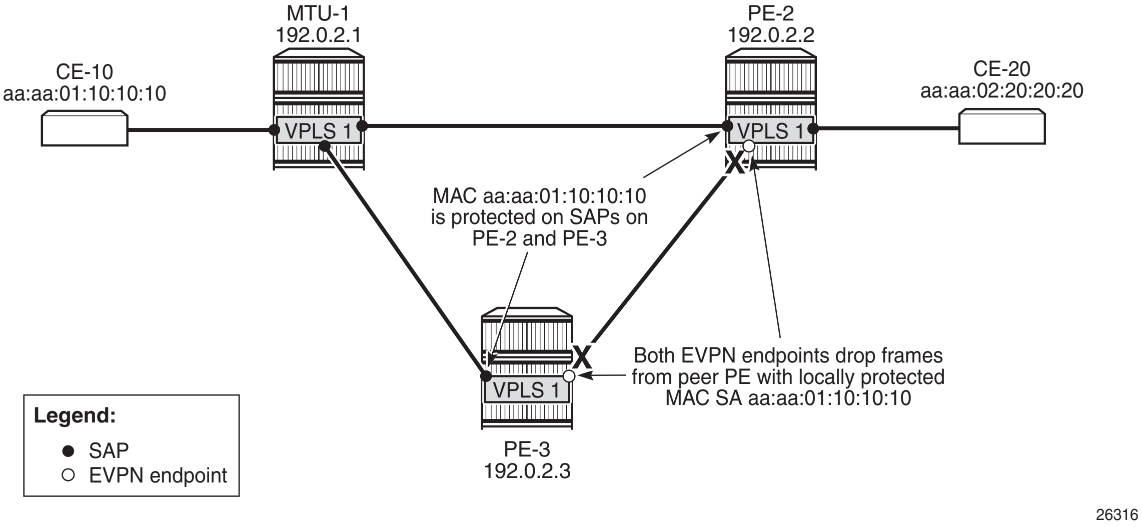
Because the MAC address was protected on the SAP on PE-2 and the BGP EVPN MAC route update had been received by PE-3 before any frame was received with this MAC SA, there will be no temporary loop. The frames with the protected MAC SA will be discarded at the SAP on PE-3, not on the EVPN MPLS destination on PE-2. In this case, there is no need to configure RPS-DF on the EVPN MPLS destinations, but it will make a difference when the MAC address is learned on both SAPs simultaneously.

ALMP and RPS-DF on SAPs, RPS-DF on EVPN MPLS Destinations, MAC Simultaneously Learned on PE-2 and PE-3

In the preceding example, the MAC addresses of CE-10 and CE-20 were first learned and protected on PE-2 and received on PE-3's SAP after the BGP update with static/sticky bit was received by PE-3. However, when the MAC address of CE-10 is learned simultaneously on both PEs, for example, because the MAC DA aa:aa:02:20:20:20 is unknown, there is a temporary loop until the MAC addresses are protected. Initially, the frames enter a SAP, are forwarded to the EVPN peer, and forwarded out of the remote SAP.

After the MAC addresses are learned and protected on the SAPs on both PEs, new frames received on a SAP with the protected MAC address will be sent to the other PE. However, they will be discarded due to RPS-DF on destination, as shown in Figure 14, because the destination PE has that same MAC address protected on its local SAP. This prevents a loop. BGP updates with the static/sticky bit set are sent to the BGP EVPN peer, but the locally learned and protected MAC address is preferred to the MAC address in a BGP update. Therefore, the FDB contains the locally learned MAC address aa:aa:01:10:10:10, not the BGP EVPN MAC address update for MAC address aa:aa:01:10:10:10.

Figure 14 **MAC Learned and Protected Simultaneously on PEs - RPS-DF on EVPN Endpoints**



The MAC addresses of the CEs are cleared from the FDBs on all nodes, as follows:

```
clear service id 1 fdb mac aa:aa:01:10:10:10
clear service id 1 fdb mac aa:aa:02:20:20:20
```

This clear command for the FDB only works for auto-learned MAC addresses, not for BGP EVPN MAC address updates. BGP EVPN MAC address withdraw updates need to be sent. In this example, BGP is configured with **rapid-update evpn**, as shown previously.

When traffic is sent from CE-10 to CE-20, MAC address aa:aa:01:10:10:10 of CE-10 is learned simultaneously on SAP 1/2/3:1 in PE-2 and PE-3 and protected on both SAPs. MAC address aa:aa:02:20:20:20 is, in this case, first learned via MAC address learning on PE-2 and advertised via a BGP EVPN MAC address route update. However, it might happen that it was learned and protected on the SAP on PE-3 first, before the MAC address was learned and protected on PE-2 and the BGP EVPN MAC address route update sent by PE-2 was received at PE-3. In the latter case, both MAC address aa:aa:01:10:10:10 and MAC address aa:aa:02:20:20:20 are learned and protected on the SAPs on both PE-2 and PE-3, and RPS-DF on the EVPN-MPLS destinations prevents loops.

However, in the present case, MAC address aa:aa:02:20:20:20 is only protected on the SAP on PE-2, because PE-3 received the EVPN MAC address update before it received a frame with MAC SA aa:aa:02:20:20:20. Therefore, the SAP on PE-3 will discard any frames with MAC SA aa:aa:02:20:20:20.

The FDB for VPLS 1 on PE-2 shows that both MAC addresses are learned locally and protected, as follows:

```
*A:PE-2# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         aa:aa:01:10:10:10 sap:1/2/3:1              LP/0     05/11/17 15:09:17
1         aa:aa:02:20:20:20 sap:1/2/1:1              LP/0     05/11/17 15:09:17
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-2#
```

The FDB for VPLS 1 on PE-3 shows that MAC address aa:aa:01:10:10:10 is learned and protected locally, but MAC address aa:aa:02:20:20:20 is protected on PE-2, which has been advertised by PE-2 in a BGP EVPN MAC update, as follows:

```
*A:PE-3# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         aa:aa:01:10:10:10 sap:1/2/3:1              LP/0     05/11/17 15:09:15
1         aa:aa:02:20:20:20 eMpls:                   EvpnS    05/11/17 15:09:15
                                                     P
                            192.0.2.2:262140
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-3#
```

Both PE-2 and PE-3 send BGP EVPN MAC updates to their BGP peers for each locally learned and protected MAC address. The following BGP EVPN MAC update is sent by PE-2 to PE-3 for MAC address aa:aa:01:10:10:10:

```
67 2017/05/11 15:09:16.95 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 96
    Flag: 0x90 Type: 14 Len: 44 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.2
        Type: EVPN-MAC Len: 33 RD: 192.0.2.2:1 ESI: ESI-0, tag: 0, mac len: 48
                  mac: aa:aa:01:10:10:10, IP len: 0, IP: NULL, label1: 4194240
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
```

```
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:1
        bgp-tunnel-encap:MPLS
        mac-mobility:Seq:0/Static
"
```

Similar BGP EVPN updates are sent to the remote PE (PE-4). The FDB for VPLS 1 on PE-4 only contains entries learned from BGP EVPN updates, as follows:

```
*A:PE-4# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         aa:aa:01:10:10:10 eMpls:                   EvpnS    05/11/17 15:09:14
                                                     P
                            192.0.2.2:262140
1         aa:aa:02:20:20:20 eMpls:                   EvpnS    05/11/17 15:09:14
                                                     P
                            192.0.2.2:262140
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-4#
```

PE-4 received BGP EVPN MAC address route updates from PE-2 and PE-3, but only installs the MAC address routes to PE-2 in its FDB, based on the lowest next-hop IP of the EVPN NLRI (192.0.2.2).

ALMP and RPS-DF on SAPs, No RPS-DF on EVPN MPLS Destinations, MAC Simultaneously Learned on PE-2 and PE-3

RPS-DF is disabled on the EVPN MPLS destinations on the PEs, as follows:

```
configure service vpls 1 bgp-evpn mpls no restrict-protected-src
```

When a frame is received at SAP 1/2/3:1 on PE-3 with protected MAC SA aa:aa:01:10:10:10, it is not dropped by the SAP, because this MAC SA has been learned and protected on this SAP on PE-3. The frame is forwarded to PE-2 where it will not be discarded by the EVPN MPLS destination because RPS-DF is disabled. The frame will be forwarded to other objects in the VPLS in PE-2. For BUM traffic, there will be a loop, because all frames will be flooded to all objects in VPLS 1 on PE-2, including the SAP toward MTU-1.

ALMP and RPS on SAPs

When ALMP is enabled on an object, the default behavior is that frames with a protected MAC SA are discarded (RPS-DF). However, it is possible to configure RPS without any parameter on the object, in this case on the SAPs on PE-2 and PE-3, as follows:

```
configure service vpls 1 sap 1/2/3:1 restrict-protected-src
```

Instead of discarding frames with MAC SAs that are protected on another object or node, the entire object (here: SAP) can be brought operationally down after a frame has been received with a MAC SA that is protected on another node.

The RPS configuration on the SAP can be shown as follows. The SAP has not been brought down yet.

```
*A:PE-2# show service id 1 sap 1/2/3:1 detail

===============================================================================
Service Access Points(SAP)
===============================================================================
Service Id         : 1
SAP                : 1/2/3:1                  Encap             : q-tag
Description        : (Not Specified)
Admin State        : Up                       Oper State        : Up
Flags              : None
---snip---
Restr MacUnpr Dst  : Disabled
Auto Learn Mac Prot: Enabled
RestMacProtSrc Act : SAP-oper-down
---snip---
```

The **RestMacProtSrc Act** parameter is set to **SAP-oper-down**, meaning that RPS is configured without any parameter, which causes the system to bring down the SAP when a duplicate MAC address is received that is protected on another object or node. When a SAP is brought down because of this, the **RxProtSrcMAC** flag will be raised and can be shown in the detailed SAP show output.

ALMP and RPS on SAPs, RPS-DF on EVPN MPLS Destinations, MAC First Learned on PE-2

RPS-DF is enabled on the EVPN MPLS destinations on the PEs, as follows:

```
configure service vpls 1 bgp-evpn mpls restrict-protected-src discard-frame
```

To simulate a scenario where the MAC addresses are first learned on PE-2, the SAP on PE-3 is shut down until the BGP EVPN MAC route updates are sent, as follows:

```
configure service vpls 1 sap 1/2/3:1 shutdown
```

The FDBs are cleared on the nodes, as follows:

```
clear service id 1 fdb mac aa:aa:01:10:10:10
clear service id 1 fdb mac aa:aa:02:20:20:20
```

Traffic is sent between CE-10 and CE-20, and the MAC addresses are learned and protected on the SAP on PE-2, as follows:

```
*A:PE-2# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         aa:aa:01:10:10:10 sap:1/2/3:1              LP/0     05/11/17 15:10:56
1         aa:aa:02:20:20:20 sap:1/2/1:1              LP/0     05/11/17 15:10:56
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-2#
```

No MAC learning took place on the SAP on PE-3, and the FDB contains the MAC addresses from the BGP EVPN updates, as follows:

```
*A:PE-3# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         aa:aa:01:10:10:10 eMpls:                   EvpnS    05/11/17 15:10:53
                                                     P
                            192.0.2.2:262140
1         aa:aa:02:20:20:20 eMpls:                   EvpnS    05/11/17 15:10:53
```

```
                                                    P
                            192.0.2.2:262140
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-3#
```

The SAP on PE-3 is enabled, as follows:

```
configure service vpls 1 sap 1/2/3:1 no shutdown
```

The operational state of the SAP is up, because no protected MAC addresses have been received yet:

```
*A:PE-3# show service id 1 sap

===============================================================================
SAP(Summary), Service 1
===============================================================================
PortId                          SvcId      Ing.  Ing.    Egr.  Egr.   Adm  Opr
                                           QoS   Fltr    QoS   Fltr
-------------------------------------------------------------------------------
1/2/3:1                         1          1     none    1     none   Up   Up
-------------------------------------------------------------------------------
Number of SAPs : 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-3#
```

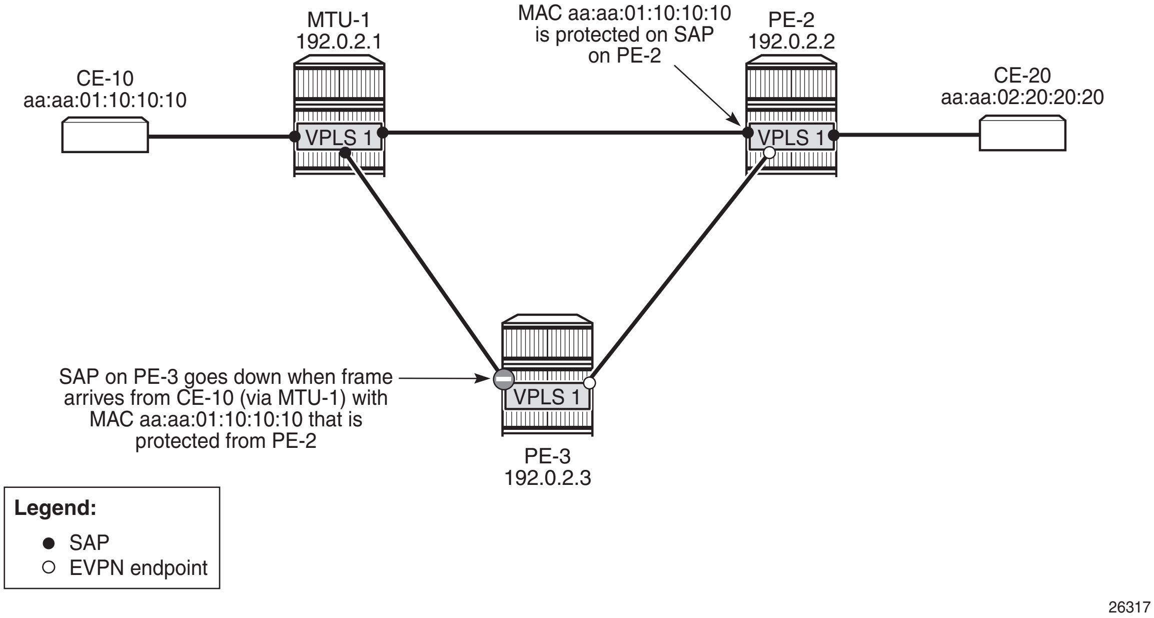
The FDB is cleared for MAC address aa:aa:02:20:20:20 on MTU-1, as follows:

```
clear service id 1 fdb mac aa:aa:02:20:20:20
```

Traffic from CE-10 toward the unknown MAC address aa:aa:02:20:20:20 reaches the SAPs on PE-2 and PE-3. When MAC SA aa:aa:01:10:10:10, which is protected on PE-2, is received on PE-3, SAP 1/2/3:1 will be brought operationally down, as shown in Figure 15, and the following alarms will be raised in log 99:

```
71 2017/05/11 15:11:48.91 UTC MINOR: SVCMGR #2208 Base
"Protected MAC aa:aa:01:10:10:10 received on SAP 1/2/3:1 in service 1.
The SAP will be disabled."

72 2017/05/11 15:11:48.91 UTC MINOR: SVCMGR #2203 Base
"Status of SAP 1/2/3:1 in service 1 (customer 1) changed to admin=up oper=down
flags=RxProtSrcMac "
```

Figure 15 **MAC Learned and Protected on SAP on PE-2 - RPS Enabled on SAP on PE-3**



The operational state of SAP 1/2/3:1 is now down, as follows:

```
*A:PE-3# show service id 1 sap

===============================================================================
SAP(Summary), Service 1
===============================================================================
PortId                          SvcId      Ing.  Ing.    Egr.  Egr.   Adm  Opr
                                           QoS   Fltr    QoS   Fltr
-------------------------------------------------------------------------------
1/2/3:1                         1          1     none    1     none   Up   Down
-------------------------------------------------------------------------------
Number of SAPs : 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-3#
```

Detailed information about this SAP shows the **RxProtSrcMAC** flag, indicating that a duplicate MAC address that is protected on a remote node has been received, as follows:

```
*A:PE-3# show service id 1 sap 1/2/3:1

===============================================================================
Service Access Points(SAP)
===============================================================================
Service Id         : 1
SAP                : 1/2/3:1                  Encap             : q-tag
Description        : (Not Specified)
Admin State        : Up                       Oper State        : Down
```

```
Flags              : RxProtSrcMac
Multi Svc Site     : None
---snip---

*A:PE-3# show service id 1 sap 1/2/3:1 detail

===============================================================================
Service Access Points(SAP)
===============================================================================
Service Id         : 1
SAP                : 1/2/3:1                  Encap               : q-tag
Description        : (Not Specified)
Admin State        : Up                       Oper State          : Down
Flags              : RxProtSrcMac
---snip---
Restr MacUnpr Dst  : Disabled
Auto Learn Mac Prot: Enabled
RestMacProtSrc Act : SAP-oper-down
---snip---
```

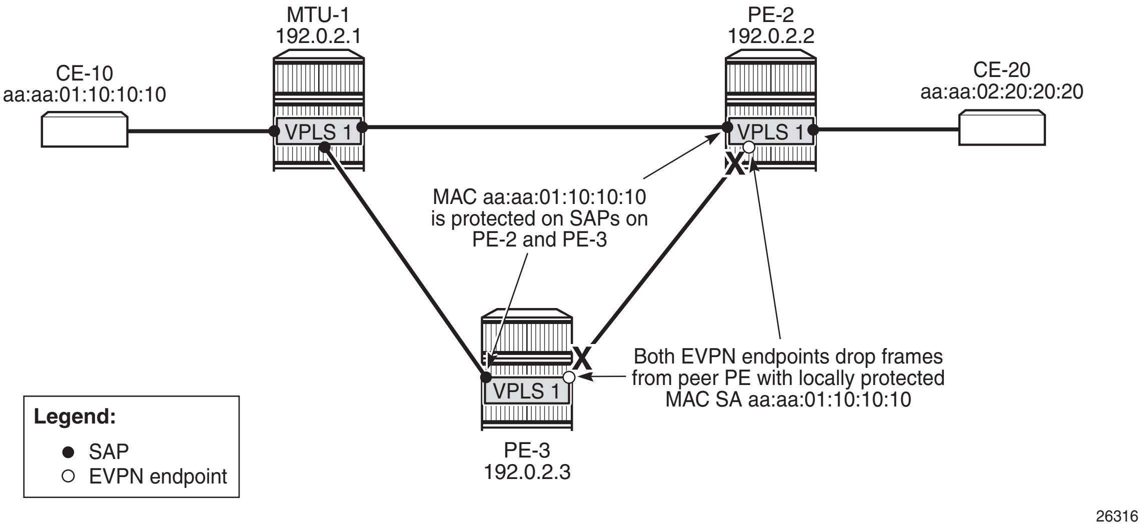
The SAP is operationally down and will not come up automatically when the FDB is cleared. To bring the SAP up, an operator needs to disable and re-enable the SAP, as follows:

```
*A:PE-3# configure service vpls 1 sap 1/2/3:1 shutdown
*A:PE-3# configure service vpls 1 sap 1/2/3:1 no shutdown
*A:PE-3# show service id 1 sap

===============================================================================
SAP(Summary), Service 1
===============================================================================
PortId                          SvcId      Ing.  Ing.    Egr.  Egr.   Adm  Opr
                                           QoS   Fltr    QoS   Fltr
-------------------------------------------------------------------------------
1/2/3:1                         1          1     none    1     none   Up   Up
-------------------------------------------------------------------------------
Number of SAPs : 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-3#
```

ALMP and RPS on SAPs, RPS-DF on EVPN MPLS Destinations, MAC Simultaneously Learned on PE-2 and PE-3

When CE-10 sends traffic to CE-20 and the destination MAC address is unknown, MAC address aa:aa:01:10:10:10 is simultaneously learned and protected on PE-2 and PE-3. No SAP will be brought down when MAC address aa:aa:01:10:10:10 is received on PE-2 or PE-3. This scenario is identical to the one with ALMP and (default) RPS-DF on the SAPs, as shown in Figure 16 (which is identical to Figure 14).

Figure 16 **RPS Enabled on SAPs - RPS-DF on EVPN Endpoints, MACs Learned Simultaneously**



A temporary loop is possible until the MAC address is protected on the SAPs. Initially, the frames enter the SAP, are forwarded to the other PEs, and are forwarded out of the other SAP (unless the MAC address is protected). When the MAC address is protected, any other frames received on the SAP will be sent to the other PE (for example, from PE-3 to PE-2, or vice versa), but they will be discarded by the receiving PE, because RPS-DF is applied on the EVPN destination. BGP EVPN updates are sent to the peer PEs with the sticky bit set. This MAC route will not be installed in the FDB of PE-2 and PE-3 because the MAC address has already been learned locally, which has a higher preference.

The FDB on PE-2 contains locally learned and protected MAC addresses, as follows:

```
*A:PE-2# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         aa:aa:01:10:10:10 sap:1/2/3:1              LP/0     05/11/17 15:10:56
1         aa:aa:02:20:20:20 sap:1/2/1:1              LP/0     05/11/17 15:10:56
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-2#
```

The FDB on PE-3 contains MAC address aa:aa:01:10:10:10 that is locally learned and protected, and MAC address aa:aa:02:20:20:20 that is protected on PE-2, as follows:

```
*A:PE-3# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         aa:aa:01:10:10:10 sap:1/2/3:1              LP/0     05/11/17 15:13:36
1         aa:aa:02:20:20:20 eMpls:                   EvpnS    05/11/17 15:13:36
                                                     P
                            192.0.2.2:262140
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-3#
```

The SAP will not be brought down if frames are received with MAC address aa:aa:01:10:10:10. However, MAC address aa:aa:02:20:20:20 was learned and protected first on PE-2 and the BGP update was received by PE-3 before the MAC address was received on PE-3. Therefore, MAC address aa:aa:02:20:20:20 will not be learned and protected on PE-3 and, if frames with a MAC SA aa:aa:02:20:20:20 were received on SAP 1/2/3:1 on PE-3, the SAP would be brought down.

ALMP and RPS on SAPs, No RPS-DF on EVPN MPLS Destinations, MAC Simultaneously Learned on PE-2 and PE-3

RPS-DF is disabled on the EVPN MPLS destinations on the PEs, as follows:

```
configure service vpls 1 bgp-evpn mpls no restrict-protected-src
```

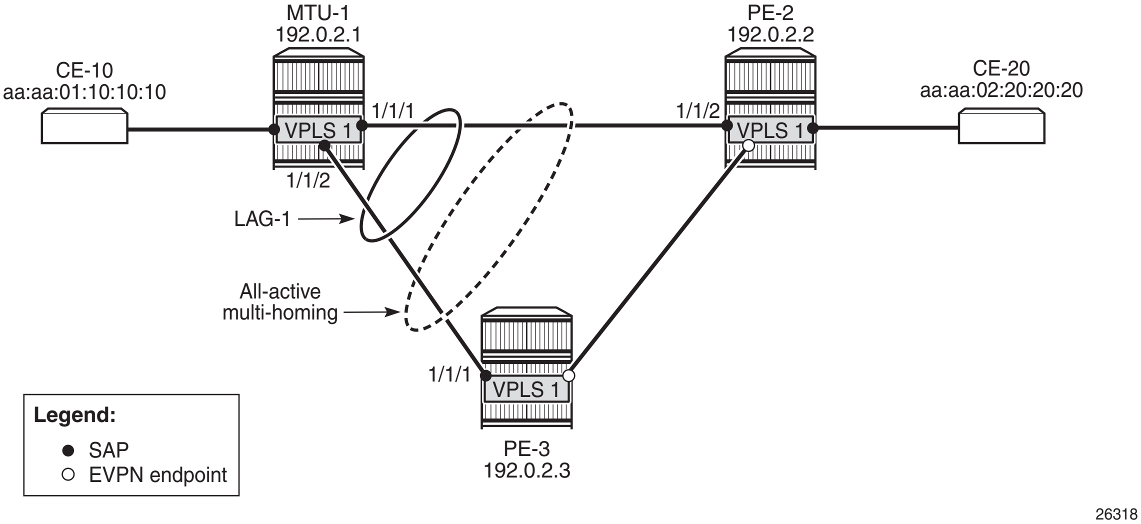
When frames are received at SAP 1/2/3:1 on PE-3 with protected MAC SA aa:aa:01:10:10:10, the SAP is not brought down, because this MAC SA has been learned and protected on this SAP. The frame is forwarded to PE-2 where it will not be discarded by the EVPN MPLS destination because RPS-DF is disabled. It will be forwarded to other objects in the VPLS. For BUM traffic, there will be a loop, because the frames will be flooded to all objects, including the SAP on PE-2 toward MTU-1.

ALMP in All-Active Multi-Homing SAPs

All-active multi-homing for EVPN MPLS is explained in chapter EVPN for MPLS Tunnels. ALMP is not required on all-active multi-homing SAPs. The following example shows that traffic can be dropped when ALMP is enabled on the SAPs and RPS-DF is enabled on the EVPN-MPLS destinations.

Figure 17 shows the example topology for all-active multi-homing.

Figure 17 **ALMP in All-Active Multi-Homing SAPs**



VPLS is configured with SAP lag-1:1 on the three nodes in the topology, as follows:

```
configure service vpls 1 sap lag-1:1 create
```

The SAPs used in the preceding scenarios are removed.

All-active multi-homing is configured on PE-2 and PE-3, as follows:

```
configure
    service
        system
            bgp-evpn
                ethernet-segment "ESI-23" create
                    esi 01:00:00:00:00:23:00:00:00:01
                    es-activation-timer 3
                    service-carving
                        mode auto
                    exit
                    multi-homing all-active
                    lag 1
                    no shutdown
                exit
            exit
        exit
    exit
exit
```

ALMP is enabled on the SAPs on PE-2 and PE-3, as follows:

```
configure service vpls 1 sap lag-1:1 auto-learn-mac-protect
```

MAC address aa:aa:01:10:10:10 is learned and protected on PE-2 and PE-3, as follows:

```
*A:PE-2# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         aa:aa:01:10:10:10 sap:lag-1:1              EvpnS    05/11/17 15:16:01
                                                     P
1         aa:aa:02:20:20:20 sap:1/2/1:1              LP/60    05/11/17 15:13:38
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-2#
```

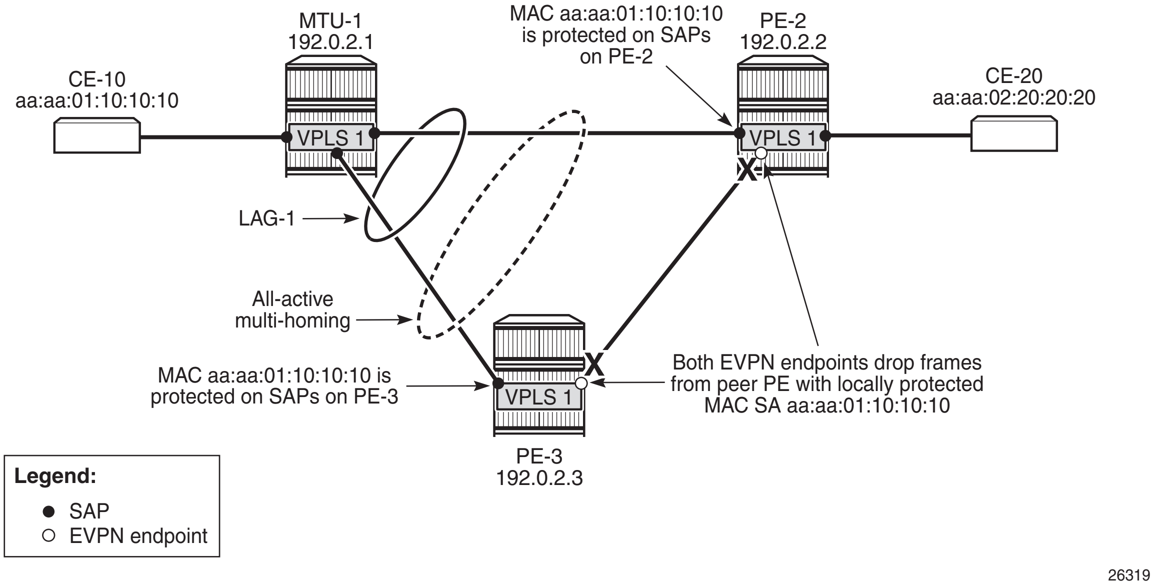
```
*A:PE-3# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         aa:aa:01:10:10:10 sap:lag-1:1              LP/0     05/11/17 15:15:59
1         aa:aa:02:20:20:20 eMpls:                   EvpnS    05/11/17 15:13:36
                                                     P
                            192.0.2.2:262140
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-3#
```

ALMP in All-Active Multi-Homing, RPS-DF on EVPN MPLS Destinations

ALMP is not recommended in all-active multi-homing because it can cause traffic loss. The following example shows when frames are dropped.

Figure 18 shows the example setup with MAC address aa:aa:01:10:10:10 protected on SAP lag-1:1 on both PE-2 and PE-3, and RPS-DF enabled on the EVPN endpoints.

Figure 18 **All-Active Multi-Homing - RPS-DF on SAPs and EVPN Endpoints**



When frames with MAC address aa:aa:01:10:10:10 are sent between PE-2 and PE-3, these frames will be dropped by the EVPN MPLS destination that has RPS-DF enabled.

The traffic flows from CE-10 and CE-20 are hashed over both links in the LAG. When the frames are sent out on MTU-1 on port 1/1/1 toward PE-2, the traffic reaches CE-20, and traffic can be sent back from CE-20 to CE-10 via the direct link between PE-2 and MTU-1. However, when traffic is sent out from MTU-1 on port 1/1/2 toward PE-3, the frames will be forwarded from PE-3 to PE-2, where they will be discarded at the EVPN MPLS destination on PE-2 because of RPS-DF. No traffic flow is possible for frames with the protected MAC SA aa:aa:01:10:10:10 via PE-3 to PE-2, or vice versa. If the MAC address is not protected yet on PE-2, the first few messages get through until the MAC address is protected on PE-2. Both multi-homing PEs, PE-2 and PE-3, protect the MAC address aa:aa:01:10:10:10 on their local all-active SAP. Therefore, PE-2 discards all frames with the MAC SA aa:aa:01:10:10:10 when they are received on the EVPN MPLS destination from the other multi-homing PE (PE-3).

An improved mechanism for EVPN loop protection in all-active multi-homing is black-hole MAC duplication, as described in chapter Black-hole MAC for EVPN Loop Protection.

For single-active multi-homing, this problem does not arise: only one designated forwarder in the Ethernet segment forwards traffic. Therefore, the CE MAC addresses will not be learned and protected on different PEs in the same Ethernet segment.

Conclusion

For security, MAC addresses learned on objects, such as SAPs, spoke/mesh-SDPs, and SHGs in EVPN services can be protected and advertised by BGP with the sticky bit set. By default, frames with a protected MAC SA are discarded if received on objects where the MAC address was not learned. Objects can be configured to be shut down when a frame is received with a protected MAC SA that has not been learned locally.

BGP Multi-Homing for VPLS Networks

This chapter describes BGP Multi-Homing (BGP-MH) for VPLS network configurations.

Topics in this chapter include:

- Applicability
- Summary
- Overview
- Configuration
- Conclusion

Applicability

Initially, the information in this chapter was based on SR OS Release 8.0.R5, with additions for SR OS Release 9.0.R1. The CLI in the current edition corresponds to SR OS Release 15.0.R2.

Summary

SR OS supports the use of Border Gateway Protocol Multi-Homing for VPLS (hereafter called BGP-MH). BGP-MH is described in *draft-ietf-l2vpn-vpls-multihoming, BGP based Multi-homing in Virtual Private LAN Service*, and provides a network-based resiliency mechanism (no interaction from the Provider Edge routers (PEs) to Multi-Tenant Units/Customer Equipment (MTU/CE)) that can be applied on service access points (SAPs) or network (pseudowires) topologies. The BGP-MH procedures will run between the PEs and will provide a loop-free topology from the network perspective (only one logical active path will be provided per VPLS among all the objects SAPs or pseudowires which are part of the same Multi-Homing site).

Each multi-homing site connected to two or more peers is represented by a site-id (2 bytes long) which is encoded in the BGP MH Network Layer Reachability Information (NLRI). The BGP peer holding the active path for a particular multi-homing site will be named as the Designated Forwarder (DF), whereas the rest of the BGP peers participating in the BGP MH process for that site will be named as non-DF and will block the traffic (in both directions) for all the objects belonging to that multi-homing site.

BGP MH uses the following rules to determine which PE is the DF for a particular multi-homing site:

1. A BGP MH NLRI with D flag = 0 (multi-homing object up) always takes precedence over a BGP MH NLRI with D flag = 1 (multi-homing object down). If there is a tie, then:
2. The BGP MH NLRI with the highest BGP Local Preference (LP) wins. If there is a tie, then:
3. The BGP MH NLRI issued from the PE with the lowest PE ID (system address) wins.

The main advantages of using BGP-MH as opposed to other resiliency mechanisms for VPLS are:

- Flexibility: BGP-MH uses a common mechanism for access and core resiliency. The designer has the flexibility of using BGP-MH to control the active/standby status of SAPs, spoke SDPs, Split Horizon Groups (SHGs) or even mesh SDP bindings.
- The standard protocol is based on BGP, a standard, scalable, and well-known protocol.
- Specific benefits at the access:
  - It is network-based, independent of the customer CE and, as such, it does not need any customer interaction to determine the active path. Consequently the operator will spend less effort on provisioning and will minimize both operation costs and security risks (in particular, this removes the requirement for spanning tree interaction between the PE and CE).
  - Easy load balancing per service (no service fate-sharing) on physical links.
- Specific benefits in the core:
  - It is a network-based mechanism, independent of the MTU resiliency capabilities and it does not need MTU interaction, therefore operational advantages are achieved as a result of the use of BGP-MH: less provisioning is required and there will be minimal risks of loops. In addition, simpler MTUs can be used.
  - Easy load balancing per service (no service fate-sharing) on physical links.

– Less control plane overhead: there is no need for an additional protocol running the pseudowire redundancy when BGP is already used in the core of the network. BGP-MH just adds a separate NLRI in the L2-VPN family (AFI=25, SAFI=65).

This chapter describes how to configure and troubleshoot BGP-MH for VPLS

Knowledge of the LDP/BGP VPLS (RFC 4762, *Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling*, and RFC 4761, *Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling*) architecture and functionality is assumed throughout this document. For further information, see the relevant Nokia documentation.

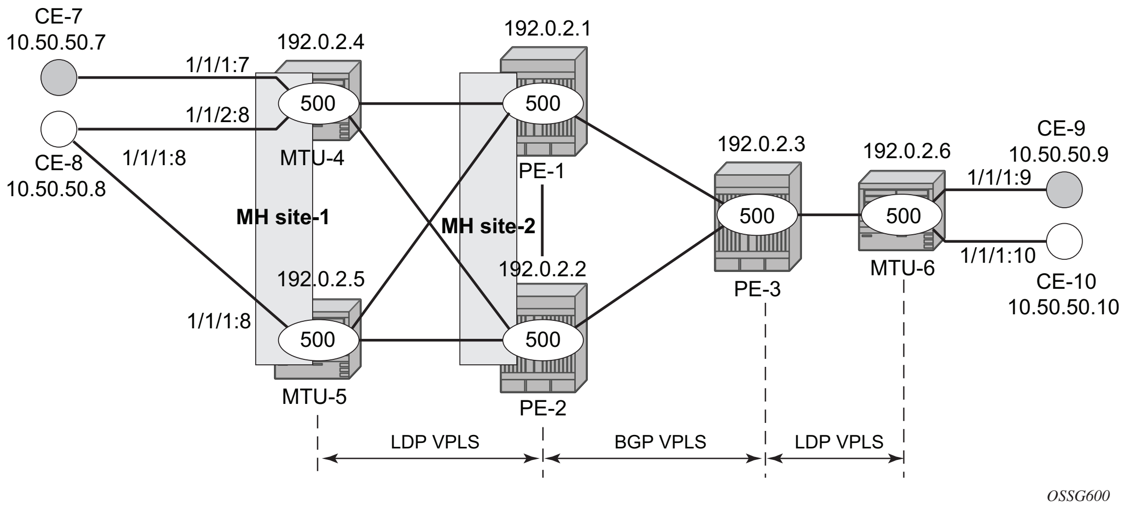
Overview

Figure 19 shows the example topology that will be used throughout the rest of the chapter.

The initial configuration includes:

- IGP — ISIS, Level 2 on all routers; area 49.0001
- RSVP-TE for transport tunnels
- FRR in the core
- No protection at the access.

Figure 19 **Example Topology**



The topology consists of three core nodes (PE-1, PE-2, and PE-3) and three Multi-Tenant Unit (MTU) nodes connected to the core. The VPLS service 500 is configured on all the six nodes with the following characteristics:

The VPLS service 500 is configured on all the six nodes with the following characteristics:

- The core VPLS instances are connected by a full mesh of BGP-signaled pseudowires (that is, pseudowires among PE-1, PE-2, and PE-3 will be signaled by BGP VPLS).
- As shown in Figure 19, the MTUs are connected to the BGP VPLS core by TLDP pseudowires. MTU-6 is connected to PE-3 by a single pseudowire, whereas MTU-4 and MTU-5 are dual-homed to PE-1 and PE-2. The following resiliency mechanisms are used on the dual-homed MTUs:
  - MTU-4 is dual-connected to PE-1 and PE-2 by an active/standby pseudowire (A/S pseudowire hereafter).
  - MTU-5 is dual-connected to PE-1 and PE-2 by two active pseudowires, one of them being blocked by BGP MH running between PE-1 and PE-2. The PE-1 and PE-2 pseudowires, set up from MTU-5, will be part of the BGP MH site MH-site-2.
  - MTU-4 and MTU-5 are running BGP MH, being SHG site-1 and SAP 1/1/1:8 on MTU-5 part of the same BGP MH site, MH-site-1.
- The CEs are connected to the network in the following way:
  - CE-7, CE-9 and CE-10 are single-connected to the network
  - CE-8 is dual connected to MTU-4 and MTU-5.
  - CE-7 and CE-8 are part of the split-horizon-group (SHG) site-1(SAPs 1/1/4:500 and 1/1/3:500 on MTU-4). Assume that CE-7 and CE-8 have a backdoor link between them so that when MTU-5 is elected as DF, CE1 does not get isolated. This configuration highlights the use of a SHG within a site configuration.

For each BGP MH site, MH-site-1 and MH-site-2, the BGP MH process will elect a DF, blocking the site objects for the non-DF nodes. In other words, based on the specific configuration explained throughout the chapter:

- For MH-site-1, MTU-4 will be elected as the DF. The non-DF-MTU-5 will block the SAP 1/1/1:8.
- For MH-site-2, PE-1 will be elected as the DF. The non-DF PE-1 will block the spoke-SDP to MTU-5.

Configuration

This section describes all the relevant configuration tasks for the setup shown in Figure 19. The appropriate associated IP/MPLS configuration is out of the scope of this chapter. In this example, the following protocols will be configured beforehand:

- ISIS-TE as IGP with all the interfaces being level-2 (OSPF-TE could have been used instead).
- RSVP-TE as the MPLS protocol to signal the transport tunnels (LDP could have been used instead).
- LSPs between core PEs will be Fast Re-Route protected (facility bypass tunnels) whereas LSP tunnels between MTUs and PEs will not be protected.

Note: The designer can choose whether to protect access link failures by means of MPLS FRR or A/S pseudowire or BGP MH. Whereas FRR provides a faster convergence (around 50ms) and stability (it does not impact on the service layer, therefore, link failures do not trigger MAC flush and flooding), some interim inefficiencies can be introduced compared to A/S pseudowire or BGP MH.

Once the IP/MPLS infrastructure is up and running, the specific service configuration including the support for BGP MH can begin.

Global BGP Configuration

BGP is used in this configuration guide for these purposes:

a. Auto-discovery and signaling of the pseudowires in the core, as per RFC 4761.
b. Exchange of multi-homing site NLRIs and redundancy handling from MTU-5 to the core.
c. Exchange of multi-homing site NLRIs and redundancy handling at the access for CE-7/CE-8.

A BGP route reflector (RR), PE-3, is used for the reflection of BGP updates corresponding to the preceding uses **a** and **b.**

A direct peering is established between MTU-4 and MTU-5 for use **c**. The same RR could have been used for the three cases, however, like in this example, the designer may choose to have a direct BGP peering between access devices. The reasons for this are:

- By having a direct BGP peering between MTU-4 and MTU-5, the BGP updates do not have to travel back and forth.
- On MTU-4 and MTU-5, BGP is exclusively used for multi-homing, therefore there will not be more BGP peers for either MTUs and a RR adds nothing in terms of control plane scalability.

On all nodes, the autonomous-system number must be configured.

```
configure router autonomous-system 65000
```

The following CLI output shows the global BGP configuration required on MTU-4. The 192.0.2.5 address will be replaced by the corresponding peer or the RR system address for PE-1 and PE-2.

```
*A:MTU-4# configure
    router
        bgp
            family l2-vpn
            rapid-withdrawal
            rapid-update l2-vpn
            group "Multi-Homing"
                neighbor 192.0.2.5
                    peer-as 65000
                exit
            exit
```

In this example, PE-3 is the BGP RR, therefore its BGP configuration will contain a cluster with all its peers included (PE-1 and PE-2):

```
*A:PE-3# configure
    router
        bgp
            family l2-vpn
            rapid-withdrawal
            rapid-update l2-vpn
            group "internal"
                cluster 1.1.1.1
                neighbor 192.0.2.1
                    peer-as 65000
                exit
                neighbor 192.0.2.2
                    peer-as 65000
                exit
            exit
```

The relevant BGP commands for BGP-MH are in bold. Some considerations about those:

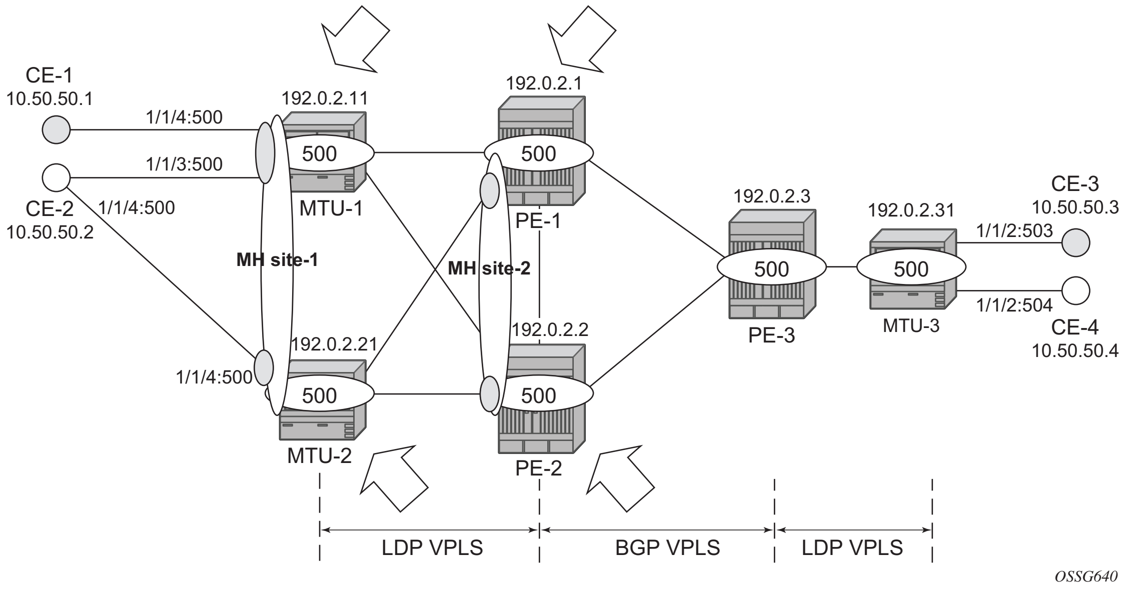
- It is required to specify **family l2-vpn** in the BGP configuration. That statement will allow the BGP peers to agree on the support for the family AFI=25 (Layer 2 VPN), SAFI=65 (VPLS). This family is used for BGP VPLS as well as for BGP MH and BGP AD.

- The **rapid-update l2-vpn** statement allows BGP MH to send BGP updates immediately after detecting link failures, without having to wait for the Minimum Route Advertisement Interval (MRAI) to send the updates in batches. This statement is required to guarantee a fast convergence for BGP MH.
- Optionally, **rapid-withdrawal** can also be added. In the context of BGP MH, this command is only useful if a particular multi-homing site is cleared. In that case, a BGP withdrawal is sent immediately without having to wait for the MRAI. A multi-homing site is cleared when the BGP MH site is removed or even the entire VPLS service.

Service Level Configuration

Once the IP/MPLS infrastructure is configured, including BGP, this section shows the configuration required at service level (VPLS 500). The focus is on the nodes involved on BGP MH, that is, MTU-4, MTU-5, PE-1, and PE-2. These nodes are highlighted in Figure 20.

Figure 20 **Nodes Involved in BGP MH**



Core PE Service Configuration

The following CLI excerpt shows the service level configuration on PE-1. The import/export policies configured on the PE nodes are identical:

```
configure
    router
        policy-options
            begin
            community "comm_core" members "target:65000:500"
            policy-statement "vsi500_export"
                entry 10
                    action accept
                        community add "comm_core"
                    exit
                exit
            exit
            policy-statement "vsi500_import"
                entry 10
                    from
                        community "comm_core"
                        family l2-vpn
                    exit
                    action accept
                    exit
                exit
                default-action drop
                exit
            exit
            commit
```

The configuration of the SDPs, PW template, and VPLS on PE-1 is as follows:

```
configure
    service
        sdp 12 mpls create
            description "SDP to transport BGP-signaled PWs"
            signaling bgp
            far-end 192.0.2.2
            lsp "LSP-PE-1-PE-2"
            path-mtu 8000
            no shutdown
        exit
        sdp 13 mpls create
            description "SDP to transport BGP-signaled PWs"
            signaling bgp
            far-end 192.0.2.3
            lsp "LSP-PE-1-PE-3"
            path-mtu 8000
            no shutdown
        exit
        sdp 14 mpls create
            far-end 192.0.2.4
            lsp "LSP-PE-1-MTU-4"
            path-mtu 8000
            no shutdown
        exit
        sdp 15 mpls create
            far-end 192.0.2.5
            lsp "LSP-PE-1-MTU-5"
            path-mtu 8000
            no shutdown
        exit
```

```
pw-template 500 use-provisioned-sdp create
exit
vpls 500 customer 1 create
    bgp
        route-distinguisher 65000:501
        vsi-export "vsi500_export"
        vsi-import "vsi500_import"
        pw-template-binding 500 split-horizon-group "CORE"
        exit
    exit
    bgp-vpls
        max-ve-id 65535
        ve-name 501
            ve-id 501
        exit
        no shutdown
    exit
    site "MH-site-2" create
        site-id 2
        spoke-sdp 15:500
        no shutdown
    exit
    spoke-sdp 14:500 create
    exit
    spoke-sdp 15:500 create
    exit
    no shutdown
exit
```

The following are general comments about the configuration of service 500:

- As seen in the preceding CLI output for PE-1, there are four provisioned SDPs that the service VPLS 500 will use in this example. SDP 14 and SDP 15 are tunnels over which the TLDP FEC128 pseudowires for service 500 will be carried (according to RFC 4762), whereas SDP 12 and SDP 13 are the tunnels for the core BGP pseudowires (based on RFC 4761).
- The BGP context provides the general service BGP configuration that will be used by BGP VPLS and BGP MH:
    - Route distinguisher (notation chosen is based on <AS_number:500 + node_id>)
    - VSI export policies are used to add the export route-targets included in all the BGP updates sent to the BGP peers.
    - VSI import policies are used to control the NLRIs accepted in the RIB, normally based on the route-targets.
    - Both VSI-export and VSI-import policies can be used to modify attributes such as the Local-Preference (LP) that will be used to influence the BGP MH Designated Forwarder (DF) election (LP is the second rule in the BGP MH election process, as previously discussed). The use of these policies will be described later in the chapter.

- The **pw-template-binding** command maps the previously defined pw-template 500 to the **split-horizon-group "CORE"**. In this way, all the BGP-signaled pseudowires will be part of this split horizon group. Although not shown in this example, the **pw-template-binding** command can also be used to instantiate pseudowires within different split horizon groups, based on different import route targets:

Note: Detailed BGP-VPLS configuration is out of the scope of this chapter. For more information, see chapter *BGP-VPLS*.

```
*A:PE-1# configure service vpls 500 bgp pw-template-binding ?
  - pw-template-binding <policy-id> [split-horizon-group <group-name>]
                                    [import-rt {ext-community,...(upto 5 max)}]
  - no pw-template-binding <policy-id>

---snip---
```

- The BGP-signaled pseudowires (from PE-1 to PE-2 and PE-3) are set up according to the configuration in the BGP context. Beside those pseudowires, the VPLS 500 also has two more pseudowires signaled by TLDP: spoke-sdp 14:500 (to MTU-4) and spoke-sdp 15:500 (to MTU-5).

The general BGP MH configuration parameters for a particular multi-homing site are shown in the following output:

```
*A:PE-1# configure service vpls ?
  - no vpls <service-id>
  - vpls <service-id> [customer <customer-id>] [create] [vpn <vpn-id>] [m-vpls] [b-
vpls|i-vpls] [etree]

---snip---


*A:PE-1# configure service vpls 500 site ?
  - no site <name>
  - site <name> [create]

 <name>                : [32 chars max]

 [no] boot-timer       - Configure/Override site boot-timer
      failed-thresho* - Configure threshold for the site to be declared down
 [no] mesh-sdp-bindi* - Enable/Disable application to all Mesh-SDP
 [no] monitor-oper-g* - Configure an Operational-Group to monitor
 [no] sap              - Configure a SAP for the site
 [no] shutdown         - Administratively enable/disable the site
 [no] site-activatio* - Configure/Override site activation timer
 [no] site-id          - Configure site identifier
 [no] site-min-down-* - Configure minimum down timer for the site
 [no] split-horizon-* - Configure a split-horizon-group
 [no] spoke-sdp        - Configure a spoke-SDP
```

Where:

- The site **name** is defined by a string of up to 32 characters.
- The **site-id** is an integer that identifies the multi-homing site and is encoded in the BGP MH NLRI. This ID must be the same one used on the peer node where the same multi-homing site is connected to. That is, MH-site-2 must use the same site-id in PE-1 and PE-2 (value = 2 in the PE-1 site configuration).
- Out of the four potential objects in a site—spoke SDP, SAP, SHG, and mesh SDP binding—only one can be used at the time on a particular site. To add more than just one SAP/spoke-SDP to the same site, a split horizon group composed of the SAP/spoke-SDP objects must be used in the site configuration. Otherwise, only one object—spoke SDP, SAP, SHG, or mesh SDP binding—is allowed per site. A CLI log message warns the operator of such fact:

```
*A:PE-1>config>service>vpls>site# mesh-sdp-binding
MINOR: SVCMGR #5855 only one object is allowed per site
```

- The **failed-threshold** command defines how many objects should be down for the site to be declared down. This command is obviously only valid for multi-object sites (split horizon groups and mesh-SDP bindings). By default, all the objects in a site must be down for the site to be declared as operationally down.

```
*A:PE-1>config>service>vpls>site# failed-threshold ?
  - failed-threshold <[1..1000]>
  - failed-threshold all
```

- The **boot-timer** specifies for how long the service manager waits after a node reboot before running the MH Procedures. The boot-timer value should be configured to allow for the BGP sessions to come up and for the NLRI information to be refreshed/exchanged. In environments with the default BGP MRAI (30 seconds), it is highly recommended to increase this value (for instance, 120 seconds for a normal configuration). The **boot-timer** is only important when a node comes back up and would become the DF. Default value: 10 seconds.

```
*A:PE-1>config>service>vpls>site# boot-timer ?
  - boot-timer <seconds>
  - no boot-timer

 <seconds>            : [0..600]
```

- The **site-activation-timer** command defines the amount of time the service manager will keep the local objects in standby (in the absence of BGP updates from remote PEs) before running the DF election algorithm to decide whether the site should be unblocked. The timer is started when one of the following events occurs only if the site is operationally up:

  - Manual site activation using the **no shutdown** command at the site-id level or at member object(s) level (SAP(s) or pseudowire(s))
  - Site activation after a failure
  - The BGP MH election procedures will be resumed upon expiration of this timer or the arrival of a BGP MH update for the multi-homing site. Default value: 2 seconds.

```
*A:PE-1>config>service>vpls>site# site-activation-timer
  - no site-activation-timer
  - site-activation-timer <seconds>
 <seconds>           : [0..100]
```

- When a BGP MH site goes down, it may be preferred that it stays down for a minimum time. This is configurable by the **site-min-down-timer**. When set to zero, this timer is disabled.

```
*A:PE-1>config>service>vpls>site# site-min-down-timer
  - no site-min-down-timer
  - site-min-down-timer <seconds>

 <seconds>           : [0..100]
```

- The **boot-timer**, **site-activation-timer** and **site-min-down-timer** commands can be provisioned at service level or at global level. The service level settings have precedence and override the global configuration. The **no** form of the commands at global level, sets the value back to the default values. The **no** form of the commands at service level, makes the timers inherit the global values.

```
*A:PE-1# configure redundancy bgp-multi-homing
  - bgp-multi-homing

 [no] boot-timer      - Configure BGP multi-homing boot-timer
 [no] site-activatio* - Configure BGP multi-homing site activation timer
 [no] site-min-down-* - Configure minimum down timer for the site
```

- The **shutdown** command controls the admin state of the site. Each site has three possible states:
  - Admin state — controlled by the shutdown command.
  - Operational state — controlled by the operational status of the individual site objects.
  - Designated-Forwarder (DF) state — controlled by the BGP MH election algorithm.

The following CLI output shows the three states for BGP MH site “MH-site-1” on MTU-5:

```
*A:MTU-5# show service id 500 site "MH-site-1"
```

```
===============================================================================
Site Information
===============================================================================
Site Name            : MH-site-1
-------------------------------------------------------------------------------
Site Id              : 1
Dest                 : sap:1/1/1:8          Mesh-SDP Bind     : no
Admin Status         : Enabled              Oper Status       : up
Designated Fwdr      : No
DF UpTime            : 0d 00:00:00          DF Chg Cnt        : 1
Boot Timer           : default              Timer Remaining   : 0d 00:00:00
Site Activation Timer: default              Timer Remaining   : 0d 00:00:00
Min Down Timer       : default              Timer Remaining   : 0d 00:00:00
Failed Threshold     : default(all)
Monitor Oper Grp     : (none)
===============================================================================
*A:MTU-5#
```

For this example, MH-site “ MH-site-2” is configured in PE-1, where the site-id is 2 and the object in the site is spoke-SDP 15:500 (pseudowire established from PE-1 to MTU-5).

The following CLI shows the service configuration for PE-2. The site-id is 2, that is, the same value configured in PE-1. The object defined in PE-2’s site is spoke-SDP 25:500 (pseudowire established from PE-2 to MTU-5).

```
configure
    service
        sdp 21 mpls create
            description "SDP to transport BGP-signaled PWs"
            signaling bgp
            far-end 192.0.2.1
            lsp "LSP-PE-2-PE-1"
            path-mtu 8000
            no shutdown
        exit
        sdp 23 mpls create
            description "SDP to transport BGP-signaled PWs"
            signaling bgp
            far-end 192.0.2.3
            lsp "LSP-PE-2-PE-3"
            path-mtu 8000
            no shutdown
        exit
        sdp 24 mpls create
            far-end 192.0.2.4
            lsp "LSP-PE-2-MTU-4"
            path-mtu 8000
            no shutdown
        exit
        sdp 25 mpls create
            far-end 192.0.2.5
            lsp "LSP-PE-2-MTU-5"
            path-mtu 8000
            no shutdown
        exit
        pw-template 500 use-provisioned-sdp create
        exit
        vpls 500 customer 1 create
            bgp
                route-distinguisher 65000:502
                vsi-export "vsi500_export"
                vsi-import "vsi500_import"
                pw-template-binding 500 split-horizon-group "CORE"
                exit
            exit
            bgp-vpls
                max-ve-id 65535
                ve-name 502
                    ve-id 502
                exit
                no shutdown
            exit
            site "MH-site-2" create
                site-id 2
                spoke-sdp 25:500
                no shutdown
            exit
            spoke-sdp 24:500 create
            exit
            spoke-sdp 25:500 create
            exit
            no shutdown
        exit
```

MTU Service Configuration

The following CLI output shows the service level configuration on MTU-4.

```
configure
    service
        sdp 41 mpls create
            far-end 192.0.2.1
            lsp "LSP-MTU-4-PE-1"
            path-mtu 8000
            no shutdown
        exit
        sdp 42 mpls create
            far-end 192.0.2.2
            lsp "LSP-MTU-4-PE-2"
            path-mtu 8000
            no shutdown
        exit
        vpls 500 customer 1 create
            split-horizon-group "site-1" create
            exit
            bgp
                route-distinguisher 65000:504
                route-target export target:65000:500 import target:65000:500
            exit
            site "MH-site-1" create
```

```
                site-id 1
                split-horizon-group site-1
                no shutdown
            exit
            endpoint "CORE" create
                no suppress-standby-signaling
            exit
            sap 1/1/1:7 split-horizon-group "site-1" create
            exit
            sap 1/1/2:8 split-horizon-group "site-1" create
                eth-cfm
                    mep 48 domain 1 association 1 direction down
                        fault-propagation-enable use-if-tlv
                        ccm-enable
                        no shutdown
                    exit
                exit
            exit
            spoke-sdp 41:500 endpoint "CORE" create
                precedence primary
            exit
            spoke-sdp 42:500 endpoint "CORE" create
            exit
            no shutdown
        exit
```

MTU-4 is configured with the following characteristics:

- The BGP context provides the general BGP parameters for service 500 in MTU-4. The **route-target** command is now used instead of the vsi-import and vsi-export commands. The intent in this example is to configure only the export and import route-targets. There is no need to modify any other attribute. If the local preference is to be modified (to influence the DF election), a **vsi-policy** must be configured.
- An A/S pseudowire configuration is used to control the pseudowire redundancy towards the core.
- The multi-homing site, MH-site-1 has a site-id = 1 and a split horizon group as an object. The split horizon group site-1 is composed of sap 1/1/1:7 and sap 1/1/2:8. As previously discussed, the site will not be declared operationally down until the two SAPs belonging to the site are down. This behavior can be changed by the **failed-threshold** command (for instance, in order to bring the site down when only one object has failed even though the second SAP is still up).
- As an example, a Y.1731 MEP with fault-propagation has been defined in SAP 1/1/2:8. As discussed later in the chapter, this MEP will signal the status of the SAP (as a result of the BGP MH process) to CE-8.

The service configuration in MTU-5 is as follows:

```
configure
    service
```

```
sdp 51 mpls create
    far-end 192.0.2.1
    lsp "LSP-MTU-5-PE-1"
    path-mtu 8000
    no shutdown
exit
sdp 52 mpls create
    far-end 192.0.2.2
    lsp "LSP-MTU-5-PE-2"
    path-mtu 8000
    no shutdown
exit
vpls 500 customer 1 create
    bgp
        route-distinguisher 65000:505
        route-target export target:65000:500 import target:65000:500
    exit
    site "MH-site-1" create
        site-id 1
        sap 1/1/1:8
        no shutdown
    exit
    sap 1/1/1:8 create
    exit
    spoke-sdp 51:500 create
    exit
    spoke-sdp 52:500 create
    exit
    no shutdown
exit
```

Influencing the Designated Forwarder (DF) Decision

As previously explained, assuming that the sites on the two nodes taking part of the same multi-homing site are both up, the two tie-breakers for electing the DF are (in this order):

1. Highest LP
2. Lowest PE ID

The LP by default is 100 in all the routers. Under normal circumstances, if the LP in any router is not changed, MTU-4 will be elected the DF for MH-site-1, whereas PE-1 will be the DF for MH-site-2. Assume in this section that this behavior is changed for MH-site-2 to make PE-2 the DF. Since changing the system address (to make PE-2's ID the lower of the two IDs) is usually not an easy task to accomplish, the vsi-export policy on PE-2 is modified with an LP of 150 with which the MH-site-2 NLRI is announced to PE-1. Because LP 150 is greater than the default 100 in PE-1, PE-2 will be elected as the DF for MH-site-2. The vsi-import policy remains unchanged and the vsi-export policy is modified as follows:

```
configure
    router
        policy-options
            begin
            community "comm_core" members "target:65000:500"
            policy-statement "vsi500_export"
                entry 10
                    action accept
                        community add "comm_core"
                        local-preference 150
                    exit
                exit
            exit
            policy-statement "vsi500_import"
                entry 10
                    from
                        community "comm_core"
                        family l2-vpn
                    exit
                    action accept
                    exit
                exit
                default-action reject
            exit
            commit
```

In PE-1, the import and export policies are not modified. The policies were already applied in the BGP context of VPLS 500, as follows:

```
*A:PE-2# configure
    service
        vpls 500 customer 1 create
            bgp
                route-distinguisher 65000:502
                vsi-export "vsi500_export"
                vsi-import "vsi500_import"
                pw-template-binding 500 split-horizon-group "CORE"
                exit
            exit
---snip---
```

The DF state of PE-2 can be verified as follows:

```
*A:PE-2# show service id 500 site "MH-site-2"

===============================================================================
Site Information
===============================================================================
Site Name            : MH-site-2
-------------------------------------------------------------------------------
Site Id              : 2
Dest                 : sdp:25:500           Mesh-SDP Bind     : no
Admin Status         : Enabled              Oper Status       : up
Designated Fwdr      : Yes
DF UpTime            : 0d 00:12:29          DF Chg Cnt        : 2
Boot Timer           : default              Timer Remaining   : 0d 00:00:00
Site Activation Timer: default              Timer Remaining   : 0d 00:00:00
```

```
Min Down Timer        : default            Timer Remaining  : 0d 00:00:00
Failed Threshold      : default(all)
Monitor Oper Grp      : (none)
===============================================================================
*A:PE-2#
```

The import and export policies are applied at service 500 level, which means that the LP changes for all the potential multi-homing sites configured under service 500. Therefore, load balancing can be achieved on a per-service basis, but not within the same service.

These policies are applied on the VPLS 500 for all the potential BGP applications: BGP VPLS, BGP MH, and BGP AD. In the example, the LP for the PE-2 BGP updates for BGP MH and BGP VPLS will be set to 150. However, this has no impact on BGP VPLS because a PE cannot receive two BGP VPLS NLRIs with the same VE-ID, which implies that a different VE-ID per PE within the same VPLS is required.

The vsi-export policy is restored to its original settings on PE-2, as follows:

```
configure
    router
        policy-options
            begin
            policy-statement "vsi500_export"
                entry 10
                    action accept
                        community add "comm_core"
                        no local-preference
                    exit
                exit
            exit
            commit
```

In all the PE nodes, the import and export policies applied in the BGP context of VPLS 500 have identical settings again, and PE-1 is the DF.

Black-Hole Avoidance

SR OS supports the appropriate MAC flush mechanisms for BGP MH, regardless of the protocol being used for the pseudowire signaling:

- LDP VPLS — The PE that contains the old DF site (the site that just experienced a DF to non-DF transition) always sends a LDP MAC **flush-all-from-me** to all LDP pseudowires in the VPLS, including the LDP pseudowires associated with the new DF site. No specific configuration is required.

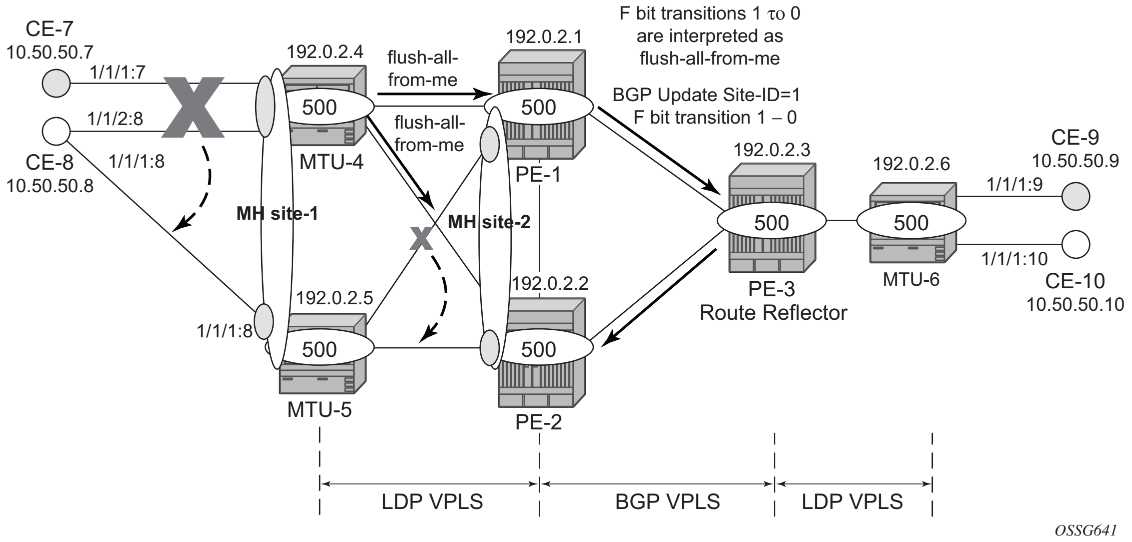
- BGP VPLS — The remote BGP VPLS PEs interpret the F bit transitions from 1 to 0 as an implicit MAC flush-all-from-me indication. If a BGP update with the flag F=0 is received from the previous DF PE, the remote PEs perform MAC flush-all-from-me, flushing all the MACs associated with the pseudowire to the old DF PE. No specific configuration is required.

Double flushing will not happen because it is expected that between any pair of PEs there will exist only one type of pseudowires—either BGP or LDP pseudowire—, but not both types.

In the example, assuming MTU-4 and PE-1 are the DF nodes:

- When MH-site-1 is brought operationally down on MTU-4 (so by default, the two SAPs must go down unless the **failed-threshold** parameter is changed so that the site is down when only one SAP is brought down), MTU-4 will issue a **flush-all-from-me** message.
- When MH-site-2 is brought operationally down on PE-1, a BGP update with F=0 and D=1 is issued by PE-1. PE-2 and PE-3 will receive the update and will flush the MAC addresses learned on the pseudowire to PE-1.

Figure 21 **MAC Flush for BGP MH**



Node failures implicitly trigger a MAC flush on the remote nodes, because the TLDP/BGP session to the failed node goes down.

Access CE/PE Signaling

BGP MH works at service level, therefore no physical ports are torn down on the non-DF, but rather the objects are brought down operationally, while the physical port will stay up and used for any other services existing on that port. Due to this reason, there is a need for signaling the standby status of an object to the remote PE or CE.

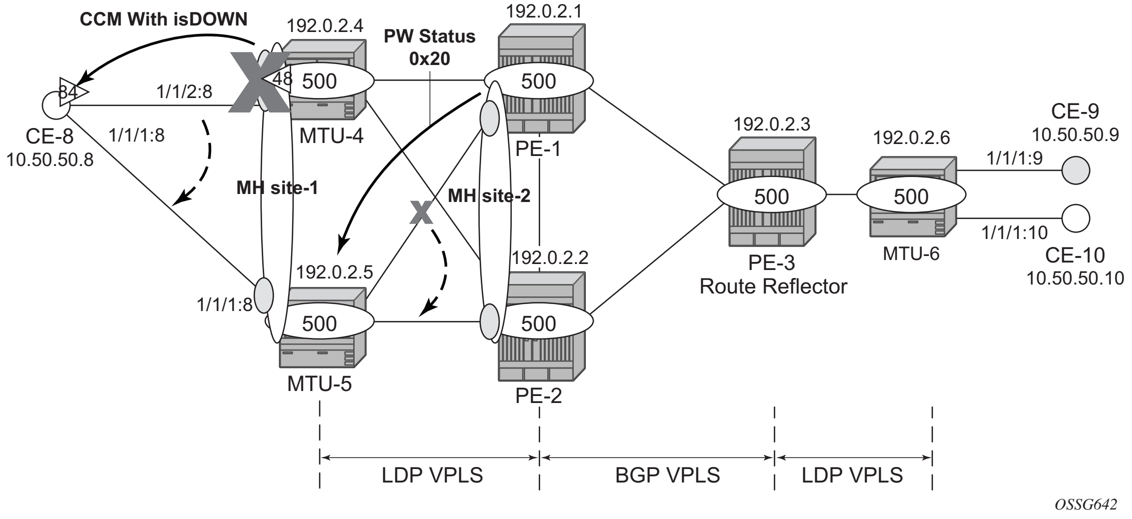
- Access PEs running BGP MH on spoke SDPs and elected non-DF, will signal pseudowire standby status (0x20) to the other end. If no pseudowire status is supported on the remote MTU, a label withdrawal is performed. If there is more than one spoke SDP on the site (part of the same SHG), the signaling is sent for all the pseudowires of the site.

Note: The **configure service vpls x spoke-sdp y:z no pw-status-signaling** parameter allows to send a TLDP label-withdrawal instead of pseudowire status bits, even though the peer supports pseudowire status.

- Multi-homed CEs connected through SAPs to the PEs running BGP MH, are signaled by the PEs using Y.1731 CFM, either by stopping the transmission of CCMs or by sending CCMs with isDown (interface status down encoding in the interface status TLV).

In this example, down MEPs on MTU-4 SAP 1/1/2:8 and CE-8 SAP 1/1/2:8 are configured. In a similar way, other MEPs can be configured on MTU-4 SAP 1/1/1:7, MTU-5 SAP 1/1/1:8, and CE-8 SAP 1/1/1:7 and SAP 1/1/1:8. Figure 22 shows the MEPs on MTU-4 SAP 1/1/2:8 and CE-8. Upon failure on the MTU-4 site MH-site-1, the MEP 48 will start sending CCMs with interface status down.

Figure 22 **Access PE/CE Signaling**



The CFM configuration required at SAP 1/1/2:8 is as follows. Down MEPs will be configured on CE-8 and MTU-5 SAPs in the same way, but in a different association. The option **fault-propagation-enable use-if-tlv** must be added. In case the CE does not understand the CCM interface status TLV, the **fault-propagation-enable suspend-ccm** option can be enabled instead. This will stop the transmission of CCMs upon site failures. Detailed configuration guidelines for Y.1731 are beyond the scope of this chapter.

```
*A:MTU-4# configure
    eth-cfm
        domain 1 format none level 3
            association 1 format icc-based name "Association48"
                bridge-identifier 500
                exit
                ccm-interval 1
                remote-mepid 84
            exit
        exit


*A:MTU-4# configure
    service
        vpls 500 customer 1 create
            sap 1/1/2:8 split-horizon-group "site-1" create
                eth-cfm
                    mep 48 domain 1 association 1 direction down
                        fault-propagation-enable use-if-tlv
                        ccm-enable
                        no shutdown
                    exit
                exit
            exit
```

If CE-8 is a service router, upon receiving a CCM with isDown, an alarm will be triggered and the SAP will be brought down:

```
61 2017/04/26 06:58:30.32 UTC MINOR: ETH_CFM #2001 Base
"MEP 1/1/84 highest defect is now defRemoteCCM"

62 2017/04/26 06:58:30.32 UTC MINOR: SVCMGR #2108 vprn8
"Status of interface int-CE-8-MTU-4 in service 8 (customer 1) changed to admin=up
oper=down"

63 2017/04/26 06:58:30.32 UTC MINOR: SVCMGR #2203 vprn8
"Status of SAP 1/1/2:8 in service 8 (customer 1) changed to admin=up oper=down
flags=OamDownMEPFault "

64 2017/04/26 06:58:30.32 UTC WARNING: SNMP #2004 vprn8 int-CE-8-MTU-4
"Interface int-CE-8-MTU-4 is not operational"
```

On CE-8, the status of the SAP can be verified as follows:

```
*A:CE-8# show service id 8 sap 1/1/2:8

===============================================================================
Service Access Points(SAP)
===============================================================================
Service Id         : 8
SAP                : 1/1/2:8                  Encap             : q-tag
Description        : (Not Specified)
Admin State        : Up                       Oper State        : Down
Flags              : OamDownMEPFault
Multi Svc Site     : None
Last Status Change : 04/26/2017 06:58:30
Last Mgmt Change   : 04/25/2017 11:37:26
===============================================================================
*A:CE-8#
```

As also depicted in Figure 22, PE-1 will signal pseudowire status standby (code 0x20) when PE-1 goes to non-DF state for MH-site-2. MTU-5 will receive that signaling and, based on the **ignore-standby-signaling** parameter, will decide whether to send the broadcast, unknown unicast, and multicast (BUM) traffic to PE-1. In case MTU-5 uses in its configuration **ignore-standby-signaling**, it will be sending BUM traffic on both pseudowires at the same time (which is not normally desired), ignoring the pseudowire status bits. The following output shows the MTU-5 spoke-SDP receiving the pseudowire status signaling. Although the spoke SDP stays operationally up, the peer Pw Bits field shows **pwFwdingStandby** and MTU-5 will not send any traffic if the **ignore-standby-signaling** parameter is disabled.

```
A:MTU-5# show service id 500 sdp 51:500 detail

===============================================================================
Service Destination Point (Sdp Id : 51:500) Details
===============================================================================
-------------------------------------------------------------------------------
 Sdp Id 51:500  -(192.0.2.1)
-------------------------------------------------------------------------------
```

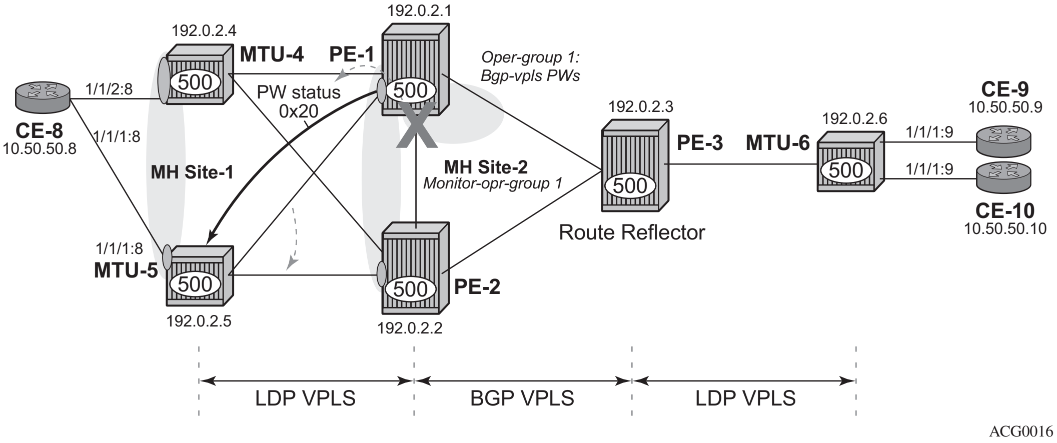
```
Description      : (Not Specified)
SDP Id             : 51:500                   Type              : Spoke
Spoke Descr      : (Not Specified)
Split Horiz Grp    : (Not Specified)
Etree Root Leaf Tag: Disabled                  Etree Leaf AC      : Disabled
VC Type            : Ether                     VC Tag             : n/a
Admin Path MTU     : 8000                      Oper Path MTU      : 8000
Delivery           : MPLS
Far End            : 192.0.2.1
Tunnel Far End     : n/a                       LSP Types          : RSVP
Hash Label         : Disabled                  Hash Lbl Sig Cap   : Disabled
Oper Hash Label    : Disabled
Entropy Label      : Disabled

Admin State        : Up                        Oper State         : Up

---snip---

Endpoint           : N/A                       Precedence         : 4
PW Status Sig      : Enabled
Force Vlan-Vc      : Disabled                  Force Qinq-Vc      : Disabled
Class Fwding State : Down
Flags              : None
Time to RetryReset : never                     Retries Left       : 3
Mac Move           : Blockable                 Blockable Level    : Tertiary
Local Pw Bits      : None
Peer Pw Bits       : pwFwdingStandby

---snip---
```

Operational Groups for BGP-MH

Operational groups (**oper-group**) introduce the capability of grouping objects into a generic group object and associating its status to other service endpoints (pseudowires, SAPs, IP interfaces) located in the same or in different service instances. The operational group status is derived from the status of the individual components using certain rules specific to the application using the concept. A number of other service entities—the monitoring objects—can be configured to monitor the operational group status and to drive their own status based on the **oper-group** status. In other words, if the operational group goes down, the monitoring objects will be brought down. When one of the objects included in the operational group comes up, the entire group will also come up, and therefore so will the monitoring objects.

This concept can be used to enhance the BGP-MH solution for avoiding black-holes on the PE selected as the Designated Forwarder (DF), if the rest of the VPLS endpoints fail (pseudowire spoke(s)/pseudowire mesh and/or SAP(s)). Figure 23 illustrates the use of operational groups together with BGP-MH. On PE-1 (and PE-2) all of the BGP-VPLS pseudowires in the core are configured under the same **oper-group** *group-1*. MH-site-2 is configured as a monitoring object. When the two BGP-VPLS pseudowires go down, **oper-group** *group-1* will be brought down, therefore MH-site-2 on PE-1 will go down as well (PE-2 will become DF and PE-1 will signal standby to MTU-5).

Figure 23 **Oper-Groups and BGP-MH**



In the preceding example, this feature provides a solution to avoid a black-hole when PE-1 loses its connectivity to the core.

Operational groups are configured in two steps:

1. Identify a set of objects whose forwarding state should be considered as a whole group, then group them under an operational group (in this case **oper-group** *group-1*, which is configured in the **bgp pw-template-binding** context).
2. Associate other existing objects (clients) with the oper-group using the **monitor-group** command (configured, in this case, in the **site** *MH-site-2*).

The following CLI excerpt shows the commands required (**oper-group**, **monitor-oper-group**).

```
*A:PE-1# configure
    service
        oper-group "group-1" create
        exit
        vpls 500
            bgp
                pw-template-binding 500 split-horizon-group "CORE"
                    oper-group "group-1"
                exit
            exit
            site "MH-site-2"
                monitor-oper-group "group-1"
            exit
```

When all the BGP-VPLS pseudowires go down, **oper-group** *group-1* will go down and therefore the monitoring object, **site** *MH-site-2*, will also go down and PE-2 will then be elected as DF. The log 99 gives information about this sequence of events:

```
*A:PE-1# configure service sdp 12 shutdown
*A:PE-1# configure service sdp 13 shutdown


*A:PE-1# show log log-id 99
---snip---
207 2017/04/26 09:20:52.74 UTC WARNING: SVCMGR #2531 Base BGP-MH
"Service-id 500 site MH-site-2 is not the designated-forwarder"

206 2017/04/26 09:20:52.74 UTC MAJOR: SVCMGR #2316 Base
"Processing of a SDP state change event is finished and the status of all affected
SDP Bindings on SDP 13 has been updated."

205 2017/04/26 09:20:52.74 UTC MINOR: SVCMGR #2306 Base
"Status of SDP Bind 15:500 in service 500 (customer 1) changed to admin=up oper=down
 flags="

204 2017/04/26 09:20:52.74 UTC MINOR: SVCMGR #2326 Base
"Status of SDP Bind 15:500 in service 500 (customer 1) local PW status bits changed
to pwFwdingStandby "

203 2017/04/26 09:20:52.74 UTC MINOR: SVCMGR #2542 Base
"Oper-group group-1 changed status to down"
```

PE-1 is no longer the DF, as follows:

```
*A:PE-1# show service id 500 site

===============================================================================
VPLS Sites
===============================================================================
Site                Site-Id   Dest                Mesh-SDP  Admin   Oper  Fwdr
-------------------------------------------------------------------------------
MH-site-2           2         sdp:15:500          no        Enabled down  No
-------------------------------------------------------------------------------
Number of Sites : 1
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

PE-2 becomes the DF.

```
*A:PE-2# show service id 500 site

===============================================================================
VPLS Sites
```

```
===============================================================================
Site                Site-Id   Dest                Mesh-SDP  Admin   Oper  Fwdr
-------------------------------------------------------------------------------
MH-site-2           2         sdp:25:500          no        Enabled up    Yes
-------------------------------------------------------------------------------
Number of Sites : 1
-------------------------------------------------------------------------------
===============================================================================
*A:PE-2#
```

The process reverts when at least one BGP-VPLS pseudowire comes back up.

Show Commands and Debugging Options

The main command to find out the status of a site is the **show service id x site** command.

```
*A:MTU-5# show service id 500 site

===============================================================================
VPLS Sites
===============================================================================
Site                Site-Id   Dest                Mesh-SDP  Admin   Oper  Fwdr
-------------------------------------------------------------------------------
MH-site-1           1         sap:1/1/1:8         no        Enabled up    No
-------------------------------------------------------------------------------
Number of Sites : 1
-------------------------------------------------------------------------------
===============================================================================
*A:MTU-5#
```

A **detail** modifier is available:

```
*A:MTU-5# show service id 500 site detail

===============================================================================
Site Information
===============================================================================
Site Name            : MH-site-1
-------------------------------------------------------------------------------
Site Id              : 1
Dest                 : sap:1/1/1:8          Mesh-SDP Bind    : no
Admin Status         : Enabled              Oper Status      : up
Designated Fwdr      : No
DF UpTime            : 0d 00:00:00          DF Chg Cnt       : 1
Boot Timer           : default              Timer Remaining  : 0d 00:00:00
Site Activation Timer: default              Timer Remaining  : 0d 00:00:00
Min Down Timer       : default              Timer Remaining  : 0d 00:00:00
Failed Threshold     : default(all)
Monitor Oper Grp     : (none)
-------------------------------------------------------------------------------
Number of Sites : 1
===============================================================================
*A:MTU-5#
```

The **detail** view of the command displays information about the BGP MH timers. The values are only shown if the global values are overridden by specific ones at service level (and will be tagged with **Ovr** if they have been configured at service level). The **Timer Remaining** field reflects the count down from the boot/site activation timers down to the moment when this router tries to become DF again. Again, this is only shown when the global timers have been overridden by the ones at service level.

The objects on the non-DF site will be brought down operationally and flagged with **StandByForMHProtocol**, for example, for SAP 1/1/1:8 on non-DF MTU-5:

```
*A:MTU-5# show service id 500 sap 1/1/1:8

===============================================================================
Service Access Points(SAP)
===============================================================================
Service Id         : 500
SAP                : 1/1/1:8                  Encap             : q-tag
Description        : (Not Specified)
Admin State        : Up                       Oper State        : Down
Flags              : StandByForMHProtocol
Multi Svc Site     : None
Last Status Change : 04/26/2017 08:21:49
Last Mgmt Change   : 04/25/2017 11:38:29
===============================================================================
*A:MTU-5#
```

For spoke SDP 25:500 on non-DF PE-2:

```
*A:PE-2# show service id 500 sdp 25:500 detail

===============================================================================
Service Destination Point (Sdp Id : 25:500) Details
===============================================================================
-------------------------------------------------------------------------------
 Sdp Id 25:500  -(192.0.2.5)
-------------------------------------------------------------------------------
Description     : (Not Specified)
SDP Id             : 25:500                   Type              : Spoke
---snip---

Admin State        : Up                       Oper State        : Down
---snip---
Flags              : StandbyForMHProtocol
---snip---
```

The BGP MH routes in the RIB, RIB-In and RIB-Out can be shown by using the corresponding **show router bgp routes** and **show router bgp neighbor x.x.x.x received-routes|advertised-routes** commands. The BGP MH routes are only shown when the operator uses the **l2-vpn** family modifier. Should the operator want to filter only the BGP MH routes out of the l2-vpn routes, the **multi-homing** filter has to be added to the **show router bgp routes** commands.

```
*A:PE-3# show router bgp routes l2-vpn
===============================================================================
 BGP Router ID:192.0.2.3        AS:65000       Local AS:65000
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP L2VPN Routes
===============================================================================
Flag  RouteType                   Prefix                                 MED
      RD                          SiteId                                 Label
      Nexthop                     VeId                   BlockSize   LocalPref
      As-Path                     BaseOffset             vplsLabelBa
                                                         se
-------------------------------------------------------------------------------
u*>i  VPLS                        -                      -           0
      65000:501                   -                                  -
      192.0.2.1                   501                    8           100
      No As-Path                  497                    262127
u*>i  MultiHome                   -                      -           0
      65000:501                   2                                  -
      192.0.2.1                   -                      -           100
      No As-Path                  -                      -
u*>i  VPLS                        -                      -           0
      65000:502                   -                                  -
      192.0.2.2                   502                    8           100
      No As-Path                  497                    262127
u*>i  MultiHome                   -                      -           0
      65000:502                   2                                  -
      192.0.2.2                   -                      -           100
      No As-Path                  -                      -
-------------------------------------------------------------------------------
Routes : 4
===============================================================================
*A:PE-3#
```

For the L2 VPN BGP routes toward site 2 (PE-1 and PE-2) in detail:

```
*A:PE-3# show router bgp routes l2-vpn multi-homing siteid 2 detail
===============================================================================
 BGP Router ID:192.0.2.3        AS:65000       Local AS:65000
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP L2VPN-MULTIHOME Routes
===============================================================================
Original Attributes

Route Type     : MultiHome
Route Dist.    : 65000:501
Site Id        : 2
```

```
Nexthop        : 192.0.2.1
From           : 192.0.2.1
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : target:65000:500
                 l2-vpn/vrf-imp:Encap=19: Flags=-DF: MTU=0: PREF=0
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.1
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h05m30s

Modified Attributes
 ---snip---
-------------------------------------------------------------------------------
Original Attributes

Route Type     : MultiHome
Route Dist.    : 65000:502
Site Id        : 2
Nexthop        : 192.0.2.2
From           : 192.0.2.2
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : target:65000:500
                 l2-vpn/vrf-imp:Encap=19: Flags=none: MTU=0: PREF=0
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.2
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h06m50s

Modified Attributes
 ---snip---
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-3#
```

The following shows the Layer 2 BGP routes on PE-1:

```
*A:PE-1# show service l2-route-table
  - l2-route-table [detail] [bgp-ad] [multi-homing] [bgp-vpls] [bgp-vpws] [all-routes]

 <detail>             : keyword - display detailed information

*A:PE-1# show service l2-route-table multi-homing

===============================================================================
Services: L2 Multi-Homing Route Information - Summary
===============================================================================
Svc Id     L2-Routes (RD-Prefix)        Next Hop        SiteId     State  DF
-------------------------------------------------------------------------------
500        65000:502                    192.0.2.2       2          up(0)  clear
-------------------------------------------------------------------------------
No. of L2 Multi-Homing Route Entries: 1
===============================================================================
*A:PE-1#
```

In case PE-3 were the RR for MTU-4 and MTU-5 as well as for PE-1 and PE-2, PE-1 would have two more L2-routes for multi-homing in this table, as follows:

```
*A:PE-1# show service l2-route-table multi-homing

===============================================================================
Services: L2 Multi-Homing Route Information - Summary
===============================================================================
Svc Id     L2-Routes (RD-Prefix)        Next Hop        SiteId     State  DF
-------------------------------------------------------------------------------
500        65000:504                    192.0.2.4       1          up(0)  set
500        65000:505                    192.0.2.5       1          up(0)  clear
500        65000:502                    192.0.2.2       2          up(0)  clear
-------------------------------------------------------------------------------
No. of L2 Multi-Homing Route Entries: 3
===============================================================================
*A:PE-1#
```

When operational groups are configured (as previously shown), the following **show** command helps to find the operational dependencies between monitoring objects and group objects.

```
*A:PE-1# show service oper-group "group-1" detail

===============================================================================
Service Oper Group Information
===============================================================================
Oper Group       : group-1
Creation Origin  : manual                          Oper Status: up
Hold DownTime    : 0 secs                          Hold UpTime: 4 secs
Members          : 2                               Monitoring : 1
===============================================================================

===================================================================
Member SDP-Binds for OperGroup: group-1
===================================================================
SdpId            SvcId      Type     IP address      Adm     Opr
-------------------------------------------------------------------
12:4294967292    500        BgpVpls  192.0.2.2       Up      Up
13:4294967293    500        BgpVpls  192.0.2.3       Up      Up
-------------------------------------------------------------------
SDP Entries found: 2
===================================================================

===============================================================================
Monitoring Sites for OperGroup: group-1
===============================================================================
SvcId     Site                Site-Id   Dest                Admin   Oper  Fwdr
-------------------------------------------------------------------------------
500       MH-site-2           2         sdp:15:500          Enabled up    Yes
-------------------------------------------------------------------------------
Site Entries found: 1
===============================================================================
*A:PE-1#
```

For debugging, the following CLI sources can be used:

- **log-id 99** — Provides information about the site object changes and DF changes.
- **debug router bgp update** — Shows the BGP updates for BGP MH, including the sent and received BGP MH NLRIs and flags.

```
*A:MTU-4# debug router bgp update
```

- **debug router ldp** commands — Provides information about the pseudowire status bits being signaled as well as the MAC flush messages.

```
*A:MTU-4# debug router ldp peer 192.0.2.1 packet init detail
*A:MTU-4# debug router ldp peer 192.0.2.1 packet label detail
```

As an example, log-id 99 shows the folloiwng debug output after shutting down MH-site-1 on MTU-4:

```
*A:MTU-4# configure service vpls 500 sap 1/1/1:7 shutdown
*A:MTU-4# configure service vpls 500 sap 1/1/2:8 shutdown


*A:MTU-4# show log log-id 99
===============================================================================
Event Log 99
===============================================================================
Description : Default System Log
Memory Log contents  [size=500   next event=91  (not wrapped)]

132 2017/04/26 12:05:01.45 UTC WARNING: SVCMGR #2531 Base BGP-MH
"Service-id 500 site MH-site-1 is not the designated-forwarder"

131 2017/04/26 12:05:01.45 UTC MINOR: SVCMGR #2203 Base
"Status of SAP 1/1/1:7 in service 500 (customer 1) changed to admin=down oper=down
flags=SapAdminDown "

---snip---
```

Log 2 has been configured to log BGP updates and LDP commands.

```
*A:MTU-4# show log log-id 2
===============================================================================
Event Log 2
===============================================================================
Description : (Not Specified)
Memory Log contents  [size=100   next event=11  (not wrapped)]


4 2017/04/26 12:05:01.45 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 86
    Flag: 0x90 Type: 14 Len: 28 Multiprotocol Reachable NLRI:
        Address Family L2VPN
        NextHop len 4 NextHop 192.0.2.5
        [MH] site-id: 1, RD 65000:505
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.5
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        1.1.1.1
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:65000:500
        l2-vpn/vrf-imp:Encap=19: Flags=-DF: MTU=0: PREF=0
"

2 2017/04/26 12:05:01.45 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 72
    Flag: 0x90 Type: 14 Len: 28 Multiprotocol Reachable NLRI:
        Address Family L2VPN
        NextHop len 4 NextHop 192.0.2.4
        [MH] site-id: 1, RD 65000:504
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:65000:500
        l2-vpn/vrf-imp:Encap=19: Flags=D: MTU=0: PREF=0
"
1 2017/04/26 12:05:01.46 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Send Address Withdraw packet (msgId 9795) to 192.0.2.1:0
Protocol version = 1
MAC Flush (All MACs learned from me)
Service FEC PWE3: ENET(5)/500 Group ID = 0 cBit = 0
"
```

Assuming all the recommended tools are enabled, a DF to non-DF transition can be shown as well as the corresponding MAC flush messages and related BGP processing.

If MH-site-2 is torn down on PE-1, the **debug router bgp update** command would allow us to see two BGP updates from PE-1:

- A BGP MH update for site-id 2 with flag D set (because the site is down).
- A BGP VPLS update for veid=501 and flag D set. This is due to the fact that there are no more active objects on the VPLS, besides the BGP pseudowires.

```
*A:PE-1# configure service vpls 500 spoke-sdp 14:500 shutdown
*A:PE-1# configure service vpls 500 spoke-sdp 15:500 shutdown


*A:PE-1# show log log-id 2

===============================================================================
Event Log 2
===============================================================================
Description : (Not Specified)
Memory Log contents  [size=100   next event=141  (wrapped)]

5 2017/04/26 12:18:23.69 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 72
    Flag: 0x90 Type: 14 Len: 28 Multiprotocol Reachable NLRI:
        Address Family L2VPN
        NextHop len 4 NextHop 192.0.2.1
        [MH] site-id: 2, RD 65000:501
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:65000:500
        l2-vpn/vrf-imp:Encap=19: Flags=D: MTU=0: PREF=0
"

4 2017/04/26 12:18:23.69 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 72
    Flag: 0x90 Type: 14 Len: 28 Multiprotocol Reachable NLRI:
        Address Family L2VPN
        NextHop len 4 NextHop 192.0.2.1
        [VPLS/VPWS] preflen 17, veid: 501, vbo: 497, vbs: 8, label-base: 262122,
                    RD 65000:501
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:65000:500
        l2-vpn/vrf-imp:Encap=19: Flags=D: MTU=1514: PREF=0
"
```

The D flag, sent along with the BGP VPLS update for veid 501, would be seen on the remote core PEs as though it was a pseudowire status fault (although there is no TLDP running in the core).

```
*A:PE-2# show service id 500 all | match Flag
Flags              : PWPeerFaultStatusBits
Flags              : None
Flags              : None
Flags              : None
*A:PE-2#
```

Conclusion

SR OS supports a wide range of service resiliency options as well as the best-of-breed system level HA and MPLS mechanisms for the access and the core. BGP MH for VPLS completes the service resiliency tool set by adding a mechanism that has some good advantages over the alternative solutions:

- BGP MH provides a common resiliency mechanism for attachment circuits (SAPs), pseudowires (spoke SDPs), split horizon groups and mesh bindings
- BGP MH is a network-based technique which does not need interaction to the CE or MTU to which it is providing redundancy to.

The examples used in this chapter illustrate the configuration of BGP MH for access CEs and MTUs. Show and debug commands have also been suggested so that the operator can verify and troubleshoot the BGP MH procedures.

BGP Virtual Private Wire Services

This chapter describes BGP Virtual Private Wire Service (VPWS) configurations.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter is applicable to SR OS and was initially written for release 11.0.R4. The CLI in the current edition is based on release 15.0.R2. There are no prerequisites for this configuration.

Introduction

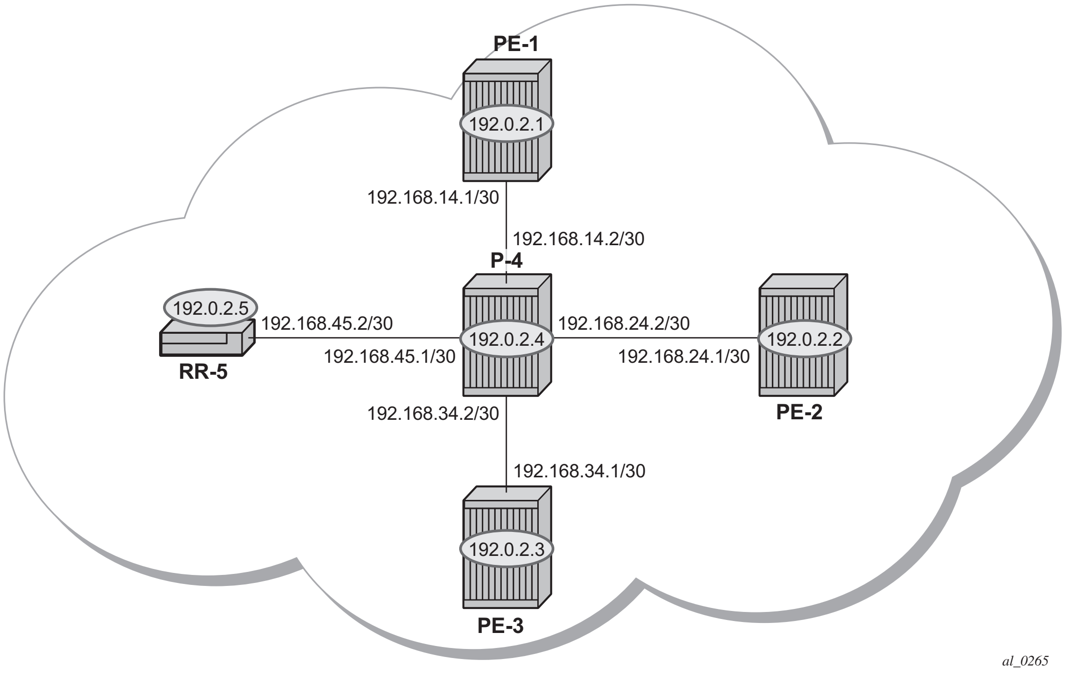
The following two IETF standards describe the provisioning of Virtual Private Wire Services (VPWS):

- RFC 4447, *Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)*, describes Label Distribution Protocol (LDP) VPWS, where VPWS pseudowires are signaled using LDP between Provider Edge (PE) Routers.
- RFC 6624, *Layer 2 Virtual Private Networks Using BGP for Auto-Discovery and Signaling*, describes the use of Border Gateway Protocol (BGP) for signaling of pseudowires between such PEs.

This chapter describes the configuration and troubleshooting for BGP VPWS.

Overview

Figure 24 **Example Topology**



The network topology is shown in Figure 24. The setup uses five SR OS routers located in the same Autonomous System (AS). There are three PE routers connected to a single P router and a route reflector (RR-5) for the AS. The PE routers are all BGP VPWS aware. The Provider (P) router is BGP VPWS unaware and also does not take part in the BGP process.

The following configuration tasks should be completed as a prerequisite:

- IS-IS or OSPF should be configured on each of the network interfaces between the PE/P routers and route reflector.
- MPLS should be configured on all interfaces between PE routers and P routers. It is not required between P-4 and RR-5.
- LDP should be configured on interfaces between PE and P routers. It is not required between P-4 and the RR-5.
- RSVP should be configured on interfaces between PE and P routers. It is not required between P-4 and the RR-5.

BGP VPWS

In this architecture, a VPWS is a collection of two (or three in case of redundancy) BGP VPWS service instances present on different PEs in a provider network.

The PEs communicate with each other at the control plane level by means of BGP updates containing BGP VPWS Network Layer Reachability Information (NLRI). Each update contains enough information for a PE to determine the presence of other BGP VPWS instances on peering PEs and to set up pseudowire connectivity for data flow between peers containing the same BGP VPWS service. Therefore, auto-discovery and pseudowire signaling is achieved using a single BGP update message.

Each PE with a BGP VPWS instance is identified by a VPWS edge identifier (VE-ID) and the presence of other BGP VPWS instances is determined using the exchange of standard BGP extended community route targets between PEs.

Each PE will advertise, via the route reflector, the presence of its BGP VPWS instance to all other PEs, along with a block of multiplexer labels (for BGP VPWS, one label per block) that can be used to communicate between each instance, plus a BGP next-hop that determines a labeled transport tunnel to be used between PEs.

Each BGP VPWS instance is configured with import and export route target extended communities for topology control, along with VE identification.

The following examples show the configuration of four BGP VPWS scenarios.

- Single homed BGP VPWS
  - using auto-provisioned SDPs
  - using pre-provisioned SDPs
- Dual homed BGP VPWS
  - with single pseudowire
  - with active/standby pseudowire

Configure MP-iBGP

The first step is to configure an MP-iBGP session between each of the PEs and the Route Reflector.

The configuration for PE-1 is as follows:

```
configure
    router
```

```
        autonomous-system 65536
        bgp
            group “INTERNAL”
                family l2-vpn
                peer-as 65536
                neighbor 192.0.2.5
                exit
            exit
        exit
    exit
```

The configuration for the other PE nodes is exactly the same. The IP addresses can be derived from Figure 24.

The configuration for the Route Reflector (RR-5) is:

```
configure
    router
        autonomous-system 65536
        bgp
            group “INTERNAL”
                family l2-vpn
                peer-as 65536
                cluster 1.1.1.1
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
            exit
        exit
    exit
```

The following command on RR-5 shows that BGP sessions with each PE are established and have a negotiated L2 VPN address family capability.

```
*A:RR-5# show router bgp summary all

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.1
Def. Instance  65536        3    0 00h00m22s 0/0/0 (L2VPN)
                            3    0
192.0.2.2
Def. Instance  65536        3    0 00h00m22s 0/0/0 (L2VPN)
                            3    0
192.0.2.3
```

```
Def. Instance  65536        3    0 00h00m22s 0/0/0 (L2VPN)
                            3    0

-------------------------------------------------------------------------------
*A:RR-5#
```

Configuration

Pseudowire Templates

BGP VPWS utilizes pseudowire (PW) templates to dynamically instantiate SDP bindings for a service to signal the egress service de-multiplexer labels used by remote PEs to reach the local PE.

The template determines the signaling parameters of the pseudowire, such as vc-type, vlan-vc-tag, hash-label, filters, etc. The following parameters are recognized by BGP VPWS; the remainder is ignored.

The following commands are supported parameters:

```
configure
  service
     [no] pw-template policy-id [use-provisioned-sdp|prefer-provisioned-sdp] [create]
       accounting-policy acct-policy-id
       no accounting-policy
       [no] collect-stats
       [no] controlword
       egress
         filter ipv6 ipv6-filter-id
         filter ip ip-filter-id
         filter mac mac-filter-id
         no filter [ip ip-filter-id] [mac mac-filter-id] [ipv6 ipv6-filter-id]
         qos network-policy-id port-redirect-group queue-group-name
                                                  [instance instance-id]
         no qos
       [no] force-vlan-vc-forwarding
       hash-label [signal-capability]
       no hash-label
       entropy-label
       ingress
         filter ipv6 ipv6-filter-id
         filter ip ip-filter-id
         filter mac mac-filter-id
         no filter [ip ip-filter-id] [mac mac-filter-id] [ipv6 ipv6-filter-id]
        qos network-policy-id fp-redirect-group queue-group-name instance instance-id
         no qos
       [no] sdp-exclude group-name
       [no] sdp-include group-name
       vc-type {ether | vlan}
```

```
        vlan-vc-tag 0..4094
        no vlan-vc-tag
```

Note that:

- The encapsulation type in the Layer-2 extended community is either 4 (Ethernet VLAN tagged mode) or 5 (Ethernet raw mode), depending on the **vc-type** parameter.
- The **force-vlan-vc-forwarding** function will add a tag (equivalent to vc-type vlan) and will allow for customer QoS transparency (dot1p + Drop Eligibility (DE) bits).

The MPLS transport tunnel between PEs can be signaled using LDP or RSVP-TE.

LDP-based SDPs can be automatically instantiated or pre-provisioned. RSVP-TE-based SDPs have to be pre-provisioned. If pre-provisioned pseudowires should be used, the PW template must be created with the **use-provisioned-sdp** parameter. Alternatively, the **prefer-provisioned-sdp** parameter can be used, in which case a pre-provisioned SDP will be used if available; if not, LDP-based SDPs can be automatically instantiated, see chapter LDP VPLS Using BGP Auto-Discovery - Prefer Provisioned SDP.

```
*A:PE-1# configure service pw-template
  - pw-template <policy-id> [create] prefer-provisioned-sdp
  - no pw-template <policy-id>
  - pw-template <policy-id> [use-provisioned-sdp] [create]
```

Pseudowire Templates for Auto-SDP Creation Using LDP

In order to use an LDP transport tunnel for data flow between PEs, it is necessary for link layer LDP to be configured between all PEs/Ps so that a transport label for each PE’s system interface is available. For example, on PE-1:

```
configure
    router
        ldp
            interface-parameters
                interface "int-PE-1-P-4"
                exit
            exit
```

Using this mechanism, SDPs can be auto-instantiated with SDP-ids starting at the higher end of the SDP numbering range, such as 17407. Any subsequent SDPs created use SDP-ids decrementing from this value.

A pseudowire template is required. The following example is created using the default values:

```
configure
    service
        pw-template 1 create
        exit
```

Pseudowire Templates for Provisioned SDPs using RSVP-TE

RSVP-TE LSPs need to be created between the PE routers on which provisioned SDPs will be used as prerequisite.

The MPLS interface and LSP configuration for PE-1 are:

```
configure
    router
        mpls
            interface "int-PE-1-P-4"
            exit
            path "dyn"
                no shutdown
            exit
            lsp "LSP-PE-1-PE-2"
                to 192.0.2.2
                primary "dyn"
                exit
                no shutdown
            exit
            lsp "LSP-PE-1-PE-3"
                to 192.0.2.3
                primary "dyn"
                exit
                no shutdown
            exit
            no shutdown
```

The MPLS and LSP configuration for PE-2 are similar to that of PE-1 with the appropriate interfaces and LSP names configured.

To use an RSVP-TE tunnel as transport between PEs, it is necessary to bind the RSVP-TE LSP between PEs to an SDP.

The SDP creation on PE-1 toward PE-2 is as follows. Similar SDPs are required on each PE to the remote PEs in the service where provisioned SDPs are to be used.

```
configure
    service
        sdp 12 mpls create
        description "SDP-PE-1-PE-2_RSVP_BGP"
        signaling bgp
        far-end 192.0.2.2
        lsp "LSP-PE-1-PE-2"
        no shutdown
    exit
```

The **signaling bgp** parameter is required. BGP VPWS instances using BGP VPWS signaling are able to use these SDPs. Conversely, SDPs that are bound to RSVP-based LSPs with signaling set to the default value of “tldp” will not be used as SDPs within BGP VPWS.

Single Homed BGP VPWS using Auto-Provisioned SDPs

Figure 25 **Single Homed BGP VPWS using Auto-Provisioned SDPs**

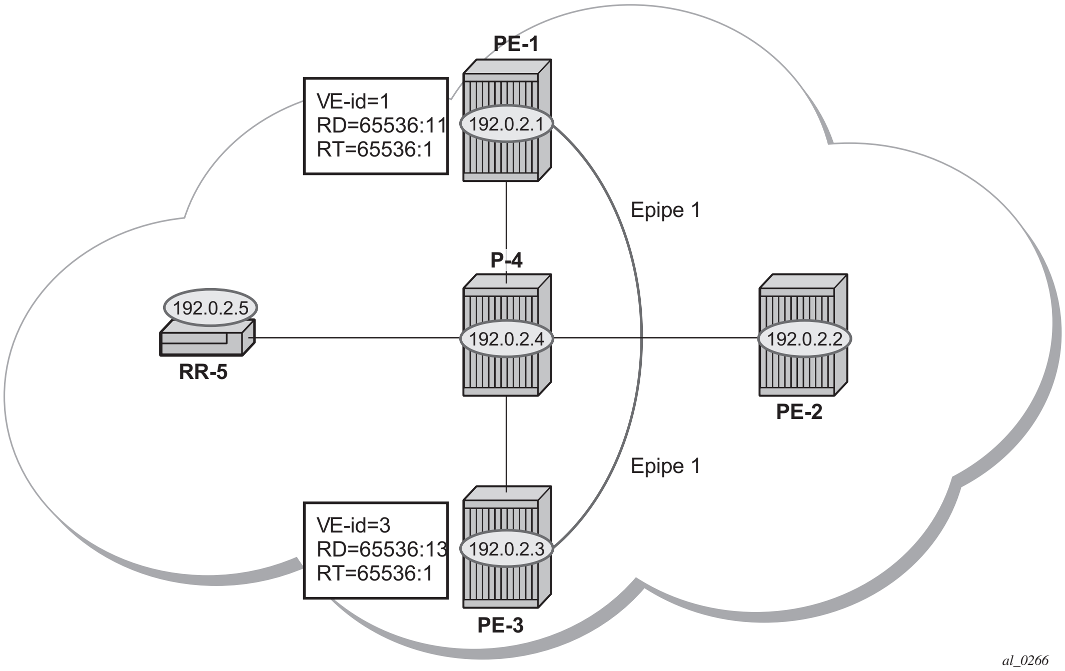


Figure 25 shows a schematic of a single homed BGP VPWS between PE-1 and PE-3 where SDPs are auto-provisioned. In this case, the transport tunnels are LDP signaled.

The following shows the configuration required on PE-1 for a BGP VPWS service using a pseudowire template configured for auto-provisioning of SDPs.

```
*A:PE-1# configure
    service
        pw-template 1 create
            vc-type vlan
        exit
        epipe 1 customer 1 create
            bgp
                route-distinguisher 65536:11
                route-target export target:65536:1 import target:65536:1
                pw-template-binding 1
                exit
            exit
            bgp-vpws
                ve-name "PE-1"
                    ve-id 1
                exit
                remote-ve-name "PE-3"
                    ve-id 3
                exit
                no shutdown
            exit
            sap 1/1/4:1 create
            exit
            no shutdown
        exit
```

The **bgp** context specifies parameters that are required for BGP VPWS.

Within the **bgp** context, parameters are configured that are used by the neighboring PEs to determine the membership of a BGP VPWS; in other words, the auto-discovery of PEs in the same BGP VPWS, the route-distinguisher is configured, along with the route target extended communities. Route target communities are used to determine membership of a BGP VPWS. The import and export route targets at the BGP level are mandatory. The PW template binding is then applied and its parameters are used for both the routes sent by this PE and the received routes matching the route target value.

Within the **bgp-vpws** context, the signaling parameters are also configured. These determine the service labels required for the data plane of the VPWS instance.

The VPWS Edge ID (VE-ID) is a numerical value assigned to each PE within a BGP VPWS. This value must be unique for a BGP VPWS, with the exception of multi-homed scenarios, where two dual-homed PEs can have the same VE-ID and are distinguishable by the site preference (or by the tie breaking rules from the multi-homing draft RFC).

Changes to the pseudowire template are not taken into account once the pseudowire has been set up (changes of route-target are refreshed though). PW-templates can be re-evaluated with the **tools perform service eval-pw-template** command. The **eval-pw-template** checks if all of the bindings using this PW template policy are still meant to be using this policy. If the template has changed and **allow-service-impact** is true, then the old binding is removed and it is re-added using the new template.

```
*A:PE-1# tools perform service eval-pw-template 1
eval-pw-template succeeded for Svc 1 Tx L2 ExtComm, Policy 1
eval-pw-template succeeded for Svc 1 17407:4294967295 Policy 1
*A:PE-1#
```

VE-ID and BGP Label Allocations

For a point-to-point VPWS, there are only two members within the BGP VPWS service, so only one label entry is required by each remote service. For dual-homed scenarios, there are two labels for the redundant site, one from each dual-homed PE.

Each PE allocates a label per BGP VPWS instance for the remote PEs, so it signals blocks with one label. It achieves this by advertising three parameters in a BGP update message. For more information about these parameters, see chapter BGP VPLS.

- A Label Base (LB) which is the lowest label in the block.
- A VE Block size (VBS) which is always 1 and cannot be changed.
- A VE Base Offset (VBO) corresponding to the first label in the label block.

PE-3 Service Creation

On PE-3 create a BGP VPWS service using pseudowire template 1. PE-3 has been allocated a VE-ID of 3. For completeness, the PW template is also shown.

```
*A:PE-3# configure
    service
        pw-template 1 create
            vc-type vlan
        exit
        epipe 1 customer 1 create
            bgp
                route-distinguisher 65536:3
                route-target export target:65536:1 import target:65536:1
                pw-template-binding 1
                exit
            exit
            bgp-vpws
                ve-name "PE-3"
                    ve-id 3
                exit
                remote-ve-name "PE-1"
                    ve-id 1
                exit
                no shutdown
            exit
            sap 1/1/4:1 create
            exit
            no shutdown
        exit
```

PE-1 Service Operation Verification

Verify that the BGP VPWS service is enabled on PE-1.

```
*A:PE-1# show service id 1 bgp-vpws

===============================================================================
BGP VPWS Information
===============================================================================
Admin State          : Enabled
VE Name              : PE-1                 VE Id            : 1
PW Tmpl used         : 1

Remote-Ve Information
-------------------------------------------------------------------------------
Remote VE Name       : PE-3                 Remote VE Id     : 3
===============================================================================
*A:PE-1#
```

Verify the BGP information used by the BGP VPWS service on PE-1.

```
*A:PE-1# show service id 1 bgp

===============================================================================
BGP Information
===============================================================================
Route Dist           : 65536:11
Oper Route Dist      : 65536:11
Oper RD Type         : configured
Rte-Target Import    : 65536:1              Rte-Target Export: 65536:1
Oper RT Imp Origin   : configured           Oper RT Import   : 65536:1
Oper RT Exp Origin   : configured           Oper RT Export   : 65536:1

PW-Template Id       : 1
BFD Template         : None
BFD-Enabled          : no                   BFD-Encap        : ipv4
Import Rte-Tgt       : None
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

Verify that the service is operationally up on PE-1.

```
*A:PE-1# show service id 1 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 1                   Vpn Id            : 0
Service Type      : Epipe
Name              : (Not Specified)
Description       : (Not Specified)
Customer Id       : 1                   Creation Origin   : manual
Last Status Change: 05/02/2017 13:30:00
Last Mgmt Change  : 05/02/2017 13:30:00
Test Service      : No
Admin State       : Up                  Oper State        : Up
MTU               : 1514
Vc Switching      : False
SAP Count         : 1                   SDP Bind Count    : 1
Per Svc Hashing   : Disabled
Force QTag Fwd    : Disabled

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/4:1                              q-tag        1578    1578    Up   Up
sdp:17407:4294967295 SB(192.0.2.3)       BgpVpws      0       1552    Up   Up
===============================================================================
*A:PE-1#
```

The SAP and SDP are all operationally up. The indication “**SB**” next to the SDP-id signifies “Spoke” and “BGP”.

The following output shows the ingress and egress labels for PE-1.

```
*A:PE-1# show service id 1 sdp

===============================================================================
Services: Service Destination Points
===============================================================================
SdpId            Type     Far End addr    Adm     Opr       I.Lbl     E.Lbl
-------------------------------------------------------------------------------
17407:4294967295 BgpVpws  192.0.2.3       Up      Up        262137    262137
-------------------------------------------------------------------------------
Number of SDPs : 1
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

The following debug output from PE-1 shows the BGP VPWS NLRI update for Epipe 1 sent by PE-1 to the route reflector (192.0.2.5). This update will then be received by the other PEs.

```
*A:PE-1# debug router bgp update


*A:PE-1# configure
    log
        log-id 2
            from debug-trace
            to memory
            no shutdown
        exit
```

```
*A:PE-1# show log log-id 2
---snip---

4 2017/05/02 13:30:17.85 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.5
"Peer 1: 192.0.2.5: UPDATE
Peer 1: 192.0.2.5 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 76
    Flag: 0x90 Type: 14 Len: 32 Multiprotocol Reachable NLRI:
        Address Family L2VPN
        NextHop len 4 NextHop 192.0.2.1
       [VPLS/VPWS] preflen 21, veid: 1, vbo: 3, vbs: 1, label-base: 262137, RD 65536:11,
                  csv: 0x00000000, type 1, len 1,
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:65536:1
        l2-vpn/vrf-imp:Encap=4: Flags=none: MTU=1514: PREF=0
"
```

The control flags within the extended community indicate the status of the BGP VPWS instance.

The control flags are the following:

```
0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+
|D|A|F|Z|Z|Z|C|S| (Z = MUST Be Zero)
+-+-+-+-+-+-+-+-+
```

- D: access circuit down indicator. D is 1 if all access circuits are down, otherwise D is 0.
- A: automatic site ID allocation, which is not supported. This is ignored on receipt and set to 0 on sending.
- F: MAC flush indicator, this relates to VPLS. This is set to 0 and ignored on receipt.
- C: presence of a control word. Control word usage is not supported. This is set to 0 on sending (control word not present) and if a non-zero value is received (indicating a control word is required), the pseudowire will not be created.
- S: sequenced delivery. Sequenced delivery is not supported. This is set to 0 on sending (no sequenced delivery) and if a non-zero value is received (indicating sequenced delivery required) the pseudowire will not be created.

The BGP VPWS NLRI is based on the BGP VPLS NLRI, but is extended with a Circuit Status Vector (CSV). The circuit status vector is used to indicate the status of both the SAP and the spoke-SDP within the local service. Because the VE block size used is 1, the most significant bit in the circuit status vector TLV value will be set to 1 if either the SAP or spoke-SDP is down; otherwise, it will be set to 0.

```
*A:PE-1# configure service epipe 1 sap 1/1/4:1 shutdown
```

```
6 2017/05/02 13:34:40.86 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.5
"Peer 1: 192.0.2.5: UPDATE
Peer 1: 192.0.2.5 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 76
    Flag: 0x90 Type: 14 Len: 32 Multiprotocol Reachable NLRI:
        Address Family L2VPN
        NextHop len 4 NextHop 192.0.2.1
        [VPLS/VPWS] preflen 21, veid: 1, vbo: 3, vbs: 1, label-base: 262137,
                    RD 65536:11, csv: 0x00000080, type 1, len 1,
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:65536:1
        l2-vpn/vrf-imp:Encap=4: Flags=D: MTU=1514: PREF=0
"
```

After shutting down the local SAP, the CSV has the most-significant bit set to 1 (0x80).
The BGP VPWS update received on PE-3 can be shown using the following command:

```
*A:PE-3# show service l2-route-table bgp-vpws detail

===============================================================================
Services: L2 Bgp-Vpws Route Information - Summary
===============================================================================

Svc Id         : 1
VeId           : 1
PW Temp Id     : 1
RD             : *65536:11
Next Hop       : 192.0.2.1
State (D-Bit)  : down(1)
Path MTU       : 1514
Control Word   : 0
Seq Delivery   : 0
Status         : active
Tx Status      : active
CSV            : 80
Preference     : 0
Sdp Bind Id    : 17407:4294967295
===============================================================================
*A:PE-3#
```

SAP 1/1/4:1 is re-enabled as follows:

```
*A:PE-1# configure service epipe 1 sap 1/1/4:1 no shutdown
```

PE-3 Service Operation Verification

Similar to PE-1, the service operation should be validated on PE-3.

Single Homed BGP VPWS using Pre-Provisioned SDP

It is possible to configure BGP VPWS instances that use RSVP-TE transport tunnels. In this case, the SDPs must be created with the MPLS LSPs mapped and with the signaling set to BGP, because the service labels are signaled using BGP. The PW template configured within the BGP VPWS instance must use the keyword **use-provisioned-sdp** (or **prefer-provisioned-sdp**).

Figure 26 **Single Homed BGP VPWS using Pre-Provisioned SDP**

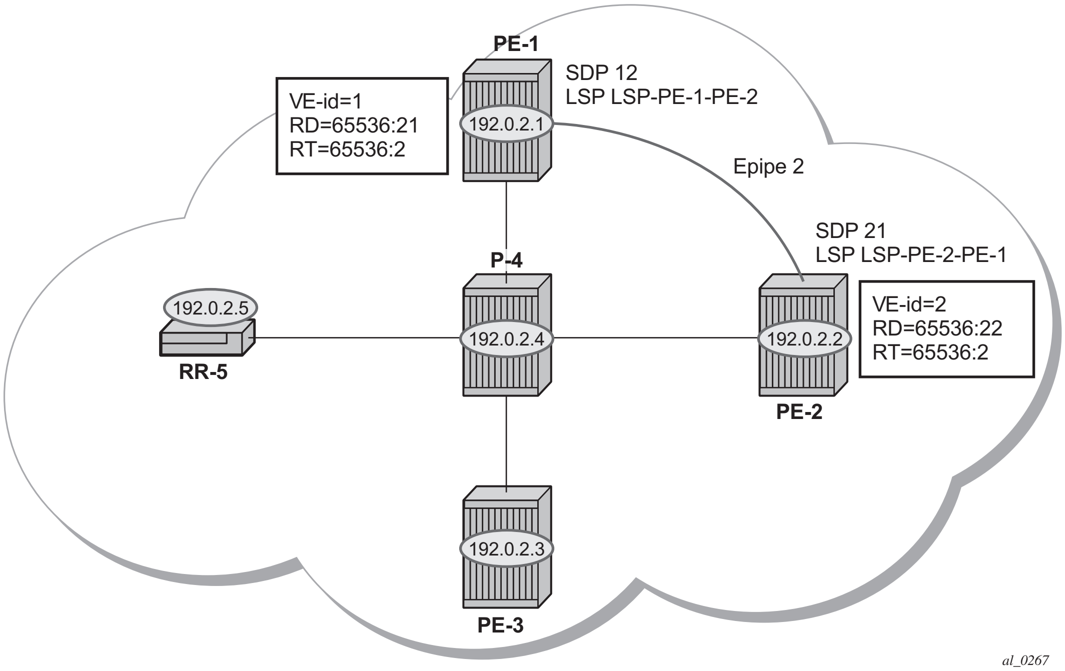


Figure 26 shows a schematic of a BGP VPWS where SDPs are pre-provisioned with RSVP-TE signaled transport tunnels.

SDP on PE-1

```
*A:PE-1# configure
    service
        sdp 12 mpls create
            description "SDP-PE-1-PE-2_RSVP_BGP"
            signaling bgp
            far-end 192.0.2.2
            lsp "LSP-PE-1-PE-2"
            no shutdown
        exit
```

SDP on PE-2

```
*A:PE-2# configure
    service
        sdp 21 mpls create
            description "SDP-PE-2-PE-1_RSVP_BGP"
            signaling bgp
            far-end 192.0.2.1
            lsp "LSP-PE-2-PE-1"
            no shutdown
        exit
```

To create a spoke SDP within a service that uses the RSVP-TE transport tunnel, a pseudowire template is required that has the **use-provisioned-sdp** parameter set.

The PW template is provisioned on both PEs as follows:

```
*A:PE-1# configure
    service
        pw-template 2 use-provisioned-sdp create
        exit
```

The following output shows the configuration required for a BGP VPWS service using a pseudowire template configured for using pre-provisioned RSVP-TE SDPs.

```
*A:PE-1# configure
    service
        epipe 2 customer 1 create
            bgp
                route-distinguisher 65536:21
                route-target export target:65536:2 import target:65536:2
                pw-template-binding 2
                exit
            exit
            bgp-vpws
                ve-name "PE-1"
                    ve-id 1
                exit
                remote-ve-name "PE-2"
                    ve-id 2
                exit
                no shutdown
            exit
            sap 1/1/4:2 create
            exit
            no shutdown
```

The route distinguisher and route target extended community values for Epipe 2 are different from that in Epipe 1. This is to differentiate between the two as their visibility is global within the BGP domain. The VE-ID values can be reused in each Epipe instance, as long as they are unique within the instance.

Similarly, the configuration is as follows on PE-2, where the VE-ID is 2:

```
*A:PE-2# configure
    service
         epipe 2 customer 1 create
            bgp
                route-distinguisher 65536:22
                route-target export target:65536:2 import target:65536:2
                pw-template-binding 2
                exit
            exit
            bgp-vpws
                ve-name "PE-2"
                    ve-id 2
                exit
                remote-ve-name "PE-1"
                    ve-id 1
                exit
                no shutdown
            exit
            sap 1/1/4:2 create
            exit
            no shutdown
```

Verify that the service is operationally up on PE-1.

```
*A:PE-1# show service id 2 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 2                   Vpn Id            : 0
Service Type      : Epipe
---snip---
Admin State       : Up                  Oper State        : Up
MTU               : 1514
Vc Switching      : False
SAP Count         : 1                   SDP Bind Count    : 1
---snip---
-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/4:2                              q-tag        1578    1578    Up   Up
sdp:12:4294967294 S(192.0.2.2)           BgpVpws      0       1552    Up   Up
===============================================================================
*A:PE-1#
```

The SDP-id is the pre-provisioned SDP 12.

For completeness, verify the service is operationally up on PE-2.

```
*A:PE-2# show service id 2 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 2                   Vpn Id            : 0
Service Type      : Epipe
---snip---
Admin State       : Up                  Oper State        : Up
MTU               : 1514
Vc Switching      : False
SAP Count         : 1                   SDP Bind Count    : 1
---snip---
-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/4:2                              q-tag        1578    1578    Up   Up
sdp:21:4294967295 S(192.0.2.1)           BgpVpws      0       1552    Up   Up
===============================================================================
*A:PE-2#
```

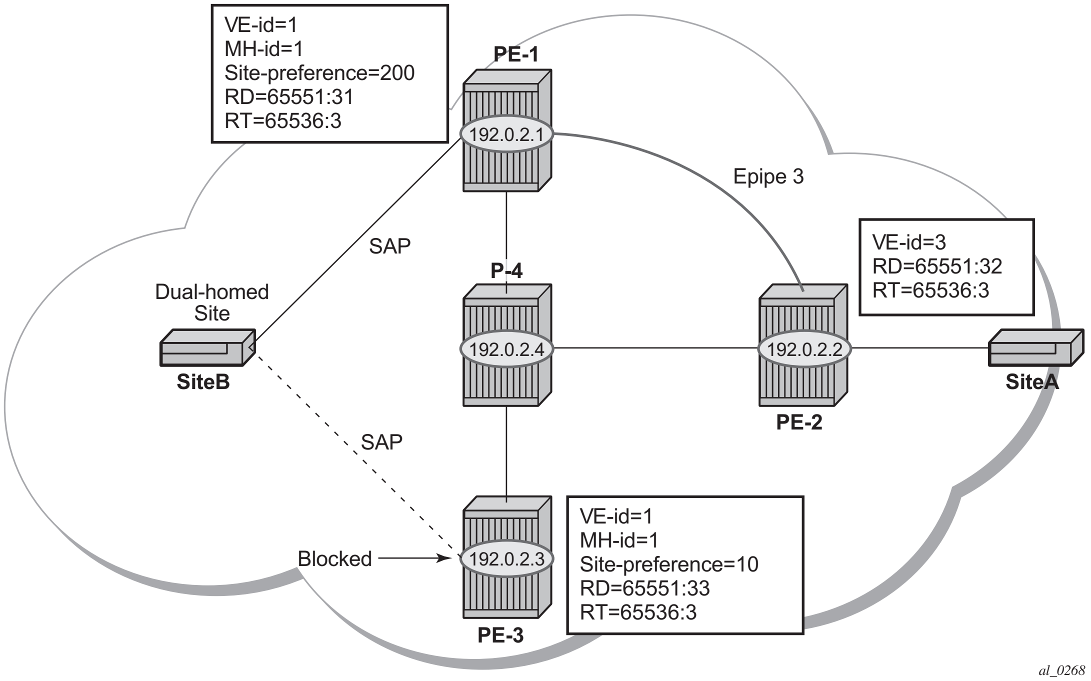
The SDP-id used is the pre-provisioned SDP 21.

Dual Homed BGP VPWS with Single Pseudowire

For access redundancy, an Epipe using a BGP VPWS service can be configured as dual-homed, as described in *draft-ietf-l2vpn-vpls-multihoming-03*. It can be configured with a single pseudowire setup, where the redundant pseudowire is not created until the initially active pseudowire is removed.

The following diagram shows a setup where an Epipe is configured on each PE. Site B is dual-homed to PE-1 and PE-3 with the remote PE-2 connected to site A; each site connection uses a SAP. A single pseudowire using Ethernet Raw Mode encapsulation connects PE-2 to PE-1 or PE-3 (but not both at the same time). The pseudowire is signaled using BGP VPWS over a tunnel LSP between the PEs.

Figure 27 **Dual Homed BGP VPWS with Single Pseudowire**



BGP multi-homing is configured for the dual-homed site B using a site-id=1. The site-preference on PE-1 is set to 200 and to 10 on PE-3, this ensures that PE-1 will be the site's Designated Forwarder (DF) and the pseudowire from PE-2 will be created to PE-1 when PE-1 is fully operational (no pseudowire is created on PE-2 to PE-3). If PE-1 fails, or the multi-homing site fails over to PE-3, then the pseudowire from PE-2 to PE-1 will be removed and a new pseudowire will be created from PE-2 to PE-3.

Epipe 3 is configured on PE-1 as follows:

```
*A:PE-1# configure
    service
         pw-template 3 create
         exit
         epipe 3 customer 1 create
            bgp
                route-distinguisher 65536:31
                route-target export target:65536:3 import target:65536:3
                pw-template-binding 3
                exit
            exit
            bgp-vpws
                ve-name "PE-1"
                    ve-id 1
                exit
                remote-ve-name "PE-2"
                    ve-id 2
                exit
                no shutdown
            exit
            site "SITEB" create
                site-id 1
                sap 1/1/4:3
                site-preference 200
                no shutdown
            exit
            sap 1/1/4:3 create
            exit
            no shutdown
        exit
```

Epipe 3 is configured on PE-3 as follows:

```
*A:PE-3# configure
    service
        pw-template 3 create
        exit
        epipe 3 customer 1 create
            bgp
                route-distinguisher 65536:33
                route-target export target:65536:3 import target:65536:3
                pw-template-binding 3
                exit
            exit
            bgp-vpws
                ve-name "PE-3"
                    ve-id 1
                exit
                remote-ve-name "PE-2"
                    ve-id 2
                exit
                no shutdown
            exit
            site "SITEB" create
                site-id 1
                sap 1/1/4:3
                site-preference 10
                no shutdown
            exit
            sap 1/1/4:3 create
            exit
            no shutdown
        exit
```

In the preceding configurations, the **remote-ve-name** for PE-2 uses VE-ID 2 on both PE-1 and PE-3.

Epipe 3 is configured on PE-2 as follows:

```
*A:PE-2# configure
    service
        pw-template 3 create
        exit
        epipe 3 customer 1 create
            bgp
```

```
                route-distinguisher 65536:32
                route-target export target:65536:3 import target:65536:3
                pw-template-binding 3
                exit
            exit
            bgp-vpws
                ve-name "PE-2"
                    ve-id 2
                exit
                remote-ve-name "PE-1 or PE-3"
                    ve-id 1
                exit
                no shutdown
            exit
            sap 1/1/4:3 create
            exit
            no shutdown
        exit
```

On PE-2, the **remote-ve-name** is configured as “PE-1 or PE-3”; this is because both of these PEs are configured with VE-ID 1.

As a result of this configuration, there are multiple route entries for Route-Target 65536:31 on PE-2. In the BGP routing table, there are two entries per partner PE, one for the BGP-MH update (with site-id=1) and the other for the BGP-VPWS update (with VE-ID=1).

```
 *A:PE-2# show router bgp routes l2-vpn rd 65536:31
===============================================================================
 BGP Router ID:192.0.2.2        AS:65536       Local AS:65536
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP L2VPN Routes
===============================================================================
Flag  RouteType                 Prefix                              MED
      RD                        SiteId                               Label
      Nexthop                   VeId                   BlockSize     LocalPref
      As-Path                   BaseOffset             vplsLabelBa
                                                       se
-------------------------------------------------------------------------------
u*>i  MultiHome                 -                      -             0
      65536:31                  1                                    -
      192.0.2.1                 -                      -             200
      No As-Path                -                      -
u*>i  VPWS                      -                      -             0
      65536:31                  -                                    -
      192.0.2.1                 1                      1             200
      No As-Path                2                      262135
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-2#
```

```
*A:PE-2# show router bgp routes l2-vpn rd 65536:33
===============================================================================
 BGP Router ID:192.0.2.2        AS:65536       Local AS:65536
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP L2VPN Routes
===============================================================================
Flag  RouteType                   Prefix                             MED
      RD                          SiteId                             Label
      Nexthop                     VeId                   BlockSize   LocalPref
      As-Path                     BaseOffset             vplsLabelBa
                                                         se
-------------------------------------------------------------------------------
u*>i  MultiHome                   -                      -           0
      65536:33                    1                                  -
      192.0.2.3                   -                      -           10
      No As-Path                  -                      -
u*>i  VPWS                        -                      -           0
      65536:33                    -                                  -
      192.0.2.3                   1                      1           10
      No As-Path                  2                      262136
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-2#
```

The route to PE-1 has the higher site preference, so it is selected as the target for the pseudowire.

```
*A:PE-2# show service l2-route-table bgp-vpws detail

===============================================================================
Services: L2 Bgp-Vpws Route Information - Summary
===============================================================================

---snip---

Svc Id         : 3
VeId           : 1
PW Temp Id     : 3
RD             : *65536:31
Next Hop       : 192.0.2.1
State (D-Bit)  : up(0)
Path MTU       : 1514
Control Word   : 0
Seq Delivery   : 0
Status         : active
Tx Status      : active
CSV            : 0
Preference     : 200
Sdp Bind Id    : 17407:4294967292
===============================================================================
*A:PE-2#
```

After disabling the SAP in the service on PE-1, BGP update messages are received. The VPLS/VPWS message received on PE-2 from PE-1 shows in the CSV that the access circuit is down (the CSV has the most-significant bit set to 1 (0x80)), so PE-2 selects the update from PE-3 to create the pseudowire. The BGP-MH update received by PE-2 from PE-1 also shows that the local site is down as indicated by the flags=D.

Note in the following debug output,

- BGP MH (multi-homing) entry uses encap-type=19.
- BGP VPWS entry uses encap-type=5 (Ethernet raw mode).

```
*A:PE-1# configure service epipe 3 sap 1/1/4:3 shutdown

37 2017/05/02 13:41:57.10 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.5
"Peer 1: 192.0.2.5: UPDATE
Peer 1: 192.0.2.5 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 90
    Flag: 0x90 Type: 14 Len: 32 Multiprotocol Reachable NLRI:
        Address Family L2VPN
        NextHop len 4 NextHop 192.0.2.1
      [VPLS/VPWS] preflen 21, veid: 1, vbo: 2, vbs: 1, label-base: 262135, RD 65536:31,
                  csv: 0x00000080, type 1, len 1,
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 0
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.1
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        1.1.1.1
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:65536:3
        l2-vpn/vrf-imp:Encap=5: Flags=D: MTU=1514: PREF=200
"

36 2017/05/02 13:41:57.10 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.5
"Peer 1: 192.0.2.5: UPDATE
Peer 1: 192.0.2.5 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 86
    Flag: 0x90 Type: 14 Len: 28 Multiprotocol Reachable NLRI:
        Address Family L2VPN
        NextHop len 4 NextHop 192.0.2.1
        [MH] site-id: 1, RD 65536:31
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 0
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.1
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        1.1.1.1
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:65536:3
        l2-vpn/vrf-imp:Encap=19: Flags=D: MTU=0: PREF=200
```

```
"
```

The result can be shown on PE-2 as now the spoke SDP is up (active) to PE-3.

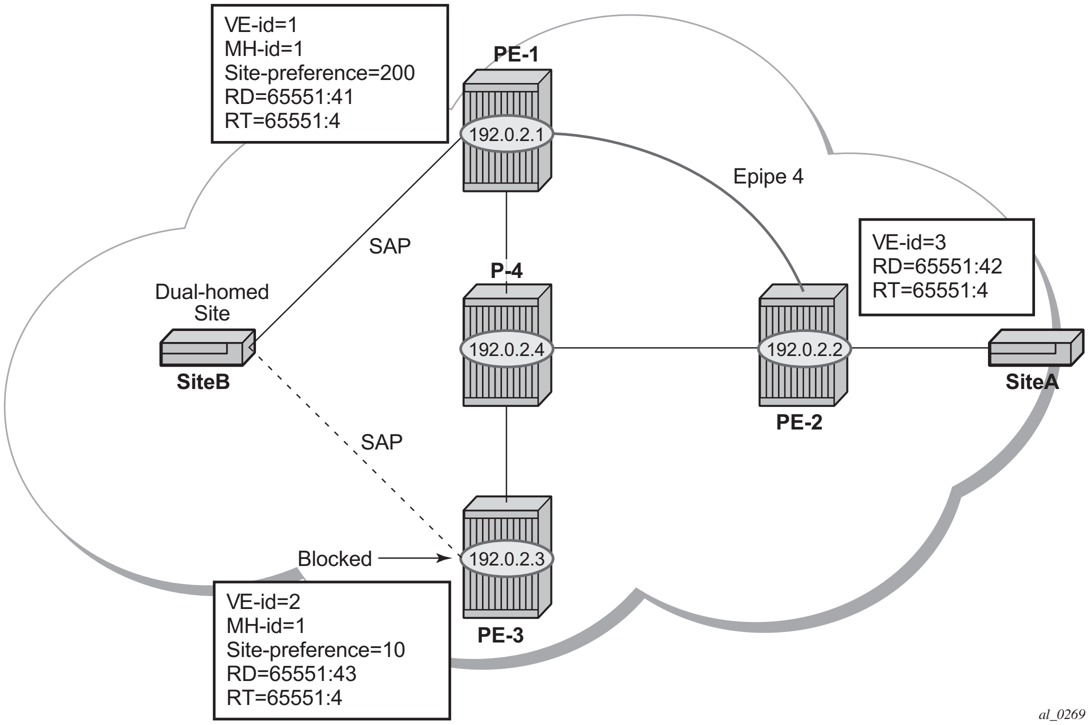
```
*A:PE-2# show service l2-route-table bgp-vpws detail

===============================================================================
Services: L2 Bgp-Vpws Route Information - Summary
===============================================================================

---snip---

Svc Id          : 3
VeId            : 1
PW Temp Id      : 3
RD              : *65536:33
Next Hop        : 192.0.2.3
State (D-Bit)   : up(0)
Path MTU        : 1514
Control Word    : 0
Seq Delivery    : 0
Status          : active
Tx Status       : active
CSV             : 0
Preference      : 10
Sdp Bind Id     : 17407:4294967291
===============================================================================
*A:PE-2#
```

Dual Homed BGP VPWS with Active/Standby Pseudowire

The second method for BGP VPWS pseudowire redundancy is an active/standby configuration. Whereas in the solution with one pseudowire, the redundant nodes use the same VE-ID for the remote PE and different preferences; in the active/standby solution, the redundant nodes use different VE-IDs for the remote PE and different preferences. The node connecting to both pseudowires (PE-2 in this example) has both remote VE-IDs configured. This allows for faster failover because the standby pseudowire is instantiated in addition to the active pseudowire. If more than two applicable BGP updates are received, at most one standby pseudowire is created (based on the BGP VPWS tie breaking rules).

Figure 28 shows a setup where an Epipe is configured on each PE. Site B is dual-homed to PE-1 and PE-3 with the remote PE-2 connected to site A; each site connection uses a SAP. The active/standby pseudowires using Ethernet raw mode encapsulation connect PE-2 to PE-1 and PE-3. The pseudowires are signaled using BGP VPWS over tunnel LSPs between the PEs.

Figure 28 **Dual Homed BGP VPWS with Active/Standby Pseudowire**



BGP Multi-Homing (MH) is configured for the dual-homed site B using a site-id=1. The site-preference on PE-1 is set to 200 and to 10 on PE-3; this ensures that PE-1 will be the site's designated forwarder for the MH site. The active pseudowire from PE-2 will be created to PE-1 with the standby pseudowire being created to PE-3. If PE-1 fails, or the multi-homing site fails over to PE-3, then the pseudowire from PE-2 to PE-3 will become active (used as the data path between site A and B).

Epipe 4 is configured on PE-1 as follows:

```
*A:PE-1# configure
    service
         pw-template 3 create
         exit
         epipe 4 customer 1 create
            bgp
                route-distinguisher 65536:41
                route-target export target:65536:4 import target:65536:4
                pw-template-binding 3
                exit
            exit
            bgp-vpws
                ve-name "PE-1"
                    ve-id 1
                exit
                remote-ve-name "PE-2"
                    ve-id 2
                exit
                no shutdown
            exit
            site "SITEB" create
                site-id 1
                sap 1/1/4:4
                site-preference 200
                no shutdown
            exit
            sap 1/1/4:4 create
            exit
            no shutdown
        exit
```

Epipe 4 is configured on PE-3 as follows:

The local VE-ID is 3 (different from previous example).

```
*A:PE-3# configure
    service
         pw-template 3 create
         exit
         epipe 4 customer 1 create
            bgp
                route-distinguisher 65536:43
                route-target export target:65536:4 import target:65536:4
                pw-template-binding 3
                exit
            exit
            bgp-vpws
                ve-name "PE-3"
                    ve-id 3
                exit
                remote-ve-name "PE-2"
                    ve-id 2
                exit
                no shutdown
            exit
            site "SITEB" create
                site-id 1
                sap 1/1/4:4
                site-preference 10
                no shutdown
            exit
            sap 1/1/4:4 create
            exit
            no shutdown
        exit
```

Epipe 4 is configured on PE-2 as follows:

There are two remote VE names configured, PE-1 and PE-3 (this is the maximum number allowed).

```
*A:PE-2# configure
    service
```

```
pw-template 3 create
exit
epipe 4 customer 1 create
   bgp
       route-distinguisher 65536:42
       route-target export target:65536:4 import target:65536:4
       pw-template-binding 3
       exit
   exit
   bgp-vpws
       ve-name "PE-2"
           ve-id 2
       exit
       remote-ve-name "PE-1"
           ve-id 1
       exit
       remote-ve-name "PE-3"
           ve-id 3
       exit
       no shutdown
   exit
   sap 1/1/4:4 create
   exit
   no shutdown
```

Compared with the single pseudowire solution, both pseudowires are signaled and up on all PEs. The pseudowire with the higher preference is forwarding traffic (to PE-1), while the Tx Status on the other one is set to inactive, as follows:

```
*A:PE-2# show service l2-route-table bgp-vpws detail

===============================================================================
Services: L2 Bgp-Vpws Route Information - Summary
===============================================================================

---snip---

Svc Id         : 4
VeId           : 1
PW Temp Id     : 3
RD             : *65536:41
Next Hop       : 192.0.2.1
State (D-Bit)  : up(0)
Path MTU       : 1514
Control Word   : 0
Seq Delivery   : 0
Status         : active
Tx Status      : active
CSV            : 0
Preference     : 200
Sdp Bind Id    : 17407:4294967288

Svc Id         : 4
VeId           : 3
PW Temp Id     : 3
RD             : *65536:43
Next Hop       : 192.0.2.3
State (D-Bit)  : up(0)
```

```
Path MTU       : 1514
Control Word   : 0
Seq Delivery   : 0
Status         : active
Tx Status      : inactive
CSV            : 0
Preference     : 10
Sdp Bind Id    : 17406:4294967289
===============================================================================
*A:PE-2#
```

The choice of pseudowire to be used to transmit traffic from PE-2 to PE-1 can also be seen in the endpoint created in the BGP VPWS service. Endpoints are automatically created for the pseudowires within a BGP VPWS service regardless of whether active/standby pseudowires are used; these endpoints are created with a system generated name that ends with the BGP VPWS service id.

```
*A:PE-2# show service id 4 endpoint

===============================================================================
Service 4 endpoints
===============================================================================
Endpoint name                : _tmnx_BgpVpws-4
Description                  : Automatically created BGP-VPWS endpoint
Creation Origin              : bgpVpws
Revert time                  : 0
Act Hold Delay               : 0
Standby Signaling Master     : false
Standby Signaling Slave      : false
Tx Active (SDP)              : 17407:4294967288
Tx Active Up Time            : 0d 00:00:48
Revert Time Count Down       : N/A
Tx Active Change Count       : 3
Last Tx Active Change        : 05/02/2017 13:46:52
-------------------------------------------------------------------------------
Members
-------------------------------------------------------------------------------
Spoke-sdp: 17406:4294967289 Prec:4                  Oper Status: Up
Spoke-sdp: 17407:4294967288 Prec:4                  Oper Status: Up
===============================================================================
===============================================================================
*A:PE-2#
```

The following command has no effect on an automatically created VPWS endpoint.

```
tools perform service id <service-id> endpoint <endpoint-name> force-switchover
```

Conclusion

BGP VPWS allows the delivery of Layer 2 virtual private wire services to customers where BGP is commonly used. This chapter shows the configuration of single and dual-homed BGP VPWS services together with the associated show output, which can be used to verify and troubleshoot them.

BGP VPLS

This chapter describes advanced BGP VPLS configurations.

Topics in this chapter include:

- Applicability
- Summary
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SR OS release 9.0.R3. The CLI in the current edition corresponds to release 15.0.R2. There are no prerequisites for this configuration.

Summary

The following two IETF standards describe the provisioning of Virtual Private LAN Services (VPLS).

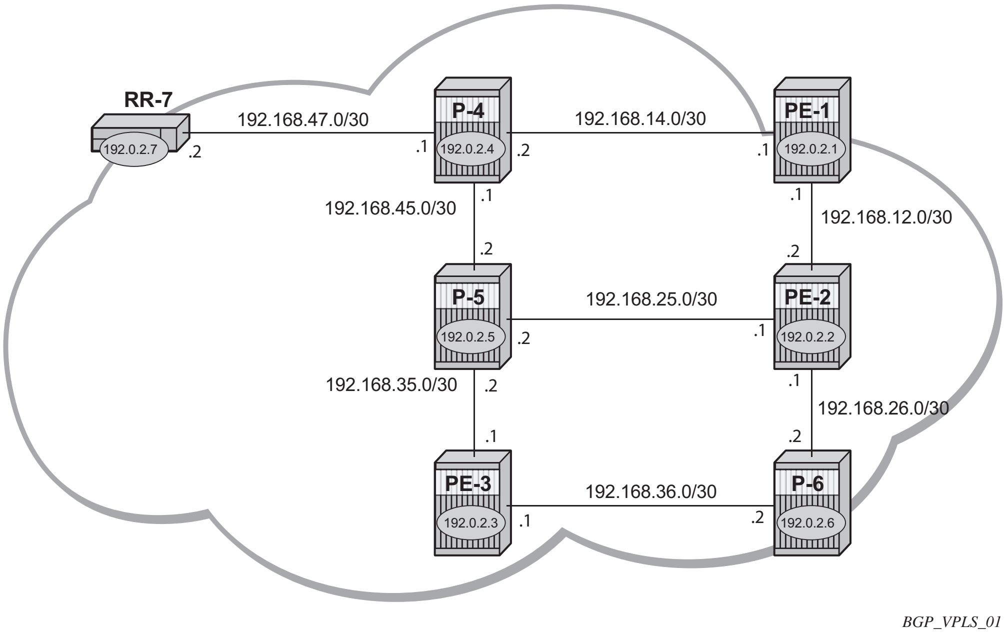
- RFC 4762, *Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling*, describes Label Distribution Protocol (LDP) VPLS, where VPLS pseudowires are signaled using LDP between VPLS Provider Edge (PE) routers, either configured manually or auto-discovered using BGP.
- RFC 4761, *Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling*, describes the use of Border Gateway Protocol (BGP) for both the auto-discovery of VPLS PEs and signaling of pseudowires between such PEs.

The purpose of this section is to describe the configuration and troubleshooting for BGP-VPLS.

Knowledge of BGP-VPLS RFC 4761 architecture and functionality is assumed throughout this chapter, as well as knowledge of Multi-Protocol BGP (MP-BGP).

Overview

Figure 29 **Network Topology**



The network topology is displayed in Figure 29. The configuration uses seven 7750 Service Router (SR) nodes located in the same Autonomous System (AS). There are three Provider Edge (PE) routers, and RR-7 will act as a Route Reflector (RR) for the AS. The PE routers are all VPLS-aware, the Provider (P) routers are VPLS unaware and do not take part in the BGP process.

The following configuration tasks should be completed as a prerequisite:

- IS-IS or OSPF on each of the network interfaces between the PE/P routers and RR.
- MPLS should be configured on all interfaces between PE routers and P routers. MPLS is not required between P-4 and RR-7.
- LDP should be configured on interfaces between PE and P routers. It is not required between P-4 and the RR-7.
- The RSVP protocol should be enabled.

BGP VPLS

In this architecture, a VPLS instance is a collection of local VPLS instances present on a number of PEs in a provider network. In this context, any VPLS-aware PE is also known as a VPLS Edge (VE) device.

The PEs communicate with each other at the control plane level by means of BGP updates containing BGP-VPLS Network Layer Reachability Information (NLRI). Each update contains enough information for a PE to determine the presence of other local VPLS instances on peering PEs and to set up pseudowire connectivity for data flow between peers containing a local VPLS within the same VPLS instance. Therefore, auto-discovery and pseudowire signaling are achieved using a single BGP update message.

Each PE within a VPLS instance is identified by a VPLS Edge identifier (ve-id) and the presence of a VPLS instance is determined using the exchange of standard BGP extended community route targets between PEs.

Each PE will advertise, via the route reflectors, the presence of each VPLS instance to all other PEs, along with a block of multiplexer labels that can be used to communicate between such instances plus a BGP next hop that determines a labeled transport tunnel between PEs.

Each VPLS instance is configured with import and export route target extended communities for topology control, along with VE identification.

Configuration

The first step is to configure an MP-iBGP session between each of the PEs and the RR.

The configuration for PE-1 is as follows:

```
*A:PE-1# configure
    router
        autonomous-system 65536
        bgp
            group "INTERNAL"
                family l2-vpn
                peer-as 65536
                neighbor 192.0.2.7
                exit
            exit
            no shutdown
        exit
```

The BGP configuration for the other PE nodes is identically the same. The IP addresses can be derived from Figure 29.

The configuration for RR-7 is as follows:

```
*A:RR-7# configure
    router
        autonomous-system 65536
        bgp
            cluster 1.1.1.1
            group "RR-INTERNAL"
                family l2-vpn
                peer-as 65536
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
            exit
            no shutdown
        exit
```

On PE-1, verify that the BGP session with RR-7 is established with address family l2-vpn capability negotiated:

```
*A:PE-1# show router bgp neighbor 192.0.2.7

===============================================================================
BGP Neighbor
===============================================================================
-------------------------------------------------------------------------------
Peer                 : 192.0.2.7
Description          : (Not Specified)
Group                : INTERNAL
-------------------------------------------------------------------------------
Peer AS              : 65536            Peer Port            : 50439
Peer Address         : 192.0.2.7
Local AS             : 65536            Local Port           : 179
Local Address        : 192.0.2.1
Peer Type            : Internal         Dynamic Peer         : No
State                : Established      Last State           : Established
Last Event           : recvKeepAlive
Last Error           : Cease (Connection Collision Resolution)
Local Family         : L2-VPN
Remote Family        : L2-VPN
Hold Time            : 90               Keep Alive           : 30
Min Hold Time        : 0
Active Hold Time     : 90               Active Keep Alive    : 30
Cluster Id           : None
Preference           : 170              Num of Update Flaps  : 0
---snip---
Local Capability     : RtRefresh MPBGP 4byte ASN
Remote Capability    : RtRefresh MPBGP 4byte ASN
---snip---
-------------------------------------------------------------------------------
Neighbors : 1
```

```
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-1#
```

On RR-7, show that BGP sessions with each PE are established, and have a negotiated the l2-vpn address family capability.

```
*A:RR-7# show router bgp summary all

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.1
Def. Instance  65536        6    0 00h01m38s 0/0/0 (L2VPN)
                            6    0
192.0.2.2
Def. Instance  65536        6    0 00h01m38s 0/0/0 (L2VPN)
                            6    0
192.0.2.3
Def. Instance  65536        6    0 00h01m38s 0/0/0 (L2VPN)
                            6    0

-------------------------------------------------------------------------------
*A:RR-7#
```

Configure a full mesh of RSVP-TE LSPs between PE routers.

The MPLS interface and LSP configuration for PE-1 is as follows:

```
*A:PE-1# configure
    router
        mpls
            interface "int-PE-1-PE-2"
            exit
            interface "int-PE-1-P-4"
            exit
            path "loose"
                no shutdown
            exit
            lsp "LSP-PE-1-PE-2"
                to 192.0.2.2
                primary "loose"
                exit
                no shutdown
            exit
            lsp "LSP-PE-1-PE-3"
                to 192.0.2.3
                primary "loose"
                exit
                no shutdown
            exit
            no shutdown
        exit
```

The MPLS and LSP configuration for PE-2 and PE-3 are similar to that of PE-1 with the appropriate interfaces and LSP names configured.

BGP VPLS PE Configuration

Pseudowire Templates

Pseudowire templates are used by BGP to dynamically instantiate SDP bindings, for a given service, to signal the egress service de-multiplexer labels used by remote PEs to reach the local PE.

The template determines the signaling parameters of the pseudowire, control word presence, MAC-pinning, filters and so on, plus other usage characteristics such as split horizon groups.

The MPLS transport tunnel between PEs can be signaled using LDP or RSVP-TE.

LDP based pseudowires can be automatically instantiated. RSVP-TE based SDPs have to be pre-provisioned.

Pseudowire Templates for Auto-SDP Creation Using LDP

In order to use an LDP transport tunnel for data flow between PEs, it is necessary for link layer LDP to be configured between all PEs/Ps, so that a transport label for each PE's system interface is available.

```
*A:PE-1# configure
    router
        ldp
            interface-parameters
                interface "int-PE-1-PE-2"
                exit
                interface "int-PE-1-P-4
                exit
            exit
        exit
```

Using this mechanism, SDPs can be auto-instantiated with SDP-IDs starting at the higher end of the SDP numbering range, such as 17407. Any subsequent SDPs created use SDP-IDs decrementing from this value.

A pseudowire template is required containing a split horizon group. Each SDP created with this template is contained within a split horizon group so that traffic cannot be forwarded between them.

```
*A:PE-1# configure
    service
        pw-template 1 create
            split-horizon-group "VPLS-SHG"
            exit
        exit
```

The pseudowire template also has the following options available when used for BGP-VPLS:

```
*A:PE-1# configure service pw-template
  ---snip---
  [no] controlword
  ---snip---
  [no] force-vlan-vc-forwarding
  ---snip---
   vc-type {ether | vlan}
  ---snip---
```

- The control word will determine whether the C flag is set in the Layer 2 extended community and, therefore, if a control word is used in the pseudowire.
- The encap type in the Layer 2 extended community is always 19 (VPLS encap), therefore, the vc-type will always be **ether** regardless of the configured value on the vc-type.
- The **force-vlan-vc-forwarding** command will add a tag (equivalent to **vc-type vlan**) and will allow for customer QoS transparency (dot1p + Drop Eligibility (DE) bits).

Pseudowire Templates for Provisioned SDPs using RSVP-TE

To use an RSVP-TE tunnel as transport between PEs, it is necessary to bind the RSVP-TE LSP between PEs to an SDP.

The following SDP is created from PE-1 to PE-2:

```
*A:PE-1# configure
    service
        sdp 12 mpls create
            description "SDP-PE-1-PE-2_RSVP_BGP"
            signaling bgp
            far-end 192.0.2.2
            lsp "LSP-PE-1-PE-2"
            no shutdown
        exit
```

The **signaling bgp** parameter is required for BGP-VPLS to be able to use this SDP. Conversely, SDPs that are bound to RSVP-based LSPs with signaling set to the default value of **tldp** will not be used as SDPs within BGP-VPLS.

BGP VPLS Using Auto-Provisioned SDPs

Figure 30 **BGP VPLS Using Auto-Provisioned SDPs**

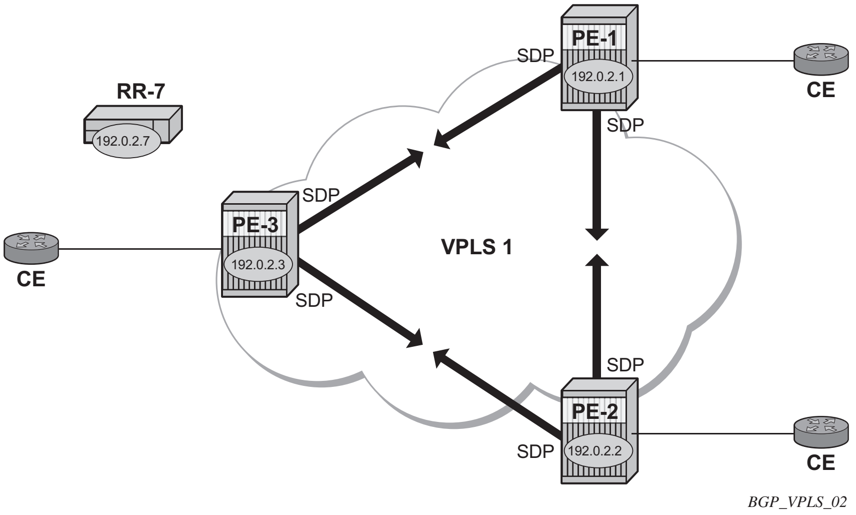


Figure 30 shows a VPLS instance where SDPs are auto-provisioned. In this case, the transport tunnels are LDP signaled.

The following output shows the configuration required on PE-1 for a BGP-VPLS service using a pseudowire template configured for auto-provisioning of SDPs.

```
*A:PE-1# configure
    service
        pw-template 1 create
            split-horizon-group "VPLS-SHG"
            exit
        exit
        vpls 1 customer 1 create
            bgp
                route-distinguisher 65536:1
                route-target export target:65536:1 import target:65536:1
                pw-template-binding 1
                exit
            exit
            bgp-vpls
                max-ve-id 10
                ve-name "PE-1"
                    ve-id 1
                exit
                no shutdown
            exit
            service-name "VPLS1_PE-1"
            sap 1/1/4:1.0 create
            exit
            no shutdown
        exit
```

The **bgp** context specifies parameters which are valid for all of the VPLS BGP applications, such as BGP-multi-homing, BGP-auto-discovery, and BGP-VPLS.

Within the **bgp** context, parameters are configured that are used by neighboring PEs to determine membership of a VPLS instance, such as the auto-discovery of PEs containing the same VPLS instance; the route-distinguisher is configured, along with the route target extended communities.

Route target communities are used to determine membership of a VPLS instance. The import route target at the BGP level is mandatory. The pseudowire template bind is then applied by the service manager on the received routes matching the route target value.

Within the **bgp-vpls** context, the signaling parameters are configured. These determine the service labels required for the data plane of the VPLS instance.

The VPLS edge ID (ve-id) is a numerical value assigned to each PE within a VPLS instance. This value should be unique for a given VPLS instance; no two PEs within the same instance should have the same ve-id values.

A more specific route target can be applied to a pseudowire template in order to define a specific pseudowire topology, rather than only a full mesh, using the command within the **bgp** context:

pw-template *template-id* [**split-horizon-group** *groupname*] [**import-rt** *import-rt-value* (up to 5 max)]

Changes to the import policies are not taken once the pseudowire has been set up (changes on route-target are refreshed though). Pseudowire templates can be re-evaluated with the command **tools perform service eval-pw-template**. The **eval-pw-template** command checks whether all the bindings using this pseudowire template policy are still meant to use this policy.

If the policy has changed and **allow-service-impact** is true, then the old binding is removed and it is re-added with the new template.

VE-ID and BGP Label Allocations

The choice of ve-id is crucial in ensuring efficient allocation of de-multiplexer labels. The most efficient choice is for ve-ids to be allocated starting at 1 and incrementing for each PE as the following section explains.

The **max-ve-id** *value* determines the range of the ve-id value that can be configured. If a PE receives a BGP-VPLS update containing a ve-id with a greater value than the configured **max-ve-id**, then the update is dropped and no service labels are installed for this ve-id.

The **max-ve-id** command also checks the locally-configured ve-id, and prevents a higher value from being used.

Each PE allocates blocks of labels per VPLS instance to remote PEs, in increments of eight labels. It achieves this by advertising three parameters in a BGP update message,

- A label base (LB) which is the lowest label in the block
- A VE Block Size (VBS) which is always eight labels, and cannot be changed
- A VE Base Offset (VBO).

This defines a block of labels in the range (LB, LB+1, ..., LB+VBS-1).

As an example, if the label base (LB) = 262128, then the range for the block is 262128 to 262135, which is exactly eight labels, as per the block size. (The last label in the block is calculated as 262128+8-1 = 262135)

The label allocated by the PE to each remote PE within the VPLS is chosen from this block and is determined by its ve-id. In this way, each remote PE has a unique de-multiplexer label for that VPLS.

To reduce label wastage, contiguous ve-ids in the range (N..N+7) per VPLS should be chosen, where N>0.

Assuming a collection of PEs with contiguous ve-ids, the following labels will be chosen by PEs from the label block allocated by PE-1 which has a ve-id =1.

Table 2 **VE-IDs and Labels**

VE-ID	Label
2	262129
3	262130
4	262131

Table 2 **VE-IDs and Labels (Continued)**

VE-ID	Label
5	262132
6	262133
7	262434
8	262135

This shows that the label allocated to a given PE is (LB+veid-1). The “1” is the VE block offset (VBO).

This means that the label allocated to a PE router within the VPLS can now be written as (LB + veid - VBO), which means that (ve-id - VBO) calculation must always be at least zero and be less than the block size, which is always 8.

For ve-id ≤ 8, a label will be allocated from this block.

For the next block of 8 ve-ids (ve-id 9 to ve-id 16) a new block of 8 labels must be allocated, so a new BGP update is sent, with a new label base, and a block offset of 9.

Table 3 shows how the choice of ve-ids can affect the number of label blocks allocated, and therefore the number of labels:

Table 3 **VE-IDs and Number of Labels**

VE-ID	Block Offset	Labels Allocated
1-8	1	8
9-16	9	8
17-24	17	8
25-32	25	8
33-40	33	8
41-48	41	8
49-56	49	8

This shows that the most efficient use of labels occurs when the ve-ids for a set of PEs are chosen from the same block offset.

If ve-ids are chosen that map to different block offsets, then each PE will have to send multiple BGP updates to signal service labels. Each PE sends label blocks in BGP updates to each of its BGP neighbors for all label blocks in which at least one ve-id has been seen by this PE (it does not advertise label blocks which do not contain an active ve-id, where active ve-id means the ve-id of this PE or any other PE in this VPLS).

The **max-ve-id** must be configured first, and determines the maximum value of the ve-id that can be configured within the PE. The ve-id value cannot be higher than this within the PE configuration, ve-id <= max-ve-id. Similarly, if the ve-id within a received NLRI is higher than the **max-ve-id** *value*, it will not be accepted as valid consequently the max-ve-id configured on all PEs must be greater than or equal to any ve-id used in the VPLS.

Only one ve-id value can be configured. If the ve-id value is changed, BGP withdraws the NLRI and sends a route-refresh.

If the same ve-id is used in different PEs for the same VPLS, a Designated Forwarder election takes place.

Executing the **shutdown** command triggers an MP-UNREACH-NLRI from the PE to all BGP peers.

The **no shutdown** command triggers an MP-REACH-NLRI to the same peers.

PE-2 Service Creation

On PE-2, a VPLS service using pseudowire template 1 is created. In order to make the label allocation more efficient, PE-2 has been allocated a ve-id value of 2. For completeness, the pseudowire template is also shown.

```
*A:PE-2# configure
    service
        pw-template 1 create
            split-horizon-group "VPLS-SHG"
            exit
        exit
        vpls 1 customer 1 create
            bgp
                route-distinguisher 65536:1
                route-target export target:65536:1 import target:65536:1
                pw-template-binding 1
                exit
            exit
            bgp-vpls
                max-ve-id 10
                ve-name "PE-2"
                    ve-id 2
                exit
                no shutdown
            exit
            service-name "VPLS1_PE-2"
            sap 1/1/4:1.0 create
            exit
            no shutdown
        exit
```

The **max-ve-id** *value* is set to 10 to allow an increase in the number of PEs that could be a part of this VPLS instance.

PE-3 Service Creation

The following configuration creates a VPLS instance on PE-3, using a ve-id value of 3.

```
*A:PE-3# configure
    service
        pw-template 1 create
            split-horizon-group "VPLS-SHG"
            exit
        exit
        vpls 1 customer 1 create
            bgp
                route-distinguisher 65536:1
                route-target export target:65536:1 import target:65536:1
                pw-template-binding 1
                exit
            exit
            bgp-vpls
                max-ve-id 10
                ve-name "PE-3"
                    ve-id 3
                exit
                no shutdown
            exit
            service-name "VPLS1_PE-3"
            sap 1/1/4:1.0 create
            exit
            no shutdown
        exit
```

PE-1 Service Operation Verification

The following command shows that the BGP-VPLS site is enabled on PE-1.

```
*A:PE-1# show service id 1 bgp-vpls

===============================================================================
BGP VPLS Information
```

```
===============================================================================
Max Ve Id            : 10                   Admin State      : Enabled
VE Name              : PE-1                 VE Id            : 1
PW Tmpl used         : 1
===============================================================================
*A:PE-1#
```

The following command shows that the service is operationally up on PE-1:

```
*A:PE-1# show service id 1 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 1                   Vpn Id            : 0
Service Type      : VPLS
Name              : VPLS1_PE-1
---snip---

Admin State       : Up                  Oper State        : Up
MTU               : 1514                Def. Mesh VC Id   : 1
SAP Count         : 1                   SDP Bind Count    : 2
---snip---

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/4:1.0                            qinq         1522    1522    Up   Up
sdp:17406:4294967294 SB(192.0.2.3)       BgpVpls      0       1556    Up   Up
sdp:17407:4294967295 SB(192.0.2.2)       BgpVpls      0       1556    Up   Up
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-1#
```

The SAP and SDPs are all operationally up. The **SB** flags for the SDPs signify Spoke and BGP.

The ingress labels for PE-2 and PE-3—the labels allocated by PE-1—can be seen as follows:

```
*A:PE-1# show service id 1 sdp

===============================================================================
Services: Service Destination Points
===============================================================================
SdpId            Type     Far End addr    Adm     Opr       I.Lbl     E.Lbl
-------------------------------------------------------------------------------
17406:4294967294 BgpVpls  192.0.2.3       Up      Up        262130    262128
17407:4294967295 BgpVpls  192.0.2.2       Up      Up        262129    262126
-------------------------------------------------------------------------------
Number of SDPs : 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

As can be seen from the following output, a BGP-VPLS NLRI update is sent to the route reflector (192.0.2.7) and is received by each PE.

The following debug trace from PE-1 shows the BGP NLRI update for VPLS 1 sent by PE-1 to the route reflector.

```
*A:PE-1# debug router bgp update

1 2017/04/26 12:17:27.18 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.7
"Peer 1: 192.0.2.7: UPDATE
Peer 1: 192.0.2.7 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 72
    Flag: 0x90 Type: 14 Len: 28 Multiprotocol Reachable NLRI:
        Address Family L2VPN
        NextHop len 4 NextHop 192.0.2.1
        [VPLS/VPWS] preflen 17, veid: 1, vbo: 1, vbs: 8, label-base: 262128,
                    RD 65536:1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:65536:1
        l2-vpn/vrf-imp:Encap=19: Flags=none: MTU=1514: PREF=0
"
```

The control flags within the extended community indicate the status of the VPLS instance.

The control flag D indicates that all attachment circuits are Down, or the VPLS is shut down. The flags are used in BGP Multi-homing when determining which PEs are designated forwarders, see chapter BGP Multi-Homing for VPLS Networks.

When flags=none, then all attachment circuits are up. In the preceding example, no flags are present, but should all SAPs become operationally down, then the control flag D would be seen in the debug message. To simulate this, the SAP 1/1/4:1 is disabled:

```
*A:PE-1# configure service vpls 1 sap 1/1/4:1.0 shutdown
```

All SAPs in VPLS 1 on PE-1 are operationally down, so PE-1 sends a BGP update message with control flag D set, as follows:

```
5 2017/04/26 12:21:31.19 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.7
"Peer 1: 192.0.2.7: UPDATE
Peer 1: 192.0.2.7 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 72
    Flag: 0x90 Type: 14 Len: 28 Multiprotocol Reachable NLRI:
        Address Family L2VPN
        NextHop len 4 NextHop 192.0.2.1
       [VPLS/VPWS] preflen 17, veid: 1, vbo: 1, vbs: 8, label-base: 262128, RD 65536:1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:65536:1
        l2-vpn/vrf-imp:Encap=19: Flags=D: MTU=1514: PREF=0
```

The SAP is re-enabled with the following command on PE-1:

```
*A:PE-1# configure service vpls 1 sap 1/1/4:1.0 no shutdown
```

The BGP VPLS signaling parameters are also present in the BGP update message, namely the ve-id of the PE within the VPLS instance, the VBO and VBS, and the label base. The target indicates the VPLS instance, which must be matched against the import route targets of the receiving PEs.

The signaling parameters can be seen within the BGP update with following command:

```
*A:PE-1# show router bgp routes l2-vpn rd 65536:1 hunt
===============================================================================
 BGP Router ID:192.0.2.1        AS:65536       Local AS:65536
===============================================================================
---snip---

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
Route Type     : VPLS
Route Dist.    : 65536:1
VeId           : 1                      Block Size     : 8
Base Offset    : 1                      Label Base     : 262128
Nexthop        : 192.0.2.1
To             : 192.0.2.7
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : target:65536:1
                 l2-vpn/vrf-imp:Encap=19: Flags=none: MTU=1514: PREF=0
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.7
Origin         : IGP
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
```

```
Source Class   : 0                      Dest Class     : 0

-------------------------------------------------------------------------------
Routes : 4
===============================================================================
*A:PE-1#
```

In this configuration example, PE-1 (192.0.2.1) with ve-id =1 has sent an update with base offset (VBO) =1, block size (VBS) = 8, and label base 262128. This means that labels 262128 (LB) to 262135 (LB+VBS-1) are available as de-multiplexer labels, egress labels to be used to reach PE-1 for VPLS 1.

PE-2 receives this update from PE-1. This is seen as a valid VPLS BGP route from PE-1 through the route reflector with nexthop 192.0.2.1.

```
*A:PE-2# show router bgp routes l2-vpn rd 65536:1
===============================================================================
 BGP Router ID:192.0.2.2        AS:65536       Local AS:65536
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP L2VPN Routes
===============================================================================
Flag  RouteType                   Prefix                             MED
      RD                          SiteId                             Label
      Nexthop                     VeId                   BlockSize   LocalPref
      As-Path                     BaseOffset             vplsLabelBa
                                                         se
-------------------------------------------------------------------------------
u*>i  VPLS                      -                        -           0
      65536:1                   -                                    -
      192.0.2.1                   1                      8           100
      No As-Path                  1                      262128
i     VPLS                      -                        -           0
      65536:1                   -                                    -
      192.0.2.2                   2                      8           100
      No As-Path                  1                      262126
u*>i  VPLS                      -                        -           0
      65536:1                   -                                    -
      192.0.2.3                   3                      8           100
      No As-Path                  1                      262128
-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:PE-2#
```

PE-2 uses this information in conjunction with its own ve-id to calculate the egress label toward PE-1, using the condition VBO ≤ ve-id < (VBO+VBS).

The ve-id of PE-2 is in the Label Block covered by VBO =1, thus,

Label calculation = label base + local ve-id - Base offset
= 262128 + 2 - 1
Egress label used = 262129

This is verified using the following command on PE-2 where the egress label toward PE-1 (192.0.2.1) is 262129.

```
*A:PE-2# show service id 1 sdp

===============================================================================
Services: Service Destination Points
===============================================================================
SdpId            Type     Far End addr    Adm     Opr       I.Lbl     E.Lbl
-------------------------------------------------------------------------------
17406:4294967294 BgpVpls  192.0.2.3       Up      Up        262128    262129
17407:4294967295 BgpVpls  192.0.2.1       Up      Up        262126    262129
-------------------------------------------------------------------------------
Number of SDPs : 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-2#
```

PE-3 also receives this update from PE-1 by the route reflector. This is seen as a valid VPLS BGP route from PE-1 with nexthop 192.0.2.1.

```
*A:PE-3# show router bgp routes l2-vpn rd 65536:1
===============================================================================
 BGP Router ID:192.0.2.3       AS:65536       Local AS:65536
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP L2VPN Routes
===============================================================================
Flag  RouteType                   Prefix                                MED
      RD                          SiteId                                Label
      Nexthop                     VeId                   BlockSize      LocalPref
      As-Path                     BaseOffset             vplsLabelBa
                                                         se
-------------------------------------------------------------------------------
u*>i  VPLS                        -                      -           0
      65536:1                     -                                  -
      192.0.2.1                   1                      8           100
      No As-Path                  1                      262128
u*>i  VPLS                        -                      -           0
      65536:1                     -                                  -
      192.0.2.2                   2                      8           100
      No As-Path                  1                      262126
i     VPLS                        -                      -           0
      65536:1                     -                                  -
      192.0.2.3                   3                      8           100
      No As-Path                  1                      262128
-------------------------------------------------------------------------------
```

```
Routes : 3
===============================================================================
*A:PE-3#
```

The ve-id of PE-3 is also in the label block covered by block offset VBO =1.

Label calculation= label base + local ve-id - VBO
= 262128 + 3 - 1
Egress label used = 262130

This is verified using the following command on PE-3 where egress label toward 192.0.2.1 is 262130.

```
*A:PE-3# show service id 1 sdp

===============================================================================
Services: Service Destination Points
===============================================================================
SdpId            Type     Far End addr    Adm     Opr       I.Lbl     E.Lbl
-------------------------------------------------------------------------------
17406:4294967294 BgpVpls  192.0.2.1       Up      Up        262128    262130
17407:4294967295 BgpVpls  192.0.2.2       Up      Up        262129    262128
-------------------------------------------------------------------------------
Number of SDPs : 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-3#
```

PE-2 Service Operation Verification

For completeness, verify the service is operationally up on PE-2.

```
*A:PE-2# show service id 1 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 1                    Vpn Id            : 0
Service Type      : VPLS
Name              : VPLS1_PE-2
---snip---
Admin State       : Up                   Oper State        : Up
MTU               : 1514                 Def. Mesh VC Id   : 1
SAP Count         : 1                    SDP Bind Count    : 2
---snip---
-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                                Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/4:1.0                             qinq         1522    1522    Up   Up
sdp:17406:4294967294 SB(192.0.2.3)        BgpVpls      0       1556    Up   Up
sdp:17407:4294967295 SB(192.0.2.1)        BgpVpls      0       1556    Up   Up
```

```
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-2#
```

PE-2 De-Multiplexer Label Calculation

In the same way that PE-1 allocates a label base (LB), block size (VBS), and base offset (VBO), PE-2 also allocates the same parameters for PE-1 and PE-3 to calculate the egress service label required to reach PE-2.

```
*A:PE-2# show router bgp routes l2-vpn rd 65536:1 hunt
===============================================================================
 BGP Router ID:192.0.2.2        AS:65536       Local AS:65536
===============================================================================
---snip---
===============================================================================
BGP L2VPN Routes
===============================================================================

---snip---
-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
Route Type     : VPLS
Route Dist.    : 65536:1
VeId           : 2                      Block Size     : 8
Base Offset    : 1                      Label Base     : 262126
Nexthop        : 192.0.2.2
To             : 192.0.2.7
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : target:65536:1
                 l2-vpn/vrf-imp:Encap=19: Flags=none: MTU=1514: PREF=0
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.7
Origin         : IGP
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
-------------------------------------------------------------------------------
Routes : 4
===============================================================================
*A:PE-2#
```

This is verified using the following command on PE-1 to show the egress label toward PE-2 (192.0.2.2) where the egress label toward PE-2 = 262126 + 1- 1 = 262126.

```
*A:PE-1# show service id 1 sdp
```

```
===============================================================================
Services: Service Destination Points
===============================================================================
SdpId            Type     Far End addr    Adm     Opr       I.Lbl     E.Lbl
-------------------------------------------------------------------------------
17406:4294967294 BgpVpls  192.0.2.3       Up      Up        262130    262128
17407:4294967295 BgpVpls  192.0.2.2       Up      Up        262129    262126
-------------------------------------------------------------------------------
Number of SDPs : 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

This is also verified using the following command on PE-3 to show the egress label toward PE-2 (192.0.2.2) where the egress label toward PE-2 = 262126 + 3 - 1 = 262128.

```
*A:PE-3# show service id 1 sdp

===============================================================================
Services: Service Destination Points
===============================================================================
SdpId            Type     Far End addr    Adm     Opr       I.Lbl     E.Lbl
-------------------------------------------------------------------------------
17406:4294967294 BgpVpls  192.0.2.1       Up      Up        262128    262130
17407:4294967295 BgpVpls  192.0.2.2       Up      Up        262129    262128
-------------------------------------------------------------------------------
Number of SDPs : 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-3#
```

PE-3 Service Operation Verification

Verify that the service is operationally up on PE-3:

```
*A:PE-3# show service id 1 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 1                   Vpn Id            : 0
Service Type      : VPLS
Name              : VPLS1_PE-3
---snip---
Admin State       : Up                  Oper State        : Up
MTU               : 1514                Def. Mesh VC Id   : 1
SAP Count         : 1                   SDP Bind Count    : 2
---snip---
-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
```

```
-------------------------------------------------------------------------------
sap:1/1/4:1.0                            qinq         1522    1522    Up   Up
sdp:17406:4294967294 SB(192.0.2.1)       BgpVpls      0       1556    Up   Up
sdp:17407:4294967295 SB(192.0.2.2)       BgpVpls      0       1556    Up   Up
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-3#

*A:PE-3# show service id 1 sdp

===============================================================================
Services: Service Destination Points
===============================================================================
SdpId            Type     Far End addr    Adm     Opr       I.Lbl     E.Lbl
-------------------------------------------------------------------------------
17406:4294967294 BgpVpls  192.0.2.1       Up      Up        262128    262130
17407:4294967295 BgpVpls  192.0.2.2       Up      Up        262129    262128
-------------------------------------------------------------------------------
Number of SDPs : 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-3#
```

PE-3 De-Multiplexer Label Verification

PE-3 also allocates the required parameters for PE-1 and PE-2 to calculate the egress service label required to reach PE-3.

This is verified using the following command on PE-1 to show the egress label toward PE-3 (192.0.2.3) (262128) where egress label toward PE-2 = 262126. The Label Base equals 262128 on PE-3 and 262126 on PE-2.

```
*A:PE-1# show service id 1 sdp

===============================================================================
Services: Service Destination Points
===============================================================================
SdpId            Type     Far End addr    Adm     Opr       I.Lbl     E.Lbl
-------------------------------------------------------------------------------
17406:4294967294 BgpVpls  192.0.2.3       Up      Up        262130    262128
17407:4294967295 BgpVpls  192.0.2.2       Up      Up        262129    262126
-------------------------------------------------------------------------------
Number of SDPs : 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

This is also verified using the following command on PE-2 to show the egress label toward PE-3 (192.0.2.3) which is using auto-provisioned SDP 17406.

```
*A:PE-2# show service id 1 sdp

===============================================================================
Services: Service Destination Points
===============================================================================
SdpId            Type      Far End addr    Adm     Opr       I.Lbl     E.Lbl
-------------------------------------------------------------------------------
17406:4294967294 BgpVpls  192.0.2.3       Up      Up        262128    262129
17407:4294967295 BgpVpls  192.0.2.1       Up      Up        262126    262129
-------------------------------------------------------------------------------
Number of SDPs : 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-2#
```

This example has shown that for VPLS instance with 3 PEs, not all labels allocated by a PE will be used by remote PEs as de-multiplexer service labels. There will be some wastage of label space, so there is a necessity to choose ve-ids that keep this waste to a minimum.

The next example will show an even more wasteful use of labels by using a random choice of ve-ids.

BGP VPLS Using Pre-Provisioned SDP

It is possible to configure BGP-VPLS instances that use RSVP-TE transport tunnels. In this case, the SDP must be created with the MPLS LSPs mapped and with signaling set to BGP, as the service labels are signaled using BGP. The pseudowire template configured within the BGP-VPLS instance must use the **use-provisioned-sdp** keyword.

This example also examines the effect of using ve-ids that are not all within the same contiguous block.

Figure 31 **BGP VPLS Using Pre-Provisioned SDP**

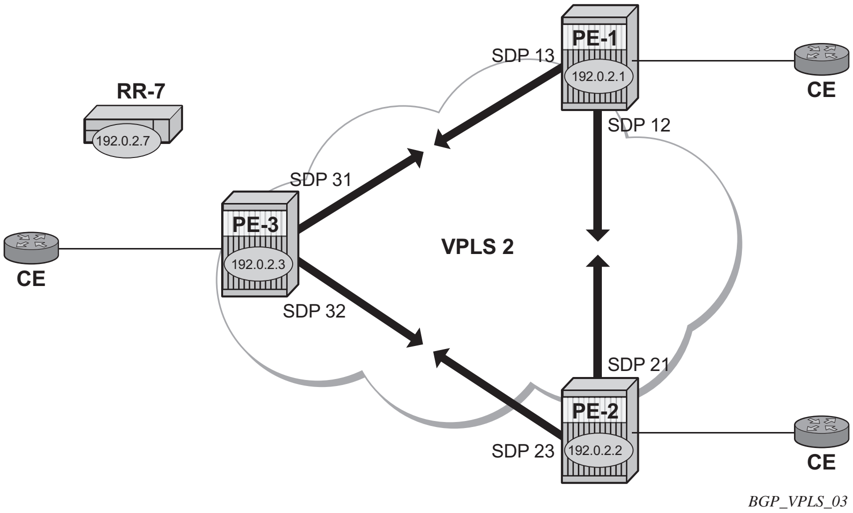


Figure 31 shows an example of a VPLS instance where SDPs are pre-provisioned with RSVP-TE signaled transport tunnels.

SDPs on PE-1

```
*A:PE-1# configure
    service
        sdp 12 mpls create
            description "SDP-PE-1-PE-2_RSVP_BGP"
            signaling bgp
            far-end 192.0.2.2
            lsp "LSP-PE-1-PE-2"
            no shutdown
        exit
        sdp 13 mpls create
            description "SDP-PE-1-PE-3_RSVP_BGP"
            signaling bgp
            far-end 192.0.2.3
            lsp "LSP-PE-1-PE-3"
            no shutdown
        exit
```

SDPs on PE-2

```
*A:PE-2# configure
    service
        sdp 21 mpls create
            description "SDP-PE-2-PE-1_RSVP_BGP"
            signaling bgp
            far-end 192.0.2.1
            lsp "LSP-PE-2-PE-1"
            no shutdown
        exit
        sdp 23 mpls create
            description "SDP-PE-2-PE-3_RSVP_BGP"
            signaling bgp
            far-end 192.0.2.3
            lsp "LSP-PE-2-PE-3"
            no shutdown
        exit
```

SDPs on PE-3

```
*A:PE-3# configure
    service
        sdp 31 mpls create
            description "SDP-PE-3-PE-1_RSVP_BGP"
            signaling bgp
            far-end 192.0.2.1
            lsp "LSP-PE-3-PE-1"
            no shutdown
        exit
        sdp 32 mpls create
            description "SDP-PE-3-PE-2_RSVP_BGP"
            signaling bgp
            far-end 192.0.2.2
            lsp "LSP-PE-3-PE-2"
            no shutdown
        exit
```

Pre-provisioned BGP-SDPs can also be used with BGP-VPLS. For reference, they are configured as follows:

```
*A:PE-3# configure
    service
        sdp 332 mpls create
            signaling bgp
            far-end 192.0.2.2
            no shutdown
        exit
```

To create an SDP within a service that uses the RSVP transport tunnel, a pseudowire template is required that has the **use-provisioned-sdp** parameter set. It is also possible to configure the **prefer-provisioned-sdp** parameter, see chapter LDP VPLS Using BGP Auto-Discovery - Prefer Provisioned SDP.

Once again, a split horizon group is included to prevent forwarding between pseudowires.

The pseudowire template must be provisioned on all PEs and looks like:

```
*A:PE-1# configure
    service
        pw-template 2 use-provisioned-sdp create
            split-horizon-group "VPLS-SHG"
            exit
        exit
```

The following output shows the configuration required for a BGP-VPLS service using a pseudowire template configured for using pre-provisioned RSVP-TE SDPs.

```
*A:PE-1# configure
    service
        vpls 2 customer 1 create
            bgp
                route-distinguisher 65536:2
                route-target export target:65536:2 import target:65536:2
                pw-template-binding 2
                exit
            exit
            bgp-vpls
                max-ve-id 100
                ve-name "PE-1"
                    ve-id 1
                exit
                no shutdown
            exit
            sap 1/1/4:2.0 create
            exit
            no shutdown
        exit
```

The route distinguisher and route target extended community values for VPLS 2 are different from the ones in VPLS 1. The ve-id value for PE-1 can be the same as the one in VPLS 1, but these must be different within the same VPLS instance on the other PEs — PE-2 should not have ve-id = 1.

On PE-2, the configuration is as follows with the ve-id value equal to 20, which will result in a label from a different block:

```
*A:PE-2# configure
    service
        vpls 2 customer 1 create
            bgp
                route-distinguisher 65536:2
                route-target export target:65536:2 import target:65536:2
                pw-template-binding 2
                exit
            exit
            bgp-vpls
                max-ve-id 100
                ve-name "PE-2"
                    ve-id 20
                exit
                no shutdown
            exit
            sap 1/1/4:2.0 create
            exit
            no shutdown
        exit
```

and on PE-3:

```
*A:PE-3# configure
    service
        vpls 2 customer 1 create
            bgp
                route-distinguisher 65536:2
                route-target export target:65536:2 import target:65536:2
                pw-template-binding 2
                exit
            exit
            bgp-vpls
                max-ve-id 100
                ve-name "PE-3"
                    ve-id 3
                exit
                no shutdown
            exit
            sap 1/1/4:2.0 create
            exit
            no shutdown
        exit
```

Verify that the service is operationally up on PE-1.

```
*A:PE-1# show service id 2 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 2                   Vpn Id            : 0
Service Type      : VPLS
---snip---
Admin State       : Up                  Oper State        : Up
MTU               : 1514                Def. Mesh VC Id   : 2
SAP Count         : 1                   SDP Bind Count    : 2
---snip---
-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/4:2.0                            qinq         1522    1522    Up   Up
sdp:12:4294967292 S(192.0.2.2)           BgpVpls      0       1556    Up   Up
sdp:13:4294967293 S(192.0.2.3)           BgpVpls      0       1556    Up   Up
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-1#
```

The SDP 12 and 13 are the pre-provisioned SDPs.

For completeness, verify the service is operationally up on PE-2

```
*A:PE-2# show service id 2 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 2                   Vpn Id            : 0
Service Type      : VPLS
---snip---
Admin State       : Up                Oper State        : Up
MTU               : 1514              Def. Mesh VC Id   : 2
SAP Count         : 1                 SDP Bind Count    : 2
---snip---
-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/4:2.0                            qinq         1522    1522    Up   Up
sdp:21:4294967293 S(192.0.2.1)           BgpVpls      0       1556    Up   Up
sdp:23:4294967292 S(192.0.2.3)           BgpVpls      0       1556    Up   Up
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-2#
```

Verify the service is operational on PE-3:

```
*A:PE-3# show service id 2 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 2                 Vpn Id            : 0
Service Type      : VPLS
---snip---
Admin State       : Up                Oper State        : Up
MTU               : 1514              Def. Mesh VC Id   : 2
SAP Count         : 1                 SDP Bind Count    : 2
---snip---
-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/4:2.0                            qinq         1522    1522    Up   Up
sdp:31:4294967293 S(192.0.2.1)           BgpVpls      0       1556    Up   Up
sdp:32:4294967292 S(192.0.2.2)           BgpVpls      0       1556    Up   Up
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-3#
```

PE-1 De-Multiplexer Label Calculation

In the case of VPLS 1, all ve-ids are in the range of a single label block. In the case of VPLS 2, the ve-ids are in different blocks, for example, the ve-id 20 is in a different block to ve-ids 1 and 3.

As the label allocation is block-dependent, multiple label blocks must be advertised by each PE to encompass this.

Consider PE-1’s BGP update NLRIs.

```
*A:PE-1# show router bgp routes l2-vpn rd 65536:2 hunt
===============================================================================
 BGP Router ID:192.0.2.1        AS:65536       Local AS:65536
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP L2VPN Routes
===============================================================================
---snip---
-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
Route Type     : VPLS
Route Dist.    : 65536:2
VeId           : 1                      Block Size     : 8
Base Offset    : 1                      Label Base     : 262120
Nexthop        : 192.0.2.1
To             : 192.0.2.7
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : target:65536:2
                 l2-vpn/vrf-imp:Encap=19: Flags=none: MTU=1514: PREF=0
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.7
Origin         : IGP
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0

Route Type     : VPLS
Route Dist.    : 65536:2
VeId           : 1                      Block Size     : 8
Base Offset    : 17                     Label Base     : 262112
Nexthop        : 192.0.2.1
To             : 192.0.2.7
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : target:65536:2
                 l2-vpn/vrf-imp:Encap=19: Flags=none: MTU=1514: PREF=0
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.7
Origin         : IGP
AS-Path        : No As-Path
```

```
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
-------------------------------------------------------------------------------
Routes : 8
===============================================================================
*A:PE-1#
```

Two NLRIs updates are sent to the route reflector, with the following label parameters:

1. LB = 262120, VBS = 8, VBO = 1

2. LB = 262112, VBS = 8, VBO = 17

PE-2 has a ve-id of 20. Applying the condition VBO ≤ ve-id < (VBO+VBS)

- Update 1: LB = 262120, VBS = 8, VBO = 1
- VBO ≤ ve-id for ve-id = 20 is true
- ve-id < (VBO+VBS) for ve-id = 20 is false.
- PE-2 cannot choose a label from this block.
- Update 2: LB = 262112, VBS = 8, VBO = 17
- VBO ≤ ve-id for ve-id = 20 is true
- ve-id < (VBO+VBS) for ve-id = 20 is true.
- PE-2 chooses label 262112 + 20 - 17 = 262115 (LB + veid - VBO)

The egress label chosen is verified by examining the egress label toward PE-1 (192.0.2.1) on PE-2.

```
*A:PE-2# show service id 2 sdp

===============================================================================
Services: Service Destination Points
===============================================================================
SdpId             Type     Far End addr    Adm     Opr       I.Lbl     E.Lbl
-------------------------------------------------------------------------------
21:4294967293     BgpVpls  192.0.2.1       Up      Up        262110    262115
23:4294967292     BgpVpls  192.0.2.3       Up      Up        262112    262115
-------------------------------------------------------------------------------
Number of SDPs : 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-2#
```

PE-3 has a ve-id of 3. Applying the condition VBO ≤ ve-id < (VBO+VBS)

- Update 1: LB = 262120, VBS = 8, VBO = 1
- VBO ≤ ve-id for ve-id = 3 is true

- ve-id < (VBO+VBS) for ve-id = 3 is true.
- PE-3 chooses label 262120 + 3 - 1 = 262122 (LB + veid - VBO)
- Update 2: LB = 262112, VBS = 8, VBO = 17
- VBO ≤ ve-id for ve-id = 3 is false
- ve-id < (VBO+VBS) for ve-id = 3 is true.
- PE-3 cannot choose a label from this block.

The egress label chosen is verified by examining the egress label toward PE-1 (192.0.2.1) on PE-3.

```
*A:PE-3# show service id 2 sdp

===============================================================================
Services: Service Destination Points
===============================================================================
SdpId            Type     Far End addr    Adm     Opr       I.Lbl     E.Lbl
-------------------------------------------------------------------------------
31:4294967293    BgpVpls  192.0.2.1       Up      Up        262120    262122
32:4294967292    BgpVpls  192.0.2.2       Up      Up        262115    262112
-------------------------------------------------------------------------------
Number of SDPs : 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-3#
```

To illustrate the allocation of label blocks by a PE, against the actual use of the same labels, consider the following. When BGP updates from each PE signal the multiplexer labels in blocks of eight, the allocated label values are added to the in-use pool. First check what label range can be allocated dynamically.

```
*A:PE-1# show router mpls-labels label-range

============================================================================
Label Ranges
============================================================================
Label Type      Start Label End Label   Aging       Available   Total
----------------------------------------------------------------------------
Static          32          18431       -           18400       18400
Dynamic         18432       524287      0           505824      505856
    Seg-Route   0           0           -           0           505856
============================================================================
*A:PE-1#
```

Verify which labels in the dynamic range are in use. The label pool of PE-1 can be verified as per the following output which shows labels used along with the associated protocol:

```
*A:PE-1# show router mpls-labels label 18432 524287 in-use
=================================================================
MPLS Labels from 18432 to 524287 (In-use)
=================================================================
Label               Label Type          Label Owner
-----------------------------------------------------------------
262110              dynamic             ILDP
262111              dynamic             ILDP
262112              dynamic             BGP
262113              dynamic             BGP
262114              dynamic             BGP
262115              dynamic             BGP
262116              dynamic             BGP
262117              dynamic             BGP
262118              dynamic             BGP
262119              dynamic             BGP
262120              dynamic             BGP
262121              dynamic             BGP
262122              dynamic             BGP
262123              dynamic             BGP
262124              dynamic             BGP
262125              dynamic             BGP
262126              dynamic             BGP
262127              dynamic             BGP
262128              dynamic             BGP
262129              dynamic             BGP
262130              dynamic             BGP
262131              dynamic             BGP
262132              dynamic             BGP
262133              dynamic             BGP
262134              dynamic             BGP
262135              dynamic             BGP
262136              dynamic             ILDP
262137              dynamic             ILDP
262140              dynamic             RSVP
262141              dynamic             RSVP
262142              dynamic             ILDP
262143              dynamic             ILDP
-----------------------------------------------------------------
In-use labels (Owner: All) in specified range   : 32
In-use labels in entire range                   : 32
=================================================================
*A:PE-1#
```

This shows that 24 labels have been allocated for use by BGP. Of this number, 16 labels have been allocated for use by PEs within VPLS 2 to communicate with PE-1, the blocks with label base 262112 and with label base 262120.

There are only two neighboring PEs within this VPLS instance, so only two labels will ever be used in the data plane for traffic destined to PE-1. These are 262115 and 262122. The remaining labels have no PE with the associated ve-id that can use them.

Once again, this case emphasizes that to reduce label wastage, contiguous ve-ids in the range (N..N+7) per VPLS should be chosen, where N>0.

Conclusion

BGP-VPLS allows the delivery of Layer 2 VPN services to customers where BGP is commonly used. The examples presented in this chapter show the configuration of BGP-VPLS together with the associated show outputs which can be used for verification and troubleshooting.

Black-hole MAC for EVPN Loop Protection

This chapter provides information about Black-hole MAC for EVPN Loop Protection.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

The information and configuration in this chapter are based on SR OS Release 15.0.R4. Black-hole MAC for EVPN loop protection is supported in SR OS Release 15.0.R1, and later.

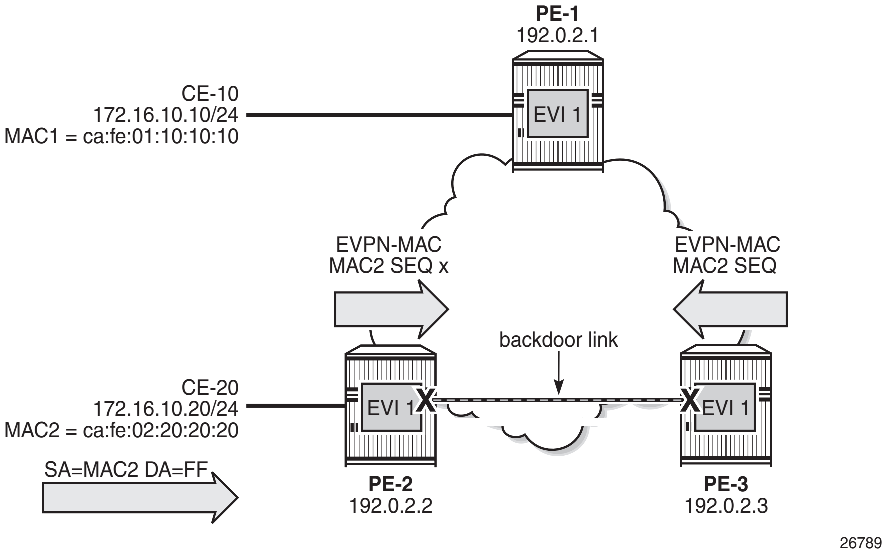
Chapters Auto-Learn MAC Protect in EVPN and Conditional Static Black-Hole MAC in EVPN are prerequisite reading.

Overview

Service providers are migrating VPLS networks to EVPN and require the same or better loop protection mechanisms, such as **mac-move** or **auto-learn-mac-protect** (ALMP). Chapter Auto-Learn MAC Protect in EVPN describes how traffic is protected in "static" networks, where the CEs do not move to a different port or PE, and MAC addresses are always learned first on the correct SAP/SDP-bindings. However, ALMP does not provide a loop protection solution in EVPN networks that require mobility and ALMP has issues with all-active multi-homing. Since mobility and all-active multi-homing are two of the key advantages of EVPN compared to VPLS, an alternate loop protection mechanism is required. This chapter describes an example for the black-hole based loop protection solution, based on *draft-snr-bess-evpn-loop-protect*.

Figure 32 shows a topology using black-hole MAC for EVPN loop protection.

Figure 32 **Black-hole MAC for EVPN Loop Protection**



VPLS 1 with EVI 1 is configured on all PEs. A backdoor link exists between PE-2 and PE-3 (in this case, caused by misconfiguration: additional SAPs are configured in VPLS 1). When CE-20 sends Broadcast, Unknown unicast, or Multicast (BUM) traffic, its source address MAC2 is learned by PE-2, which sends an EVPN-MAC route for MAC2 to its BGP peers. PE-2 floods the frame to its EVPN-MPLS destinations (PE-1 and PE-3) as well as its local SAPs (including the backdoor link to PE-3).

PE-3 receives the EVPN-MAC route from PE-2, but due to the backdoor link, it also learns MAC2 on its local SAP. Following the MAC mobility procedures, PE-3 advertises MAC2 with a higher sequence number to its BGP peers. PE-3 floods the frame to its EVPN-MPLS destinations and to its local SAPs.

Note: The preceding simplified description assumes that PE-3 receives the EVPN-MAC route prior to learning MAC2 from the backdoor link, which may or may not be the case. Regardless of how MAC2 is learned, the MAC duplication procedures are invoked.

PE-2 and PE-3 keep learning and advertising MAC2 until the configured number of MAC moves (**num-moves**) has been reached. Then, MAC2 is detected as duplicate and will not be advertised again until the **retry** interval has expired.

If the **mac-duplication black-hole-dup-mac** option is configured, MAC2 will be added to the FDB as black-hole MAC, so traffic with MAC DA = MAC2 will be discarded. Also, MAC addresses assigned to a black-hole destination are considered as protected, so traffic with MAC SA = MAC2 will not be forwarded due to one of the following reasons:

- When the SAPs/SDP-bindings or BGP-EVPN MPLS/VXLAN destinations are configured with **restrict-protected-source discard-frame** (default), the frames are discarded before any MAC SA is learned or the MAC DA is looked up.
- When the SAP/SDP-binding is configured with **restrict-protected-source**, an incoming frame with MAC SA = black-hole MAC causes the system to bring down the corresponding SAP/SDP-binding.

Assuming PE-3 detects MAC2 as duplicate and installs it as black-hole MAC, PE-3 will discard the broadcast frames with MAC SA = MAC2, so the loop is broken, whereas the legitimate traffic between CE-10 and CE-20 is allowed (assuming PE-2 does not black-hole MAC2).

Black-hole MAC duplication is enabled with the **black-hole-dup-mac** keyword in the **mac-duplication** context, as follows:

```
*A:PE-3# configure service vpls 1 bgp-evpn mac-duplication
  - mac-duplication

 [no] black-hole-dup* - Enable/disable BGP-EVPN black-hole duplicate MAC traffic
      detect          - Configure BGP EVPN Mac Duplication Detection
 [no] retry           - Configure BGP EVPN Mac Duplication Retry

*A:PE-3# configure service vpls 1 bgp-evpn mac-duplication black-hole-dup-mac
```

When enabled, the operation is as follows:

- Each node that learns a MAC address that has been advertised by a BGP peer will send an EVPN-MAC route for that MAC address with a higher sequence number. When the number of MAC moves exceeds the configured threshold (by default, five MAC moves in three minutes), the MAC address is detected as duplicate and no EVPN-MAC routes will be sent for that MAC address until the retry interval (default nine minutes) has elapsed.
- When MAC2 is detected as duplicate, the system will:
  - Add MAC2 to the duplicate MAC list
  - Add MAC2 in the FDB as protected MAC associated with a black-hole endpoint (type **EvpnD:P** and source identifier **black-hole**)
    - Incoming frames with MAC DA = MAC2 will be discarded based on a MAC lookup in the FDB.

- MAC addresses assigned to a black-hole destination are protected and incoming frames with MAC SA = MAC2 will be discarded or the system will bring down the SAP/SDP-binding, depending on the **restrict-protected-src** setting on the SAP/SDP/EVPN endpoint.

The following output shows the FDB with black-hole MAC address ca:fe:02:20:20:20 (type EvpnD:P):

```
*A:PE-3# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         ca:fe:01:10:10:10 eMpls:                   Evpn     08/17/17 07:02:10
                            192.0.2.2:262140
1         ca:fe:02:20:20:20 black-hole               EvpnD:P  08/17/17 07:02:18
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-3#
```

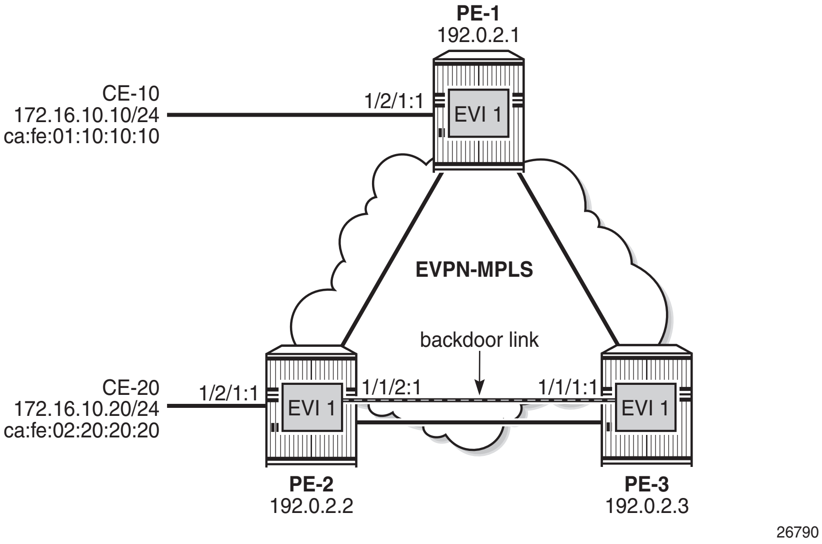
The duplicate MAC address will be removed from the FDB and the process will be restarted in the following cases:

- Retry interval events:
  - When the retry interval expires.
  - When the user configures **no retry** on the service that detected the duplicate MAC address.
- MAC relearning events:
  - When the remote PE withdraws the MAC address (due to aging or **clear service fdb**). Local attempts to clear a black-hole MAC (via **clear service fdb**) will fail because the type of the MAC entry is not "learned", but "EvpnD:P".
  - When configuring a local conditional static MAC address (CStatic:P) prevents the EvpnD:P entry for the same MAC address from being installed in the FDB as black-hole, if the SAP/SDP-binding where the MAC is configured is operationally up.
- CPM switchover event

Configuration

Figure 33 shows the example topology with three PEs and two CEs. A loop will occur when CE-20 sends Broadcast, Unknown unicast, or Multicast (BUM) traffic. Traffic between PE-2 and PE-3 will be sent over the regular router interfaces between the PEs, but also over the backdoor link (SAP 1/1/2:1 in VPLS 1 on PE-2 and SAP 1/1/1:1 in VPLS 1 on PE-3).

Figure 33 **Example Topology**



The initial configuration includes:

- Cards, MDAs, ports
- Router interfaces
- IS-IS on all router interfaces (alternatively, OSPF can be used)
- LDP on all router interfaces

Enable Black-hole MAC Duplication Detection in EVPN

BGP is configured for address family EVPN on all PEs with PE-3 as route reflector. The following is the BGP configuration on PE-3:

```
configure
    router
        autonomous-system 64500
        bgp
            min-route-advertisement 1
            rapid-withdrawal
            split-horizon
            rapid-update evpn
            group "internal"
                family evpn
                cluster 1.1.1.1
                peer-as 64500
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
            exit
        exit
```

VPLS 1 is configured on all PEs with BGP-EVPN and MAC duplication enabled, as follows:

```
configure
    service
        vpls 1 customer 1 create
            bgp
            exit
            bgp-evpn
                evi 1
                mac-duplication
                    detect num-moves 3 window 1
                    retry 2
                    black-hole-dup-mac
                exit
                mpls
                    restrict-protected-src discard-frame
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            sap 1/1/2:1 create                         # backdoor link to PE-3
            exit
            sap 1/2/1:1 create                         # to CE-20
            exit
            no shutdown
        exit
```

To speed up MAC duplication detection, MAC duplication is detected after three MAC moves (default: five MAC moves). To shorten the retry interval, the time window is reduced to one minute (default: three minutes). When a MAC address has been detected as duplicated, the system removes the duplicate MAC entry after a retry interval of two minutes (default: nine minutes). The retry interval must be at least twice the time window for MAC duplication detection.

On the EVPN-MPLS endpoints, **restrict-protected-src discard-frame** must be configured. When MAC address ca:fe:02:20:20:20 is detected on PE-3 as a duplicate MAC address that is black-holed, the EVPN-MPLS endpoints on PE-3 should discard all frames with MAC SA ca:fe:02:20:20:20.

The configuration on the other PEs is similar; only the SAPs are different. VPLS 1 on PE-1 has SAP 1/2/1:1 to CE-10, but no SAP to a backdoor link; VPLS 1 on PE-3 has SAP 1/1/1:1 to the backdoor link to PE-2, but no SAP to a CE.

When CE-20 sends BUM traffic, its MAC SA ca:fe:02:20:20:20 is learned by PE-2 and advertised in EVPN-MAC routes. Because of the backdoor link to PE-3, PE-3 also learns MAC SA ca:fe:02:20:20:20 and advertises it to its BGP peers. The MAC-mobility sequence number is increased until the threshold of three MAC moves is reached. The following BGP EVPN-MAC route with sequence number 2 is sent by PE-2 to PE-3:

```
40 2017/08/17 07:08:17.730 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 96
    Flag: 0x90 Type: 14 Len: 44 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.2
        Type: EVPN-MAC Len: 33 RD: 192.0.2.2:1 ESI: ESI-
0, tag: 0, mac len: 48 mac: ca:fe:02:20:20:20, IP len: 0, IP: NULL, label1: 4194240
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:1
        bgp-tunnel-encap:MPLS
        mac-mobility:Seq:2
"
```

The FDB on PE-2 shows that MAC ca:fe:02:20:20:20 has been learned on local SAP 1/2/1:1, as follows:

```
*A:PE-2# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         ca:fe:01:10:10:10 eMpls:                   Evpn     08/17/17 07:07:58
                            192.0.2.3:262140
1         ca:fe:02:20:20:20 sap:1/1/2:1              L/0      08/17/17 07:08:18
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
```

```
===============================================================================
*A:PE-2#
```

The following FDB on PE-3 shows that MAC ca:fe:02:20:20:20 has been detected as a duplicate and protected MAC (type EvpnD:P) associated with a black-hole endpoint:

```
*A:PE-3# show service id 1 fdb mac ca:fe:02:20:20:20

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         ca:fe:02:20:20:20 black-hole               EvpnD:P  08/17/17 07:16:28
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-3#
```

The following BGP-EVPN information for VPLS 1 on PE-3 shows the settings for MAC duplication detection, and the number of and list of detected duplicate MAC addresses:

```
*A:PE-3# show service id 1 bgp-evpn

===============================================================================
BGP EVPN Table
===============================================================================
MAC Advertisement  : Enabled            Unknown MAC Route  : Disabled
CFM MAC Advertise  : Disabled
VXLAN Admin Status : Disabled           Creation Origin    : manual
MAC Dup Detn Moves : 3                  MAC Dup Detn Window: 1
MAC Dup Detn Retry : 2                  Number of Dup MACs : 1
MAC Dup Detn BH    : Enabled
IP Route Advert    : Disabled

EVI                : 1
Ing Rep Inc McastAd: Enabled
Accept IVPLS Flush : Disabled
Send EVPN Encap    : Enabled
BGP Instance       : 1

-------------------------------------------------------------------------------
Detected Duplicate MAC Addresses             Time Detected
-------------------------------------------------------------------------------
ca:fe:02:20:20:20                            08/17/2017 07:16:28
-------------------------------------------------------------------------------
===============================================================================
---snip---
```

The following message is logged in log 99 on PE-3 when VPLS 1 has detected duplicate MACs:

```
50 2017/08/17 07:16:28.176 UTC MINOR: SVCMGR #2331 Base
"VPLS Service 1 has MAC(s) detected as duplicates by EVPN mac-duplication
detection."
```

MAC address ca:fe:02:20:20:20 remains in the FDB as duplicate and black-holed until the retry interval expires, as follows:

```
*A:PE-3# configure service vpls 1 bgp-evpn mac-duplication retry
  - no retry
  - retry <minutes>

 <minutes>            : [2..60]
```

By default, the retry interval is nine minutes, but in this example, it is set to two minutes, which is the minimum value. The retry interval must be at least twice the time window for MAC duplication detection, which is by default three minutes, but reduced to one minute in this example. The following error is raised when attempting to configure a retry interval of two minutes for a detection time window of three minutes:

```
*A:PE-3# configure service vpls 1 bgp-evpn mac-duplication retry 2
MINOR: SVCMGR #1003 Inconsistent value - mac-duplication detection retry time
should be atleast twice that of detect window
```

After the retry interval expires, the MAC duplication is released.

Log 99 shows the following message when VPLS 1 no longer has duplicate MAC addresses:

```
55 2017/08/17 07:18:28.932 UTC MINOR: SVCMGR #2332 Base
"VPLS Service 1 no longer has MAC(s) detected as duplicates by EVPN mac-duplication
detection."
```

MAC address ca:fe:02:20:20:20 remains in the FDB with type Evpn instead of EvpnD:P. BGP routes only disappear after a withdraw message has been received, whereas locally learned MAC addresses are flushed.

```
*A:PE-3# show service id 1 fdb mac ca:fe:02:20:20:20

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         ca:fe:02:20:20:20 eMpls:                   Evpn     08/17/17 07:18:28
                            192.0.2.2:262140
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-3#
```

Clear Commands

The following FDB entry on PE-3 of type EvpnD:P cannot be cleared with a normal FDB **clear** command:

```
*A:PE-3# show service id 1 fdb mac ca:fe:02:20:20:20

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         ca:fe:02:20:20:20 black-hole               EvpnD:P  08/17/17 07:19:47
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-3#
```

The following error is raised when attempting to clear this FDB entry:

```
*A:PE-3# clear service id 1 fdb mac ca:fe:02:20:20:20
MAJOR: LOG #1202 Cannot perform clear operation - Entry is not of learned type
```

Log 99 shows the following message:

```
92 2017/08/17 07:20:07.378 UTC INDETERMINATE: LOGGER #2010 Base Clear SVCMGR
"Clear function clearSvcIdFdbMac has been run with parameters: svc-id="1"
mac="ca:fe:02:20:20:20".  The completion result is: failure.  Additional error text,
 if any, is: Entry is not of learned type"
```

The following **clear** command releases the MAC duplication from the entry in the FDB, but it does not remove the entry from the FDB if it was learned from EVPN. The type is changed from EvpnD:P to Evpn.

```
*A:PE-3# clear service id 1 evpn mac-dup-detect ca:fe:02:20:20:20
*A:PE-3# show service id 1 fdb mac ca:fe:02:20:20:20

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         ca:fe:02:20:20:20 eMpls:                   Evpn     08/17/17 07:20:03
                            192.0.2.2:262140
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-3#
```

When this MAC duplication is released, VPLS 1 no longer has duplicate MAC addresses detected, as follows:

```
*A:PE-3# show service id 1 bgp-evpn | match "Detected" pre-lines 2 post-lines 5
-------------------------------------------------------------------------------
Detected Duplicate MAC Addresses             Time Detected
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
===============================================================================
===============================================================================
*A:PE-3#
```

Instead of clearing the MAC duplication state for one specific MAC address, all duplicate MAC addresses can be cleared by the following command:

```
*A:PE-3# clear service id 1 evpn mac-dup-detect all
```

Log 99 shows the following messages related to the **clear** commands:

```
*A:PE-3# show log log-id 99 application "logger"
===============================================================================
Event Log 99
===============================================================================
Description : Default System Log
Memory Log contents  [size=500   next event=99  (not wrapped)]

98 2017/08/17 07:21:47.600 UTC INDETERMINATE: LOGGER #2010 Base Clear SVCMGR
"Clear function cliClearSvcIdEvpnDupDetMacAll has been run with parameters: svc-
id="1".  The completion result is: success.  Additional error text, if any, is: "

96 2017/08/17 07:21:02.576 UTC INDETERMINATE: LOGGER #2010 Base Clear SVCMGR
"Clear function cliClearSvcIdEvpnDupDetMac has been run with parameters: svc-
id="1"mac="ca:fe:02:20:20:20".  The completion result is: success.  Additional error
 text, if any, is: "
```

Restrict Protected Source Option

By default, the frames with MAC SA or DA equal to the duplicate MAC address are discarded, but the SAP/SDP-binding where the frame enters the VPLS remains operationally up. With the **restrict-protected-src** option, the system will bring the SAP/SDP-binding down where the frame with duplicate source MAC enters. The configuration on PE-2 and PE-3 is modified with **restrict-protected-src** on the SAP to the backdoor link, as follows:

```
*A:PE-2# configure service vpls 1 sap 1/1/2:1 restrict-protected-src
*A:PE-3# configure service vpls 1 sap 1/1/1:1 restrict-protected-src
```

When CE-20 sends BUM traffic, PE-3 detects MAC ca:fe:02:20:20:20 as duplicate. Log 99 shows that a duplicate MAC address has been detected, that protected MAC address ca:fe:02:20:20:20 has been received on SAP 1/1/1:1 in VPLS 1, and that the status of SAP 1/1/1:1 in VPLS 1 is changed to operationally down, with flag **RxProtSrcMac** indicating that a protected source MAC has been received.

```
*A:PE-3# show log log-id 99 count 3

===============================================================================
Event Log 99
===============================================================================
Description : Default System Log
Memory Log contents  [size=500   next event=103  (not wrapped)]

102 2017/08/17 11:29:07.597 UTC MINOR: SVCMGR #2203 Base
"Status of SAP 1/1/1:1 in service 1 (customer 1) changed to admin=up oper=down
flags=RxProtSrcMac "

101 2017/08/17 11:29:07.597 UTC MINOR: SVCMGR #2208 Base
"Protected MAC ca:fe:02:20:20:20 received on SAP 1/1/1:1 in service 1.
The SAP will be disabled."

100 2017/08/17 11:29:07.597 UTC MINOR: SVCMGR #2331 Base
"VPLS Service 1 has MAC(s) detected as duplicates by EVPN mac-duplication
detection."
```

The following shows that SAP 1/1/1:1 in VPLS 1 on PE-3 is operationally down with flag RxProtSrcMac:

```
*A:PE-3# show service id 1 sap 1/1/1:1

===============================================================================
Service Access Points(SAP)
===============================================================================
Service Id         : 1
SAP                : 1/1/1:1                  Encap             : q-tag
Description        : (Not Specified)
Admin State        : Up                       Oper State        : Down
Flags              : RxProtSrcMac
Multi Svc Site     : None
Last Status Change : 08/17/2017 11:29:08
Last Mgmt Change   : 08/17/2017 11:27:46
===============================================================================
*A:PE-3#
```

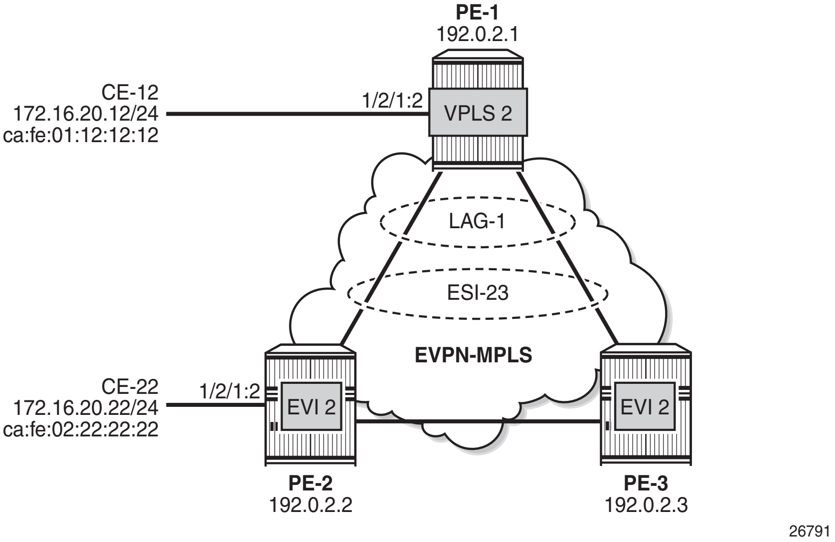
The only way to re-enable the SAP is to disable and enable the SAP, as follows:

```
*A:PE-3# configure service vpls 1 sap 1/1/1:1 shutdown
*A:PE-3# configure service vpls 1 sap 1/1/1:1 no shutdown
*A:PE-3# show service id 1 sap
===============================================================================
SAP(Summary), Service 1
===============================================================================
PortId                          SvcId      Ing.  Ing.    Egr.  Egr.   Adm  Opr
                                           QoS   Fltr    QoS   Fltr
-------------------------------------------------------------------------------
1/1/1:1                         1          1     none    1     none   Up   Up
-------------------------------------------------------------------------------
Number of SAPs : 1
-------------------------------------------------------------------------------
===============================================================================
*A:PE-3#
```

Black-hole MAC Duplication in All-active Multi-homing

Figure 34 shows the example topology with all-active multi-homing.

Figure 34 **Example Topology with All-active Multi-homing**

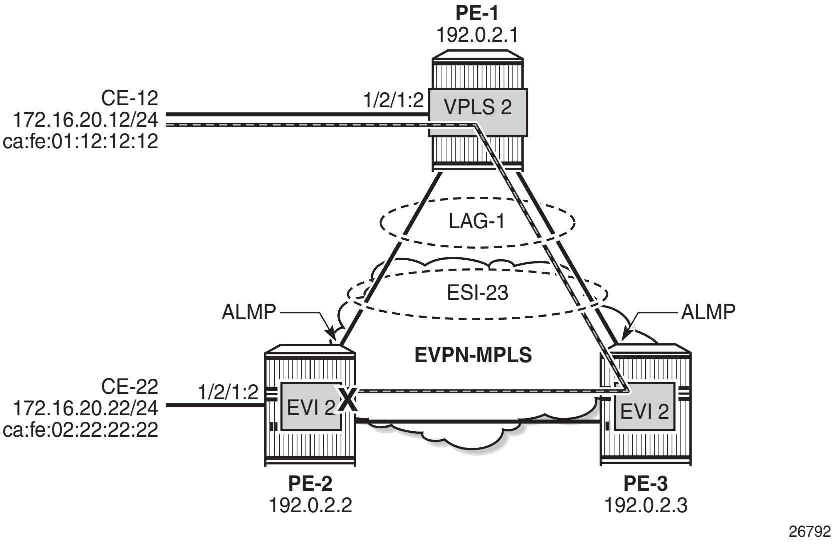


In this topology, the backdoor link is removed. On PE-1, VPLS 2 is configured without EVPN; on PE-2 and PE-3, VPLS 2 is configured with EVPN-MPLS. LAG 1 is configured on the PEs and Ethernet Segment (ES) ESI-23 is created on PE-2 and PE-3, as follows:

```
configure
    service
        system
            bgp-evpn
                ethernet-segment "ESI-23" create
                    esi 01:00:00:00:00:23:00:00:00:01
                    es-activation-timer 3
                    service-carving
                        mode auto
                    exit
                    multi-homing all-active
                    lag 1
                    no shutdown
                exit
```

The reason why black-hole MAC duplication should be configured instead of ALMP is the following. When ALMP is configured on SAP lag-1:2 on PE-2 and PE-3, MAC address ca:fe:01:12:12:12 of CE-12 is learned and protected on the SAP on both PEs. Traffic sent from CE-12 to CE-22 that is hashed over the direct link between PE-1 and PE-2 will reach its destination. Traffic that is hashed over the link between PE-1 and PE-3 will be forwarded by PE-3 to PE-2, but PE-2 will drop the traffic because it contains a MAC SA that is protected locally, as shown in Figure 35.

Figure 35 **Traffic Dropped when ALMP is Configured in All-active Multi-homing**



When black-hole MAC duplication is configured instead of ALMP, traffic hashed on the link to PE-3 is forwarded to PE-2 and to CE-22. This is because MAC duplication is ES-aware and the same MAC seen on the same ES in two different PEs will never be detected as duplicate.

The configuration of VPLS 2 in PE-2 is as follows:

```
configure
    service
        vpls 2
            bgp
            exit
            bgp-evpn
                evi 2
                mac-duplication
                    black-hole-dup-mac
                exit
                mpls
                    restrict-protected-src discard-frame
                    auto-bind-tunnel
            resolution any
        exit
        no shutdown
    exit
exit
sap 1/2/1:2 create
exit
sap lag-1:2 create
exit
no shutdown
```

The configuration of VPLS 2 on PE-3 is similar.

Conclusion

Black-hole MAC for EVPN MAC duplication protects EVPN services against customer-created backdoors or loops, while supporting MAC mobility and all-active multi-homing.

Conditional Static Black-Hole MAC in EVPN

This chapter provides information about Conditional Static Black-Hole MAC in EVPN.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SR OS Release 14.0.R6, but the CLI in the current edition is based on SR OS Release 15.0.R2. Conditional static black-hole MAC is supported on EVPN services only, including EVPN-VXLAN and EVPN-MPLS, in SR OS Release 14.0.R1, and later.

Overview

A static black-hole MAC is a local FDB record associated with a black-hole instead of a SAP or SDP-binding. Black-hole MACs offer a scalable way to filter frames in the data plane based on MAC DA or SA, regardless of how the frame is arriving in the system. Black-hole MACs can be configured in EVPN in the following ways:

- Static configured black-hole MAC
- Anti-spoof MAC in proxy Address Resolution Protocol/Neighbor Discovery (proxy-ARP/ND)
- MAC-duplication black-hole (supported in SR OS Release 15.0.R1, and later), see chapter Black-hole MAC for EVPN Loop Protection

When a specific MAC is configured as a static black-hole MAC, all frames with MAC DA equal to this black-hole MAC will be dropped. Also, black-hole MACs are treated as protected MAC addresses, which allows filtering on MAC SA; see chapter Auto-Learn MAC Protect in EVPN.

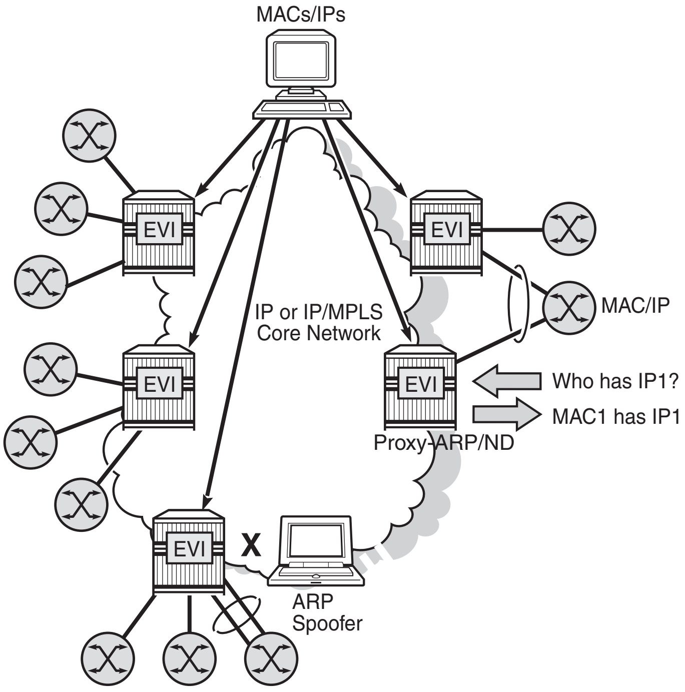
The default behavior on the SAP/SDP-bindings is Restricted Protected Source Discard Frame (RPS-DF). Therefore, all frames with MAC SA equal to the black-hole MAC will, by default, be dropped on the SAP/SDP-binding where the frames enter the service. Instead of dropping the frames, the entire SAP/SDP-binding can be brought operationally down, if the SAP/SDP-binding is explicitly configured with Restricted Protected Source (RPS) without any parameter. The SAP/SDP-binding can only be brought up manually by disabling (shutdown) and re-enabling (no shutdown) the SAP/SDP-binding. On the EVPN endpoints between PEs, it is possible to configure RPS-DF, not RPS. When configured, the EVPN endpoint will drop frames with MAC SA equal to the black-hole MAC.

Black-hole MACs can be used as an alternative to MAC filters, which simplifies the deployment of proxy-ARP/ND with anti-spoof MACs. ARP/ND spoofing is a technique whereby an attacker sends fake ARP/ND messages to a broadcast domain. Generally, the aim is to get the routers in the broadcast domain to associate the attacker's MAC address with the IP address of another host, causing any traffic destined to that IP address to be sent to the attacker instead. To prevent this from happening, a proxy-ARP/ND with duplicate IP detection monitors the number of times the MAC changes for an offending IP address. When a certain number of MAC moves are detected in a defined period, the system flags the proxy-ARP entry as duplicate for a defined hold time and an alarm is sent to log 99.

Chapter EVPN for MPLS Tunnels describes the proxy-ARP/ND configuration with the option to define an anti-spoof MAC (AS-MAC) for EVPN-MPLS networks using MAC filters, including some recommended settings. The AS-MAC will be advertised with the duplicate IP in gratuitous ARP (GARP) and ARP replies to all CEs in the EVPN (in the case of proxy-ND, unsolicited Neighbor Advertisement messages are sent instead of GARP messages).

ARP/ND broadcast traffic is a security issue for Internet eXchange Providers (IXPs) and service providers with large Layer 2 domains. In such networks, administrators try to avoid ARP/ND flooding. Figure 36 shows the proxy-ARP/ND feature where local ARP/ND requests are responded by the system on behalf of the IP interface owners.

Figure 36 **Proxy-ARP/ND and ARP Spoofing**



EVPN can suppress ARP/ND flooding within an EVPN service if all the attached hosts advertise their presence. Therefore, EVPN is preferred in IXPs to mitigate and even eliminate the ARP/ND flooding issue. The proxy-ARP/ND agent responds to local ARP/ND requests using a proxy-ARP/ND table per service. This table is populated by EVPN entries (MAC-IP pairs), static entries configured in the service, and dynamic entries snooped from ARP/GARP/ND messages sent by the ISP routers. The static entries and snooped dynamic entries are also advertised in EVPN-MAC routes.

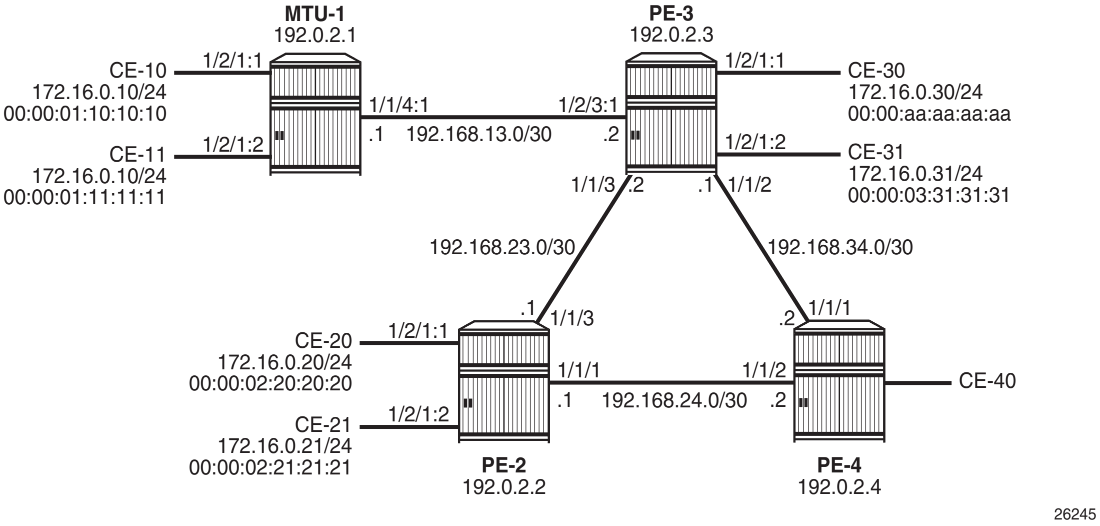
As well as the proxy-ARP/ND, SR OS supports an anti-spoofing mechanism that can detect and block an ARP spoofing attack or a misconfigured duplicated IP address. When using MAC filters, the same anti-spoof-mac option must be configured in all the PEs and this filter may be configured on all the PE SAPs/SDP-bindings to discard all the frames with MAC DA equal to the anti-spoof-mac. This requires a lot of configuration and is prone to configuration errors.

Conditional static black-hole MACs can be configured for the anti-spoof-mac so that frames with MAC DA equal to the anti-spoof-mac can be discarded based on a MAC lookup in the FDB, as opposed to a MAC filter entry. Less configuration is required and this simplifies the deployment of proxy-ARP/ND with AS-MAC. The configuration example in this chapter includes proxy-ARP, but the behavior is similar for proxy-ND.

Configuration

Figure 37 shows the example topology. Traffic will be sent between the CEs and may be dropped in the PEs if the MAC DA or MAC SA matches a black-hole MAC. IP address 172.16.0.10/24 is duplicate (CE-10 and CE-11).

Figure 37 **Example Topology**



The initial configuration on the nodes includes:

- Cards, MDAs, ports
- Router interfaces
- IS-IS between PEs
- LDP between PEs

BGP is configured between the PEs for address family EVPN with PE-2 as route reflector (RR). Instead of an RR, a full mesh can also be configured between the PEs. The BGP configuration on PE-2 is as follows:

```
configure
    router
        autonomous-system 64500
        bgp
            min-route-advertisement 1
            rapid-withdrawal
            split-horizon
            rapid-update evpn
            group "internal"
                family evpn
                cluster 1.1.1.1
                peer-as 64500
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
            exit
        exit
```

VPLS 1 is configured on all PEs and on MTU-1 (MTU-1's VPLS 1 is connected to PE-3 by a SAP). The VPLS configuration on the PEs includes EVPN-MPLS; for example, for PE-3:

```
configure
    service
        vpls 1 customer 1 create
            bgp
            exit
            bgp-evpn
                evi 1
                mpls
                    ingress-replication-bum-label
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            sap 1/2/1:1 create
            exit
            sap 1/2/3:1 create
            exit
            no shutdown
        exit
```

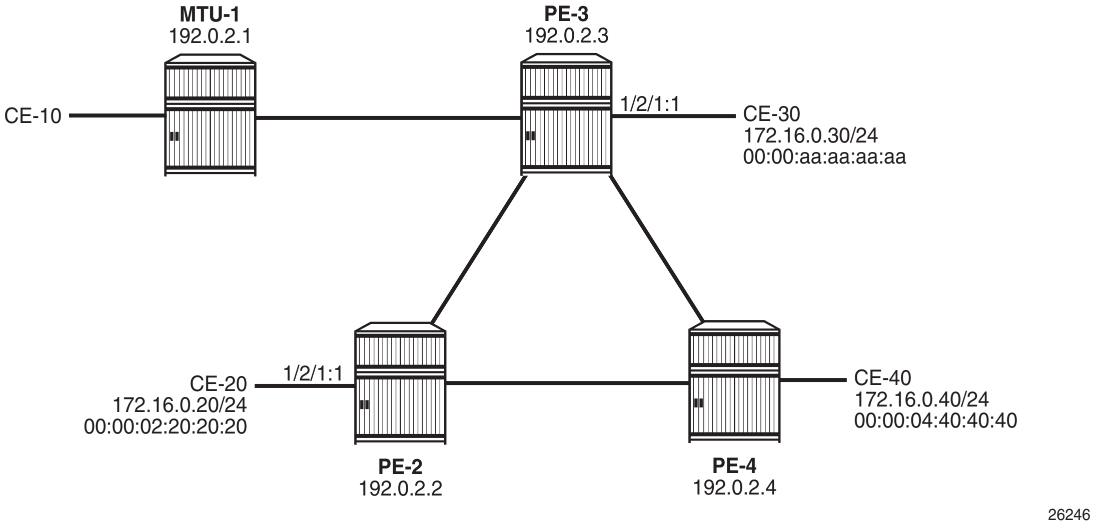
Conditional Static Black-Hole MAC

Conditional static black-hole MAC is an extension to the conditional static MAC, but with the **black-hole** keyword. It is a scalable way to filter MAC DA or SA in the data plane, regardless of how the frame is arriving at the system (SAP/SDP-bindings or EVPN termination endpoints).

When the static black-hole MAC is added to the FDB, all Ethernet frames with MAC DA equal to the black-hole MAC are dropped. Filtering based on the MAC SA is explained in the next section: Conditional Static Black-Hole MAC in Combination with Restrict Protected Source.

Figure 38 shows the example setup with conditional static black-hole MAC 00:00:aa:aa:aa:aa.

Figure 38   **Conditional Static Black-Hole MAC**



When no conditional static black-hole MAC is configured, CE-30 can receive and send traffic from and to the other CEs; for instance, from and toward CE-20, as follows:

```
*A:PE-2# ping router 10 172.16.0.30
PING 172.16.0.30 56 data bytes
64 bytes from 172.16.0.30: icmp_seq=1 ttl=64 time=0.836ms.
64 bytes from 172.16.0.30: icmp_seq=2 ttl=64 time=0.841ms.
---snip---

*A:PE-3# ping router 10 172.16.0.20
PING 172.16.0.20 56 data bytes
64 bytes from 172.16.0.20: icmp_seq=1 ttl=64 time=3.69ms.
64 bytes from 172.16.0.20: icmp_seq=2 ttl=64 time=0.814ms.
---snip---
```

In this example, CE-20 and CE-30 correspond to VPRN 10 configured on PE-2 and PE-3 (using a hairpin to loop the traffic back to the PE).

Conditional static black-hole MAC 00:00:aa:aa:aa:aa (which corresponds to the MAC address of CE-30) is configured in VPLS 1 on PE-3 as follows:

```
*A:PE-3# configure service vpls 1 static-mac mac 00:00:aa:aa:aa:aa create black-hole
```

The black-hole MAC is added as a conditional static (CStatic) MAC that is protected (P), as follows:

```
*A:PE-3# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
```

```
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         00:00:aa:aa:aa:aa black-hole               CStatic: 05/15/17 13:41:03
                                                     P
---snip---
```

The source identifier is black-hole and it is applicable to frames that enter the VPLS on this node, regardless of how they enter the VPLS (SAP, SDP-binding, or EVPN endpoint).

The conditional static black-hole MAC is advertised to the BGP peers in a BGP-EVPN MAC route with the sticky/static bit set, as follows:

```
6 2017/05/15 13:41:03.13 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 96
    Flag: 0x90 Type: 14 Len: 44 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.3
        Type: EVPN-MAC Len: 33 RD: 192.0.2.3:1 ESI: ESI-0, tag: 0, mac len: 48
              mac: 00:00:aa:aa:aa:aa, IP len: 0, IP: NULL, label1: 4194240
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:1
        bgp-tunnel-encap:MPLS
        mac-mobility:Seq:0/Static
"
```

The MAC route is added to the FDB on the other PEs as a static (S) and protected (P) MAC; for example, on PE-2, as follows:

```
*A:PE-2# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         00:00:aa:aa:aa:aa eMpls:                   EvpnS:P  05/15/17 13:41:08
                            192.0.2.3:262140
---snip---
```

When CE-20 sends an ICMP request to CE-30, the MAC DA 00:00:aa:aa:aa:aa is black-holed on PE-3, and no ICMP request succeeds, as follows:

```
*A:PE-2# ping router 10 172.16.0.30
PING 172.16.0.30 56 data bytes
```

```
Request timed out. icmp_seq=1.
Request timed out. icmp_seq=2.
Request timed out. icmp_seq=3.
Request timed out. icmp_seq=4.
Request timed out. icmp_seq=5.

---- 172.16.0.30 PING Statistics ----
5 packets transmitted, 0 packets received, 100% packet loss
*A:PE-2#
```

The port statistics show that the traffic was sent from PE-2 to PE-3, where it entered on port 1/1/3, then got discarded. To verify this, the port statistics are cleared on PE-2 and PE-3, then 1000 ICMP packets are sent from CE-20, as follows:

```
*A:PE-2# clear port 1/[1..2]/[1..4] statistics
*A:PE-3# clear port 1/[1..2]/[1..4] statistics
*A:PE-2# ping router 10 172.16.0.30 rapid count 1000
---snip---
1000 packets transmitted, 0 packets received, 100% packet loss
```

The 1000 packets are received at SAP 1/2/1:1 on PE-2, as follows:

```
*A:PE-2# show port 1/2/1 statistics

===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/2/1                      1000          106000               0               0
===============================================================================
```

These packets are forwarded to port 1/1/3 toward PE-3, as follows:

```
*A:PE-2# show port 1/1/3 statistics

===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/3                        17            1698            1018          125861
===============================================================================
```

On the interfaces between the PEs, other packets are sent besides the ICMP requests, such as IS-IS messages; therefore, the number of packets is slightly greater than 1000.

On PE-3, these packets are received on port 1/1/3, as follows:

```
*A:PE-3# show port 1/1/3 statistics
```

```
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                     Ingress         Ingress         Egress          Egress
Id                       Packets          Octets         Packets         Octets
-------------------------------------------------------------------------------
1/1/3                       1051          129115              51           5016
===============================================================================
```

The FDB entry for this MAC DA is black-holed and no traffic is received on SAP 1/2/1:1 toward CE-30; therefore, the statistics for port 1/2/1 are empty and nothing is displayed, as follows:

```
*A:PE-3# show port 1/2/1 statistics
*A:PE-3#
```

It is possible to configure the black-hole MAC on a different PE; for example, on PE-4 instead of PE-3. The conditional static black-hole MAC configuration in VPLS 1 on PE-3 is removed, as follows:

```
*A:PE-3# configure service vpls 1 static-mac no mac 00:00:aa:aa:aa:aa
```

The conditional static black-hole MAC is configured on PE-4 instead, as follows:

```
*A:PE-4# configure service vpls 1 static-mac mac 00:00:aa:aa:aa:aa create black-hole
```

PE-4 sends EVPN-MAC updates to its peers. PE-2 learns that all traffic with MAC DA 00:00:aa:aa:aa:aa should be redirected to PE-4, as shown in the FDB on PE-2:

```
*A:PE-2# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         00:00:aa:aa:aa:aa eMpls:                   EvpnS:P  05/16/17 18:27:16
                            192.0.2.4:262140
-------------------------------------------------------------------------------
No. of MAC Entries: 1
```

The port statistics are cleared on all PEs and 1000 ICMP packets are sent from CE-20 to CE-30, as follows:

```
*A:PE-2# ping router 10 172.16.0.30 rapid count 1000
---snip---
1000 packets transmitted, 0 packets received, 100% packet loss
```

On PE-2, traffic is not forwarded on the direct link (port 1/1/3) toward PE-3, but redirected to PE-4 (port 1/1/1) instead, as follows:

```
*A:PE-2# show port 1/1/[1..3] statistics

================================================================================
Port Statistics on Slot 1
================================================================================
Port                    Ingress          Ingress         Egress           Egress
Id                      Packets           Octets         Packets          Octets
--------------------------------------------------------------------------------
1/1/1                        16             1534            1017          125718
================================================================================

================================================================================
Port Statistics on Slot 1
================================================================================
Port                    Ingress          Ingress         Egress           Egress
Id                      Packets           Octets         Packets          Octets
--------------------------------------------------------------------------------
1/1/3                        17             1607              16            1614
================================================================================
```

On PE-4, traffic is received on port 1/1/2, then discarded because the MAC DA equals the static black-hole MAC in the FDB, as follows. No traffic is forwarded to PE-3, where CE-30 is attached.

```
*A:PE-4# show port 1/[1..2]/[1..4] statistics

================================================================================
Port Statistics on Slot 1
================================================================================
Port                    Ingress          Ingress         Egress           Egress
Id                      Packets           Octets         Packets          Octets
--------------------------------------------------------------------------------
1/1/1                        75             7494              74            7369
================================================================================
================================================================================
Port Statistics on Slot 1
================================================================================
Port                    Ingress          Ingress         Egress           Egress
Id                      Packets           Octets         Packets          Octets
--------------------------------------------------------------------------------
1/1/2                      1081           131950              81            7971
================================================================================
*A:PE-4#
```

The configuration is restored with conditional static black-hole MAC in VPLS 1 on PE-3, not on PE-4, as follows:

```
*A:PE-3# configure service vpls 1 static-mac mac 00:00:aa:aa:aa:aa create black-hole
*A:PE-4# configure service vpls 1 static-mac no mac 00:00:aa:aa:aa:aa
```

Conditional Static Black-Hole MAC in Combination with Restrict Protected Source

For Ethernet frames with MAC SA equal to the static black-hole MAC, the treatment is the same as for protected MACs (see chapter Auto-Learn MAC Protect in EVPN), but for conditional static black-hole MACs, ALMP need not be enabled on the SAP or SDP-binding:

- When a frame is received with MAC SA equal to the black-hole MAC, it is dropped, because RPS-DF is enabled on the SAP or SDP-binding, by default. RPS-DF need not be enabled explicitly. The default is **no restrict-protected-src**, which operates as RPS-DF. An error message is raised when the following command is entered:

```
*A:PE-3# configure service vpls 1 sap 1/2/1:1 restrict-protected-src discard-frame
MINOR: SVCMGR #7888 Cannot be configured/enabled with EVPN
```

- When RPS is enabled instead of RPS-DF, the SAP or SDP-binding where the frame was received, with MAC SA equal to the black-hole MAC, is brought operationally down. The SAP or SDP-binding can be brought up manually by disabling (shutdown) and re-enabling (no shutdown) the SAP or SDP-binding. RPS is enabled on SAP 1/2/1:1 as follows:

```
configure service vpls 1 sap 1/2/1:1 restrict-protected-src
```

- Optionally, RPS-DF can be enabled on the EVPN-MPLS endpoint or EVPN-VXLAN endpoint. When enabled, the EVPN endpoint will discard frames with MAC SA equal to the black-hole MAC. RPS cannot be configured instead of RPS-DF on EVPN endpoints. It is not an option to bring the EVPN endpoint down when a frame is received with MAC SA equal to the static black-hole MAC. The commands to enable RPS-DF on the EVPN-MPLS endpoints and EVPN-VXLAN endpoints are as follows:

```
configure service vpls 1 bgp-evpn mpls restrict-protected-src
  - no restrict-protected-src
  - restrict-protected-src discard-frame

 <discard-frame>      : keyword - discard frame and  trap on a protected MAC

configure service vpls 1 vxlan vni 1 restrict-protected-src
  - no restrict-protected-src
  - restrict-protected-src discard-frame
```

With the default configuration (RPS-DF on SAP/SDP-bindings), the behavior is as follows for conditional static black-hole MAC 00:00:aa:aa:aa:aa configured in VPLS 1 on PE-3. All traffic from CE-30 with MAC SA 00:00:aa:aa:aa:aa is black-holed on SAP 1/2/1:1 on PE-3, because the default behavior on SAP 1/2/1:1 is RPS-DF, and the frame is discarded. The packets are received on port 1/2/1 (SAP 1/2/1:1) and dropped. No packets are forwarded to port 1/1/3 toward PE-2 or any other port.

```
*A:PE-3# clear port 1/[1..2]/[1..4] statistics
*A:PE-3# ping router 10 172.16.0.20 rapid count 1000
---snip---
1000 packets transmitted, 0 packets received, 100% packet loss
*A:PE-3# show port 1/[1..2]/[1..4] statistics

===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress          Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/2                        15            1555              16            1628
===============================================================================

===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress          Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/3                        16            1650              18            1665
===============================================================================

===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress          Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/2/1                      1000          106000               0               0
===============================================================================
```

If the static MAC is configured in VPLS 1 on PE-4 and not on PE-3, PE-3 will still discard the packets with MAC SA 00:00:aa:aa:aa:aa arriving on SAP 1/2/1:1, because it learned from the EVPN-MAC updates that MAC 00:00:aa:aa:aa:aa is a protected MAC on PE-4. Therefore, traffic with this MAC SA is not expected and not allowed on PE-3, as follows:

```
*A:PE-4# configure service vpls 1 static-mac mac 00:00:aa:aa:aa:aa create black-hole


*A:PE-3# configure service vpls 1 static-mac no mac 00:00:aa:aa:aa:aa
*A:PE-3# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         00:00:aa:aa:aa:aa eMpls:                   EvpnS    05/16/17 18:58:19
                                                     P
                            192.0.2.4:262140
-------------------------------------------------------------------------------
No. of MAC Entries: 1

*A:PE-3# ping router 10 172.16.0.20 rapid count 1000
```

```
---snip---
1000 packets transmitted, 0 packets received, 100% packet loss

*A:PE-3# show port 1/[1..2]/[1..4] statistics

===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/2                        12            1170              14            1378
===============================================================================

===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/3                        14            1360              15            1433
===============================================================================

===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/2/1                      1000          106000               0               0
===============================================================================
```

The configuration is restored as follows:

```
*A:PE-3# configure service vpls 1 static-mac mac 00:00:aa:aa:aa:aa create black-hole
*A:PE-4# configure service vpls 1 static-mac no mac 00:00:aa:aa:aa:aa
```

Optionally, RPS-DF can be configured on the EVPN-MPLS endpoints on the PEs, as follows:

```
configure service vpls 1 bgp-evpn mpls restrict-protected-src discard-frame
```

When RPS-DF is configured on the EVPN-MPLS endpoints, frames with MAC SA 00:00:aa:aa:aa:aa can be discarded by the EVPN endpoints between the PEs. However, in this example this is not required, because any frame with MAC SA 00:00:aa:aa:aa:aa will be dropped by the local SAP before it can be forwarded to an EVPN endpoint.

It is possible to configure RPS without any parameters on SAP 1/2/1:1 on PE-3, as follows:

```
*A:PE-3# configure service vpls 1 sap 1/2/1:1 restrict-protected-src
```

When CE-30 sends traffic with MAC SA equal to a protected MAC address (black-hole or not), the entire SAP 1/2/1:1 will be brought operationally down, as follows:

```
*A:PE-3# ping router 10 172.16.0.20
PING 172.16.0.20 56 data bytes
Request timed out. icmp_seq=1.
Request timed out. icmp_seq=2.
---snip---
---- 172.16.0.20 PING Statistics ----
5 packets transmitted, 0 packets received, 100% packet loss
*A:PE-3# show service id 1 sap
===============================================================================
SAP(Summary), Service 1
===============================================================================
PortId                          SvcId      Ing.  Ing.    Egr.  Egr.   Adm  Opr
                                           QoS   Fltr    QoS   Fltr
-------------------------------------------------------------------------------
1/2/1:1                         1          1     none    1     none   Up   Down
-------------------------------------------------------------------------------
Number of SAPs : 1
-------------------------------------------------------------------------------
===============================================================================
*A:PE-3#
```

The following information for SAP 1/2/1:1 in VPLS 1 shows that this SAP is operationally down because a protected source MAC address was received on this SAP (Flags: RxProtSrcMac), as follows:

```
*A:PE-3# show service id 1 sap 1/2/1:1

===============================================================================
Service Access Points(SAP)
===============================================================================
Service Id         : 1
SAP                : 1/2/1:1                  Encap             : q-tag
Description        : (Not Specified)
Admin State        : Up                       Oper State        : Down
Flags              : RxProtSrcMac
Multi Svc Site     : None
Last Status Change : 05/15/2017 12:30:49
Last Mgmt Change   : 05/15/2017 12:30:49
===============================================================================
*A:PE-3#
```

Log 99 shows that a protected MAC was received on SAP 1/2/1:1 and the SAP went operationally down with flag RxProtSrcMac, as follows:

```
56 2017/05/15 12:30:49.65 UTC MINOR: SVCMGR #2203 Base
"Status of SAP 1/2/1:1 in service 1 (customer 1) changed to admin=up oper=down
flags=RxProtSrcMac "

59 2017/05/15 12:31:06.65 UTC MINOR: SVCMGR #2208 Base Slot 1
"Protected MAC 00:00:aa:aa:aa:aa received on SAP 1/2/1:1 in service 1. "
```

The SAP can only be brought up manually by disabling and re-enabling the SAP, as follows:

```
*A:PE-3# configure service vpls 1 sap 1/2/1:1 shutdown
*A:PE-3# configure service vpls 1 sap 1/2/1:1 no shutdown
*A:PE-3# show service id 1 sap

===============================================================================
SAP(Summary), Service 1
===============================================================================
PortId                          SvcId      Ing.  Ing.    Egr.  Egr.   Adm  Opr
                                           QoS   Fltr    QoS   Fltr
-------------------------------------------------------------------------------
1/2/1:1                         1          1     none    1     none   Up   Up
-------------------------------------------------------------------------------
Number of SAPs : 1
-------------------------------------------------------------------------------
===============================================================================
*A:PE-3#
```

The default behavior of SAP 1/2/1:1 is RPS-DF, which is configured by removing the RPS configuration, as follows:

```
*A:PE-3# configure service vpls 1 sap 1/2/1:1 no restrict-protected-src
```

The conditional static black-hole MAC configuration is removed as follows:

```
*A:PE-3# configure service vpls 1 static-mac no mac 00:00:aa:aa:aa:aa
```

Black-Hole MAC in Services with Proxy-ARP/ND

In this example, only proxy-ARP is shown, not proxy-ND. However, the configuration and procedures for proxy-ND would be equivalent.

First, the implementation of proxy-ARP and AS-MAC is described without static black-hole MACs. MAC filters will be required to drop or redirect traffic, but these are not shown in the example. Configuring MAC filters and applying them on SAP/SDP-bindings is labor-intensive and can be error-prone. Afterward, the implementation with AS-MAC as static black-hole is described.

Services with Proxy-ARP and AS-MAC - No Static Black-Hole MAC

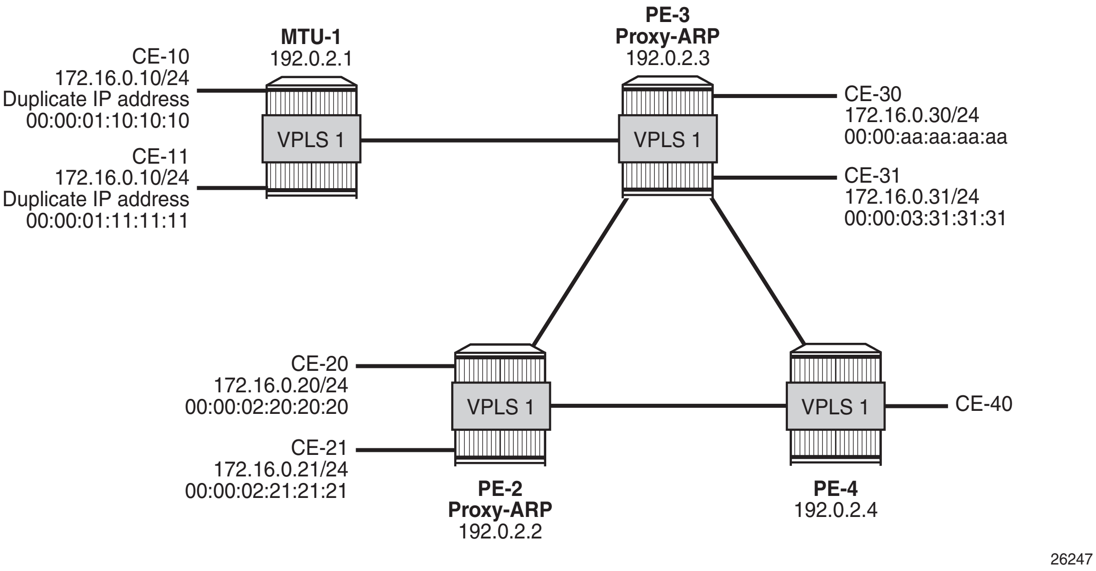
IP duplication works when the IP address moves between:

- Dynamic (learned on SAP) and EVPN
- EVPN and dynamic
- Dynamic and dynamic

The following example shows IP address moves from dynamic to dynamic between SAP 1/2/1:1 (to CE-10) and SAP 1/2/1:2 (to CE-11) in VPLS 1 on MTU-1. However, the duplicate IP address could have been in PE-3 and MTU-1 instead (EVPN or dynamic) and still the IP address would have been detected as duplicate.

Figure 39 shows the example setup with duplicate IP address 172.16.0.10/24 for CE-10 and CE-11. VPLS 1 is configured with proxy-ARP with duplicate IP detection in PE-2 and PE-3 (and possibly also in other PEs). MAC address 00:00:bb:bb:bb:bb is configured as AS-MAC, which will be used when a duplicate IP address has been detected.

Figure 39 **VPLS 1 with Proxy-ARP and AS-MAC**



For IP duplication detection, the following parameters can be customized so that the system can react to particular conditions in the network. The syntax is as follows:

```
*A:PE-3# configure service vpls 1 proxy-arp dup-detect
  - dup-detect [anti-spoof-mac <mac-address>] window <minutes> num-moves <count>
             hold-down <minutes|max>
  - dup-detect anti-spoof-mac <mac-address> window <minutes> num-moves <count>
             hold-down <minutes| max> [static-black-hole]

 <mac-address>         : xx-xx-xx-xx-xx-xx or xx:xx:xx:xx:xx:xx (hex chars)
 <minutes>             : [1..15] minutes - default:3
 <count>               : [3..10] - default:5
 <minutes|max>         : [2..60] default=9 | max - permanent hold
 <static-black-hole>  : keyword
```

In VPLS 1 on PE-3, a proxy-ARP with duplicate IP detection is configured, including an optional anti-spoof MAC (AS-MAC) 00:00:bb:bb:bb:bb for offending IP addresses, as follows:

```
configure
    service
        vpls 1
            proxy-arp
                dup-detect window 3 num-moves 3 hold-down max
                           anti-spoof-mac 00:00:bb:bb:bb:bb
                dynamic-arp-populate
                static 172.16.0.20 00:00:02:20:20:20
                no shutdown
            exit
```

The proxy-ARP table contains one static entry (for IP 172.16.0.20). In this case, dynamic ARP populate is enabled. Therefore, the proxy-ARP table will be updated with ARP entries for IP 172.16.0.10 and MAC 00:00:01:10:10:10 or MAC 00:00:01:11:11:11 for frames originating from CE-10 or CE-11.

When a duplicate IP is detected for IP 172.16.0.10 (after three changes of MAC for IP 172.16.0.10 in a period of three minutes), the corresponding ARP entry contains the duplicate IP address 172.16.0.10 and the AS-MAC 00:00:bb:bb:bb:bb and its type is duplicate (dup). Therefore, this ARP entry is always active until it is removed. Until now, this configuration does not include a static black-hole MAC, and this option is by default disabled. This configuration for duplicate IP detection can be used in combination with MAC filters. The configuration with static black-hole MAC is shown in the section Services with Proxy-ARP and AS-MAC Configured as Static Black-Hole MAC..

The configured AS-MAC will be advertised in an EVPN-MAC route with the sticky/static bit set and without any IP address (because there is no IP duplication detected yet), as follows:

```
23 2017/05/15 06:44:52.44 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 131
    Flag: 0x90 Type: 14 Len: 79 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.3
        Type: EVPN-MAC Len: 33 RD: 192.0.2.3:1 ESI: ESI-0, tag: 0, mac len: 48
              mac: 16:0b:ff:00:03:3a, IP len: 0, IP: NULL, label1: 4194240
        Type: EVPN-MAC Len: 33 RD: 192.0.2.3:1 ESI: ESI-0, tag: 0, mac len: 48
              mac: 00:00:bb:bb:bb:bb, IP len: 0, IP: NULL, label1: 4194240
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:1
        bgp-tunnel-encap:MPLS
        mac-mobility:Seq:0/Static
"
```

Without the option static black-hole, the configured AS-MAC is not added to the local FDB, but this MAC address is treated as a local MAC. The FDB on PE-3 does not contain AS-MAC 00:00:bb:bb:bb:bb, as follows:

```
*A:PE-3# show service id 1 fdb mac 00:00:bb:bb:bb:bb

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================
*A:PE-3#
```

Debugging is enabled for proxy-ARP for IP address 172.16.0.10 in VPLS 1 on PE-3, as follows:

```
*A:PE-3# debug service id 1 proxy-arp ip 172.16.0.10
```

When traffic is sent from CE-11 to CE-21, a dynamic ARP entry for IP address 172.16.0.10 and MAC 00:00:01:11:11:11 is added to the proxy-ARP table for VPLS 1 in PE-3, and an EVPN-MAC update is sent to the peer PEs, as follows:

```
59 2017/05/16 10:14:11.06 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 mac: 00:00:01:11:11:11 Static: N evpn advertise"

60 2017/05/16 10:14:11.06 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 type: Dyn mac: 00:00: 01:11:11:11 Added"

61 2017/05/16 06:51:11.44 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 92
    Flag: 0x90 Type: 14 Len: 48 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.3
        Type: EVPN-MAC Len: 37 RD: 192.0.2.3:1 ESI: ESI-0, tag: 0, mac len: 48
              mac: 00:00:01:11:11:11, IP len: 4, IP: 172.16.0.10, label1: 4194240
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:64500:1
        bgp-tunnel-encap:MPLS
"
```

There is no duplicate IP detected yet.

The following GARP update is sent locally:

```
62 2017/05/16 10:14:11.19 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 type: Dyn mac: 00:00:01:11:11:11 Gratuitous Update"
```

CE-10 and CE-11 have the same IP address for different MAC addresses. When CE-10 sends traffic to CE-20, the ARP entry for IP 172.16.0.10 changes MAC from 00:00:01:11:11:11 to 00:00:01:10:10:10, and an EVPN-MAC withdraw message is sent, as follows:

```
63 2017/05/16 10:14:36.55 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 mac: 00:00:01:11:11:11 Static: N evpn withdraw"

64 2017/05/16 10:14:36.55 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 Mac Change: 00:00:01:11:11:11->00:00:01:10:10:10

65 2017/05/16 10:14:36.55 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 46
    Flag: 0x90 Type: 15 Len: 42 Multiprotocol Unreachable NLRI:
        Address Family EVPN
        Type: EVPN-MAC Len: 37 RD: 192.0.2.3:1 ESI: ESI-0, tag: 0, mac len: 48
              mac: 00:00:01:11:11:11, IP len: 4, IP: 172.16.0.10, label1: 0
"
```

When the MAC changes, the system sends an ARP request for confirmation of the old MAC 00:00:01:11:11:11 for IP 172.16.0.10, as follows:

```
66 2017/05/16 10:14:36.69 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 mac: 00:00:01:11:11:11 confirm"
```

When MAC 00:00:01:11:11:11 is confirmed, the MAC in the ARP entry is changed once again to 00:00:01:11:11:11 and another ARP request is sent asking to confirm MAC 00:00:01:10:10:10 for IP 172.16.0.10, as follows:

```
67 2017/05/16 10:14:36.69 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 Mac Change: 00:00:01:10:10:10->00:00:01:11:11:11 "

68 2017/05/16 10:14:36.79 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 mac: 00:00:01:10:10:10 confirm"
```

When CE-10 confirms MAC 00:00:01:10:10:10 for IP 172.16.0.10, IP duplication is detected for IP address 172.16.0.10 (after three MAC moves in a detection period of three minutes), and the following message is raised in log 99 after a duplicate proxy-ARP entry was detected for IP 172.16.0.10:

```
60 2017/05/16 10:14:56.19 UTC MINOR: SVCMGR #2346 Base
"A duplicate proxy ARP entry was detected with new MAC 00:00:01:10:10:10 for entry I
P 172.16.0.10 MAC 00:00:01:11:11:11 in service 1"
```

The following proxy-ARP debug messages show that the ARP entry for IP 172.16.0.10 in the proxy-ARP table changed MAC to the AS-MAC 00:00:bb:bb:bb:bb, and the type from dynamic to duplicate:

```
69 2017/05/16 10:14:36.79 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 mac: 00:00:bb:bb:bb:bb Static: Y evpn advertise"

70 2017/05/16 10:14:36.79 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 Mac Change: 00:00:01:11:11:11->00:00:bb:bb:bb:bb
 Type Change: Dyn->Dup "

71 2017/05/16 10:14:36.79 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 type: Dup    Dup Detected"
```

If a duplicate IP is detected, AS-MAC 00:00:bb:bb:bb:bb is advertised with duplicate IP address 172.16.0.10 in an EVPN-MAC update to the BGP peers with the sticky/static bit set, as follows:

```
72 2017/05/16 10:14:36.79 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 100
    Flag: 0x90 Type: 14 Len: 48 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.3
        Type: EVPN-MAC Len: 37 RD: 192.0.2.3:1 ESI: ESI-0, tag: 0, mac len: 48
              mac: 00:00:bb:bb:bb:bb, IP len: 4, IP: 172.16.0.10, label1: 4194240
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:1
        bgp-tunnel-encap:MPLS
        mac-mobility:Seq:0/Static
"
```

The difference with the first EVPN-MAC update for AS-MAC is the IP address. Immediately after the AS-MAC was configured, it was also advertised to the BGP-EVPN peers, but without any IP address.

The proxy-ARP entry is shown with type duplicate (dup) and active status in the proxy-ARP table for VPLS 1 on PE-3, as follows:

```
*A:PE-3# show service id 1 proxy-arp detail
-------------------------------------------------------------------------------
Proxy Arp
-------------------------------------------------------------------------------
Admin State       : enabled
Dyn Populate      : enabled
Age Time          : disabled            Send Refresh      : disabled
Table Size        : 250                 Total             : 3
Static Count      : 1                   EVPN Count        : 1
Dynamic Count     : 0                   Duplicate Count   : 1

Dup Detect
-------------------------------------------------------------------------------
Detect Window     : 3 mins              Num Moves         : 3
Hold down         : max
Anti Spoof MAC    : 00:00:bb:bb:bb:bb

EVPN
-------------------------------------------------------------------------------
Garp Flood        : enabled             Req Flood         : enabled
Static Black Hole : disabled
-------------------------------------------------------------------------------

===============================================================================
VPLS Proxy Arp Entries
===============================================================================
IP Address          Mac Address         Type      Status    Last Update
-------------------------------------------------------------------------------
172.16.0.10         00:00:bb:bb:bb:bb   dup       active    05/16/2017 10:14:36
172.16.0.20         00:00:02:20:20:20   stat      inActv    05/16/2017 10:09:40
-------------------------------------------------------------------------------
Number of entries : 2
===============================================================================
*A:PE-3#
```

A duplicate entry is always active, regardless of the AS-MAC. When the entry with the duplicate IP and the AS-MAC are installed in the proxy-ARP table as active, every ARP request for the duplicate IP will be replied by the system. The entry in the proxy-ARP table is treated as active, even if the AS-MAC is not in the FDB (AS-MACs do not consume FDB space). The AS-MAC, along with the duplicate IP, is advertised in EVPN with the sticky/static bit set, as shown earlier. GARP messages with AS-MAC/IP information are flooded locally to make the CEs update their ARP caches to use AS-MAC for traffic to the duplicate IP 172.16.0.10, as follows.

```
73 2017/05/16 10:14:36.89 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 type: Dup mac: 00:00:bb:bb:bb:bb Gratuitous Update"
```

Note: The AS-MAC will always be "unique" in the system. When the AS-MAC is configured, the system will flush any entry with the same MAC learned through EVPN or dynamic sources. Conditional static MACs or OAM MACs with the same value as the AS-MAC are only allowed when they are configured as black-hole, which is not the case yet.

When the duplicate proxy-ARP entry is cleared from the list (hold-down timer expires, or clear command, or replacement of the duplicate entry for a static entry), an ARP request asking who has IP 172.16.0.10 is flooded by the proxy-ARP agent. This ARP refresh triggers an ARP reply from the IP owner, which will be learned in the proxy-ARP table and advertised in EVPN. The system will also send a GARP to local SAP/SDP-bindings. This will correct all host ARP caches in the network. In this example, the duplicate proxy-ARP entry is manually cleared, as follows:

```
*A:PE-3# clear service id 1 proxy-arp duplicate
```

Log 99 shows that the clear function has been run and the duplicate proxy-ARP entry 172.16.0.10 is cleared. The system forces a refresh and, if the condition with the duplicate IP address remains, this is detected almost immediately and a message is logged that a duplicate proxy-ARP entry was detected, as follows:

```
67 2017/05/16 10:14:55.93 UTC INDETERMINATE: LOGGER #2010 Base Clear SVCMGR
"Clear function clearSvcIdProxyArpDups has been run with parameters: svc-id="1" ip-
address="".  The completion result is: success. Additional error text, if any, is: "

68 2017/05/16 10:14:55.93 UTC MINOR: SVCMGR #2347 Base
"A duplicate proxy ARP entry 172.16.0.10 is cleared in service 1"

69 2017/05/16 10:14:56.19 UTC MINOR: SVCMGR #2346 Base
"A duplicate proxy ARP entry was detected with new MAC 00:00:01:11:11:11 for entry I
P 172.16.0.10 MAC 00:00:01:10:10:10 in service 1"
```

The following debug messages for proxy-ARP on PE-3 show the process in more detail. Initially, an EVPN-MAC route withdraw message is sent and the proxy-ARP entry is deleted.

```
74 2017/05/16 10:14:55.93 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 mac: 00:00:bb:bb:bb:bb Static: Y evpn withdraw"

75 2017/05/16 10:14:55.93 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 type: Dup mac: 00:00:bb:bb:bb:bb Deleted"
```

The following BGP-EVPN MAC update is sent by PE-3 to indicate that the AS-MAC is withdrawn for IP 172.16.0.10 (multiprotocol unreachable NLRI):

```
77 2017/05/16 10:14:55.93 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 46
    Flag: 0x90 Type: 15 Len: 42 Multiprotocol Unreachable NLRI:
        Address Family EVPN
        Type: EVPN-MAC Len: 37 RD: 192.0.2.3:1 ESI: ESI-0, tag: 0, mac len: 48
              mac: 00:00:bb:bb:bb:bb, IP len: 4, IP: 172.16.0.10, label1: 0
"
```

Removing the active duplicate entry from the proxy-ARP table triggers an ARP flooding request asking who has IP 172.16.0.10 in VPLS 1, as follows:

```
76 2017/05/16 10:14:55.93 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 flood request"
```

The result of the ARP flooding request is that the IP owners reply with their MAC, at the local or a remote PE. In this case, the reply from CE-10 is received first (IP 172.16.0.10 - MAC 00:00:01:10:10:10), a dynamic proxy-ARP entry is added, and the MAC/IP route is advertised, as follows:

```
78 2017/05/16 10:14:55.92 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 mac: 00:00:01:10:10:10 Static: N evpn advertise"

79 2017/05/16 10:14:55.92 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 type: Dyn mac: 00:00:01:10:10:10 Added"
```

When CE-11 answers with its MAC 00:00:01:11:11:11, the MAC/IP route is withdrawn for IP 172.16.0.10, and the MAC address in the proxy-ARP entry for IP 172.16.0.10 is changed from MAC 00:00:01:10:10:10 to MAC 00:00:01:11:11:11, as follows:

```
80 2017/05/16 10:14:55.92 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 mac: 00:00:01:10:10:10 Static: N evpn withdraw"

81 2017/05/16 10:14:55.92 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 Mac Change: 00:00:01:10:10:10->00:00:01:11:11:11 "
```

Any change of MAC address in a proxy-ARP entry triggers an ARP request asking for confirmation of the old MAC address for IP 172.16.0.10, in this case for MAC 00:00:01:10:10:10, as follows:

```
82 2017/05/16 10:14:56.08 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 mac: 00:00:01:10:10:10 confirm"
```

MAC 00:00:01:10:10:10 is confirmed for IP 172.16.0.10; therefore, the MAC address is changed in the proxy-ARP entry from 00:00:01:11:11:11 to 00:00:01:10:10:10, and an ARP confirmation is asked for the old MAC 00:00:01:11:11:11, as follows:

```
83 2017/05/16 10:14:56.08 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 Mac Change: 00:00:01:11:11:11->00:00:01:10:10:10 "

84 2017/05/16 10:14:56.18 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 mac: 00:00:01:11:11:11 confirm"
```

MAC 00:00:01:11:11:11 is confirmed and, therefore, three MAC moves occurred within three minutes. Duplicate IP 172.16.0.10 is detected and the proxy-ARP entry has the AS-MAC 00:00:bb:bb:bb:bb and type duplicate (Dup), as follows:

```
85 2017/05/16 10:14:56.18 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 mac: 00:00:bb:bb:bb:bb Static: Y evpn advertise"

86 2017/05/16 10:14:56.18 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 Mac Change: 00:00:01:10:10:10->00:00:bb:bb:bb:bb
Type Change: Dyn->Dup "

87 2017/05/16 10:14:56.18 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 type: Dup    Dup Detected"

88 2017/05/16 10:14:56.19 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 100
    Flag: 0x90 Type: 14 Len: 48 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.3
        Type: EVPN-MAC Len: 37 RD: 192.0.2.3:1 ESI: ESI-0, tag: 0, mac len: 48
              mac: 00:00:bb:bb:bb:bb, IP len: 4, IP: 172.16.0.10, label1: 4194240
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:1
        bgp-tunnel-encap:MPLS
        mac-mobility:Seq:0/Static
"
```

A GARP update is sent for IP 172.16.0.10 and AS-MAC 00:00:bb:bb:bb:bb, as follows:

```
89 2017/05/16 10:14:56.29 UTC MINOR: DEBUG #2001 Base proxy arp
"proxy arp:
svc: 1 ip: 172.16.0.10 type: Dyn mac: 00:00:bb:bb:bb:bb Gratuitous Update"
```

The AS-MAC is optionally configured and populates all the host ARP caches when a duplicate IP is detected. All traffic destined to the suspicious IP address 172.16.0.10 will have the AS-MAC 00:00:bb:bb:bb:bb as MAC DA. The user can configure MAC filters on all SAP/SDP-bindings where the CEs are connected to drop, log, or redirect traffic destined to the AS-MAC. This will block any interception or man-in-the-middle attack (due to ARP-spoofing) in the network.

The AS-MAC is independently configured on each PE for the same service. When a different AS-MAC is configured per PE for the same service, the user will need to filter all the AS-MACs in the service at each PE, which increases the complexity of the filters. Nokia recommends using the same AS-MAC for the same service in all the PES where duplicate detect is active and MAC filters need to be configured. However, this recommendation is suspended when the AS-MAC is configured as static black-hole MAC, as described in the following section.

Services with Proxy-ARP and AS-MAC Configured as Static Black-Hole MAC.

With the AS-MAC configured as static black-hole MAC, MAC-filters do not need to be configured to discard frames with MAC DA equal to the AS-MAC. Instead, the user can decide whether to use the same AS-MAC on all the PEs. This scalability is not limited by the number of filters, but by the number of FDB entries.

The **static-black-hole** parameter is optional and disabled by default. In the example, the static-black-hole option is not configured yet for the AS-MAC and the behavior is as follows:

- The AS-MAC is added to the MAC DB as local, but not programmed in the FDB.
- The AS-MAC is advertised in EVPN (initially without an IP address, and with an IP address as soon as the IP is detected as duplicate).
- The AS-MAC cannot be overridden by any other MAC.
- The AS-MAC value cannot be configured on a static MAC, because that MAC value is reserved for the proxy-ARP, as follows:

```
*A:PE-3# configure service vpls 1 static-mac mac 00:00:bb:bb:bb:bb create
sap 1/2/3:1 monitor fwd-status
MINOR: SVCMGR #7875 Cannot create conditional static mac - Mac reserved by proxy

*A:PE-3# configure service vpls 1 static-mac mac 00:00:bb:bb:bb:bb create black-hole
MINOR: SVCMGR #7875 Cannot create conditional static mac - Mac reserved by proxy
```

When the static-black-hole option is not configured, the AS-MAC is considered as a local MAC and cannot be overridden. The MAC priority is as follows:

1. Local MACs (including AS-MACs without static-black-hole, es-bmacs, src-bmacs, OAM, and so on)
2. Conditional static MACs (including AS-MACs with static-black-hole)
3. Auto-Learn Protected MACs
4. EVPN-MACs with sticky/static bit set
5. Data plane learned MACs (regular learning on SAP/SDP-binding)
6. EVPN-MACs without sticky/static bit set

To configure an AS-MAC with static-black-hole option, a static black-hole MAC needs to be configured first. The following error is raised when no static black-hole MAC has been configured for AS-MAC 00:00:bb:bb:bb:bb:

```
*A:PE-3# configure service vpls 1 proxy-arp dup-detect window 3 num-moves 5 hold-
down max anti-spoof-mac 00:00:bb:bb:bb:bb static-black-hole
MINOR: SVCMGR #8007 Cannot modify proxy arp - black-hole mac not configured on
service
```

In that case, the AS-MAC needs to be removed from the proxy-ARP configuration, as follows:

```
configure
    service
        vpls 1
            proxy-arp
                shutdown
                dup-detect window 3 num-moves 5 hold-down max
            exit
```

Then, the static black-hole MAC can be created as follows:

```
configure service vpls 1 static-mac mac 00:00:bb:bb:bb:bb create black-hole
```

After the conditional static black-hole MAC is configured, duplicate IP detection cannot be configured with AS-MAC, unless the static-black-hole option is added, as follows:

```
*A:PE-3# configure service vpls 1 static-mac mac 00:00:bb:bb:bb:bb create black-
hole
*A:PE-3# configure service vpls 1 proxy-arp dup-detect window 3 num-moves 5 hold-
down max anti-spoof-mac 00:00:bb:bb:bb:bb
MINOR: SVCMGR #8007 Cannot modify proxy arp - conditional static mac configured on
service
```

When the static black-hole MAC 00:00:bb:bb:bb:bb is configured, the AS-MAC can only be configured with the static-black-hole-option in VPLS 1 on PE-2 and PE-3, as follows:

```
configure
    service
```

```
        vpls 1
            static-mac
                mac 00:00:bb:bb:bb:bb create black-hole
            exit
            proxy-arp
                dup-detect window 3 num-moves 5 hold-down max
                           anti-spoof-mac 00:00:bb:bb:bb:bb static-black-hole
                dynamic-arp-populate
                static 172.16.0.20 00:00:02:20:20:20
                no shutdown
            exit
```

When the AS-MAC is configured with the static black-hole option, the AS-MAC will be added not only to the MAC DB, but also to the FDB as CStatic, and associated with a black-hole endpoint, as follows:

```
*A:PE-3# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         00:00:bb:bb:bb:bb black-hole               CStatic: 12/09/16 10:17:00
                                                     P
---snip---
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-3#
```

Any frame with MAC DA equal to the AS-MAC with static black-hole will be dropped, regardless of the ingress endpoint and without any need for a filter. This mechanism is the only way to filter MAC DAs on EVPN endpoints, because MAC filters cannot be configured on EVPN endpoints.

The AS-MAC with static black-hole will be advertised in EVPN with the sticky/static bit set, as follows:

```
96 2017/05/16 10:17:00.10 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 96
    Flag: 0x90 Type: 14 Len: 44 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.3
        Type: EVPN-MAC Len: 33 RD: 192.0.2.3:1 ESI: ESI-0, tag: 0, mac len: 48
              mac: 00:00:bb:bb:bb:bb, IP len: 0, IP: NULL, label1: 4194240
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:1
        bgp-tunnel-encap:MPLS
        mac-mobility:Seq:0/Static
"
```

When a duplicate IP address is detected, the EVPN-MAC update contains the IP address 172.16.0.10, as follows:

```
126 2017/05/16 11:04:37.65 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 100
    Flag: 0x90 Type: 14 Len: 48 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.3
        Type: EVPN-MAC Len: 37 RD: 192.0.2.3:1 ESI: ESI-0, tag: 0, mac len: 48
              mac: 00:00:bb:bb:bb:bb, IP len: 4, IP: 172.16.0.10, label1: 4194240
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:1
        bgp-tunnel-encap:MPLS
        mac-mobility:Seq:0/Static
"
```

The local CEs receive a GARP update with the AS-MAC address. The ARP table of CE-30 and CE-31 have an entry for the duplicate IP address 172.16.0.10 with the AS-MAC 00:00:bb:bb:bb:bb, as follows:

```
*A:PE-3# show router 10 arp

===============================================================================
ARP Table (Service: 10)
===============================================================================
IP Address      MAC Address        Expiry    Type   Interface
-------------------------------------------------------------------------------
172.16.0.10     00:00:bb:bb:bb:bb 03h43m02s Dyn[I] int-CE-30-PE-3
172.16.0.20     00:00:02:20:20:20 03h41m19s Dyn[I] int-CE-30-PE-3
172.16.0.30     00:00:aa:aa:aa:aa 00h00m00s Oth[I] int-CE-30-PE-3
-------------------------------------------------------------------------------
No. of ARP Entries: 3
===============================================================================


*A:PE-3# show router 11 arp

===============================================================================
ARP Table (Service: 11)
===============================================================================
IP Address      MAC Address        Expiry    Type   Interface
-------------------------------------------------------------------------------
172.16.0.10     00:00:bb:bb:bb:bb 03h47m43s Dyn[I] int-CE-31-PE-3
172.16.0.31     00:00:03:31:31:31 00h00m00s Oth[I] int-CE-31-PE-3
-------------------------------------------------------------------------------
```

```
No. of ARP Entries: 2
===============================================================================
A:PE-3#
```

CE-30 and CE-31 cannot reach CE-10 or CE-11, because the MAC DA will be the AS-MAC and all traffic to this MAC DA is black-holed instead of forwarded to SAP 1/2/3:1 toward CE-10 or CE-11. When 1000 ICMP packets are sent by CE-30, they arrive in SAP 1/2/1:1 on PE-3 and are then discarded, as follows:

```
*A:PE-3# clear port 1/[1..2]/[1..4] statistics
*A:PE-3# ping router 10 172.16.0.10 rapid count 1000
PING 172.16.0.10 56 data bytes
---snip---
---- 172.16.0.10 PING Statistics ----
1000 packets transmitted, 0 packets received, 100% packet loss
*A:PE-3# show port 1/[1..2]/[1..4] statistics

===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets        Packets           Octets
-------------------------------------------------------------------------------
1/1/2                        37            3808             38             3759
===============================================================================

===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets        Packets           Octets
-------------------------------------------------------------------------------
1/1/3                        41            4043             41             4104
===============================================================================

===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets        Packets           Octets
-------------------------------------------------------------------------------
1/2/1                      1000          106000              0                0
===============================================================================
*A:PE-3#
```

No packets were forwarded to SAP 1/2/3:1 toward CE-10; therefore, there are no statistics for port 1/2/3.

Conclusion

Static black-hole MACs can be applied in EVPN for security as a scalable alternative to MAC filters. Static black-hole MACs are programmed in the FDB and all frames with MAC DA equal to the static black-hole MAC are dropped, regardless of how the frame arrived at the system (SAP/SDP-binding or EVPN endpoint). Also, static black-hole MACs are treated like protected MACs and, in combination with RPS(-DF), filtering on MAC SA is performed in the data plane. Black-hole MACs can be an option for an AS-MAC in services with proxy-ARP/ND enabled, which simplifies the configuration because MAC filters are not required.

EVPN for MPLS Tunnels

This chapter provides information about EVPN for MPLS tunnels.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SR OS Release 13.0.R6, but the CLI in the current edition corresponds to release 15.0.R2. A prerequisite is to read the EVPN for VXLAN Tunnels (Layer 2) chapter.

Overview

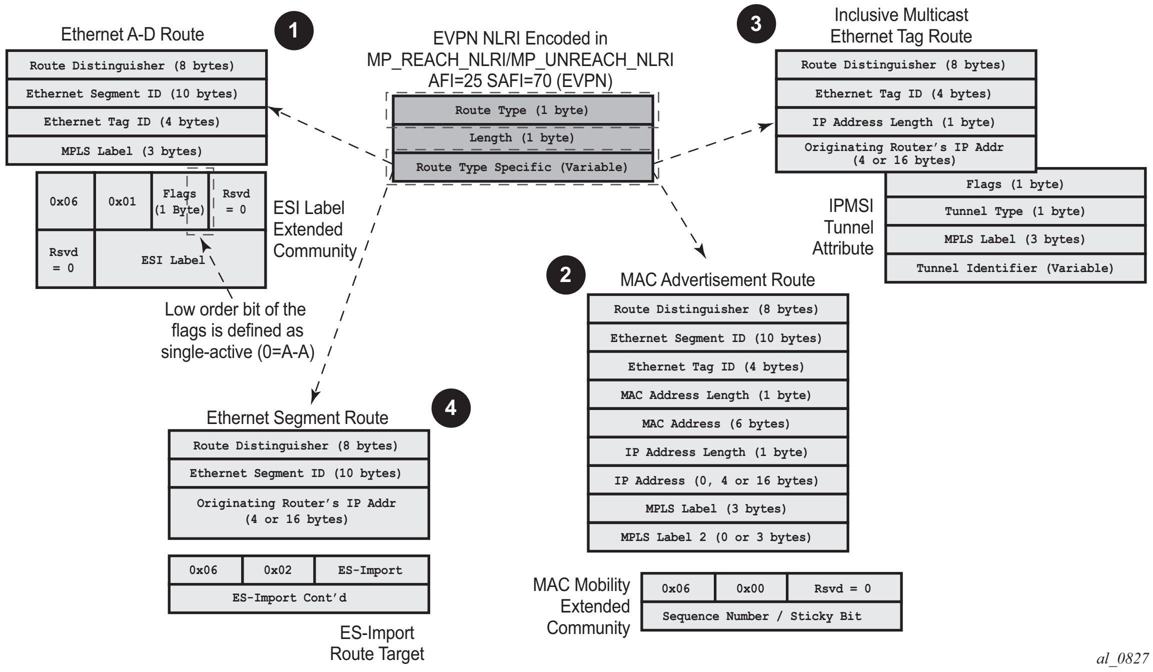
EVPN-MPLS is standardized in RFC 7432, *BGP MPLS-Based Ethernet VPN*, as a Layer 2 VPN technology that can supplement VPLS for E-LAN services. Besides the optimizations introduced by EVPN, a significant number of service providers offering E-LAN services today are requesting EVPN for their multi-homing capabilities. EVPN supports all-active multi-homing (per-flow load-balancing multi-homing) as well as single-active multi-homing (per-service load-balancing multi-homing). In addition to those superior multi-homing capabilities, EVPN also provides a number of significant benefits, such as:

- IP-VPN-like operation and control for E-LAN services.
- Reduction and (in some cases) suppression of the Broadcast, Unknown unicast, and Multicast (BUM) traffic in the network.
- Simple provisioning and management.
- New set of tools to control the distribution of MAC addresses and Address Resolution Protocol (ARP) entries in the network.

The EVPN for Virtual eXtensible Local Area Network (VXLAN) tunnels (Layer 2) chapter focuses on the use of EVPN as a control plane for VXLAN tunnels, whereas this chapter provides configuration guidelines for EVPN when used for MPLS tunnels. Similar to EVPN-VXLAN services, VPLS services with EVPN for MPLS tunnels are referred to as EVPN-MPLS services.

As a reference, the EVPN route types and NLRIs (Network Layer Reachability Information messages) used by the EVPN family in RFC 7432 are shown in Figure 40.

Figure 40 **EVPN Route Types and NLRIs**


1
Ethernet A-D Route
Route Distinguisher (8 bytes)
Ethernet Segment ID (10 bytes)
Ethernet Tag ID (4 bytes)
MPLS Label (3 bytes)
0x06
0x01
Flags (1 Byte)
Rsvd = 0
Rsvd = 0
ESI Label
ESI Label Extended Community
Low order bit of the flags is defined as single-active (0=A-A)
EVPN NLRI Encoded in MP_REACH_NLRI/MP_UNREACH_NLRI AFI=25 SAFI=70 (EVPN)
Route Type (1 byte)
Length (1 byte)
Route Type Specific (Variable)
3
Inclusive Multicast Ethernet Tag Route
Route Distinguisher (8 bytes)
Ethernet Tag ID (4 bytes)
IP Address Length (1 byte)
Originating Router's IP Addr (4 or 16 bytes)
IPMSI Tunnel Attribute
Flags (1 byte)
Tunnel Type (1 byte)
MPLS Label (3 bytes)
Tunnel Identifier (Variable)
2
MAC Advertisement Route
Route Distinguisher (8 bytes)
Ethernet Segment ID (10 bytes)
Ethernet Tag ID (4 bytes)
MAC Address Length (1 byte)
MAC Address (6 bytes)
IP Address Length (1 byte)
IP Address (0, 4 or 16 bytes)
MPLS Label (3 bytes)
MPLS Label 2 (0 or 3 bytes)
MAC Mobility Extended Community
0x06
0x00
Rsvd = 0
Sequence Number / Sticky Bit
4
Ethernet Segment Route
Route Distinguisher (8 bytes)
Ethernet Segment ID (10 bytes)
Originating Router's IP Addr (4 or 16 bytes)
0x06
0x02
ES-Import
ES-Import Cont'd
ES-Import Route Target
al_0827

When no EVPN multi-homing is used in the network, only the base routes are used. Route types 2 and 3 are considered the base and mandatory routes:

- Route type 2 - MAC/IP route: This route advertises MAC addresses to be installed in the remote FDBs, or MAC/IP address pairs to be installed in the remote proxy-ARP/ND (Neighbor Discovery) tables.
- Route type 3 - Inclusive multicast route: This route advertises the multicast tree that the advertising PE intends to use for sending BUM traffic for an EVPN Instance (EVI). Ingress Replication, Point-to-multipoint multicast Label Distribution Protocol (P2MP mLDP), and composite tunnels are supported as tunnel types in route type 3 when BGP-EVPN MPLS is enabled. The ingress replication information, as well as the downstream MPLS label (for remote PEs to send BUM traffic to the advertising PE) are encoded in the Provider Multicast Service Interface Tunnel Attribute (PTA).

When EVPN multi-homing is used in an EVI, routes type 1 and 4 are used (where type 1 has two different purposes):

- Route type 1 - Auto-discovery per Ethernet segment (AD per ES) route: This route is advertised per ES from the PE, carries the Ethernet Segment Identifier (ESI) label (used for split-horizon) in multi-homing mode, and can affect procedures such as the Designated Forwarder (DF) election, as well as the aliasing/backup path/mass withdrawal on remote PEs.
- Route type 1 - Auto-discovery per EVPN instance (AD per-EVI) route: This route allows the remote PEs to provide aliasing and a backup path to the PEs part of the ES.
- Route type 4 - Ethernet Segment (ES) route: This route advertises a local configured ES. The exchange of this route can discover remote PEs that are part of the same ES and the DF election algorithm among them.

The AD per-EVI, MAC/IP, and inclusive multicast routes are considered service-level BGP-EVPN routes. Their RT/RD (Route-Target/Route-Distinguisher) are taken from the VPLS configuration.

The AD per-ES and the ES routes are considered base-level BGP-EVPN routes. However, their RT/RD are taken differently:

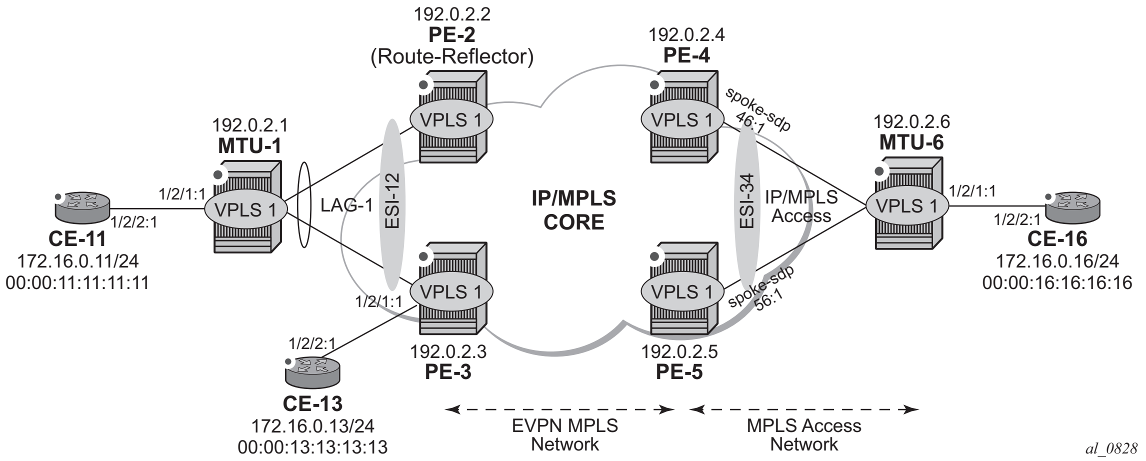
- The ES route RD is taken from the **service>system>bgp-evpn** configuration. The ES route RT is auto-derived from the Ethernet segment.
- The AD per-ES route RD is taken from the system level RD or service level RD. The RT extended community is taken from the service level RT or an RT set for the services defined on the Ethernet segment.

Configuration

This section describes the configuration of EVPN-MPLS for Layer 2 services on SR OS, as well as the available troubleshooting and show commands, and EVPN multi-homing.

Figure 41 shows the topology used throughout this chapter. The network consists of a core with four EVPN PEs (PE-2, PE-3, PE-4, and PE-5) and two MTU devices that are dual-homed to the EVPN network. For MTU-1, all-active multi-homing is used, whereas MTU-6 is connected via single-active multi-homing to the EVPN network. Three CEs are connected to VPLS 1 in MTU-1, PE-3, and MTU-6 in order to test the connectivity.

Figure 41 **EVPN-MPLS for VPLS Services**



As part of the network infrastructure configuration, the following settings and protocols must be added to the configuration before starting with the EVPN-specific configuration for the services:

- The ports interconnecting the four PEs in the core are configured as network ports (or hybrid) and will have router network interfaces defined in them. The ports on PE-2/PE-3 connected to MTU-1 can be access or hybrid ports, whereas the ports on PE-4/PE-5 connected to MTU-6 can be network or hybrid ports. In case of hybrid ports, no LACP can be configured.
- The four PEs in the core (as well as MTU-6 in the access MPLS network) are running IS-IS and establishing point-to-point adjacencies for the exchange of the system IP addresses.
- LDP is used as the MPLS protocol to signal transport tunnel labels among PE-2, PE-3, PE-4, PE-5, and MTU-6. There is no LDP running between MTU-1 and the rest of the network, that is, MTU-1 is a pure Ethernet aggregation device.
- EVPN uses MP-BGP for exchanging reachability at service level. Therefore, BGP peering sessions must be established among the core PEs for the EVPN family. Although typically a separate router is used, in this chapter, PE-2 is used as BGP RR (route reflector) for EVPN routes. For example, the following output shows the configuration of BGP in the RR and one of the BGP clients. The relevant commands for EVPN are shown in bold.

  The configuration on the route reflector PE-2 is as follows:

```
configure
    router
        autonomous-system 64500
        bgp
            vpn-apply-import
            vpn-apply-export
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            split-horizon
            rapid-update evpn
            group "internal"
                family evpn
                cluster 1.1.1.1
                peer-as 64500
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
                neighbor 192.0.2.5
                exit
            exit
```

The BGP configuration on the clients PE-3, PE-4, and PE-5 is as follows:

```
configure
    router
        autonomous-system 64500
        bgp
            vpn-apply-import
            vpn-apply-export
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            split-horizon
            rapid-update evpn
            group "internal"
                family evpn
                peer-as 64500
                neighbor 192.0.2.2
                exit
            exit
```

Note: The **def-recv-evpn-encap** command is not used in the preceding configuration because the default MPLS configuration is sufficient to have a correct interpretation of the received EVPN encapsulations.

The EVPN encapsulation type can be configured as MPLS or VXLAN, as follows:

```
*A:PE-3# configure router bgp group "internal" neighbor 192.0.2.2 def-recv-evpn-encap
 - no def-recv-evpn-encap
 - def-recv-evpn-encap <encap-type>

 <encap-type>         : mpls|vxlan
```

The default EVPN encapsulation setting is as follows:

```
*A:PE-3# configure router bgp group "internal" neighbor 192.0.2.2
*A:PE-3>config>router>bgp>group>neighbor# info detail | match def-recv
                    no def-recv-evpn-encap
```

EVPN routes type 1 (auto-discovery per-EVI route), type 2 (MAC/IP route), type 3 (inclusive multicast route), and type 5 (IP-prefix route) are always sent with the RFC 5512, *the BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute*, BGP encapsulation extended community that indicates the associated encapsulation of the route. Because the use of this extended community is not mandatory in RFC 7432, the def-recv-evpn-encap command indicates to the system what encapsulation is associated with routes received without any encapsulation. When interoperating with third-party EVPN vendors in mixed MPLS and EVPN-VXLAN networks, this command should be revised accordingly.

EVPN-MPLS Configuration without Multi-Homing

After the base infrastructure (interfaces, IGP, LDP, BGP protocols) is configured, the service and EVPN can be enabled. When no multi-homing is used, the EVPN-MPLS configuration in a VPLS service looks similar to the configuration of EVPN-VXLAN for Layer 2, except for the commands related to the MPLS data plane. The following output shows the VPLS-1 configuration in PE-3 as an example:

```
configure
    service
        vpls 1 customer 1 create
            bgp
            exit
            bgp-evpn
                evi 1
                mpls
                    ingress-replication-bum-label
                    ecmp 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            sap 1/2/1:1 create
            exit
            sap lag-1:1 create
            exit
            no shutdown
```

Where the following commands are relevant for a basic EVPN configuration:

- **bgp** enables the context for the BGP configuration relevant to the service. If a manual (non-auto-derived) RD/RT, as well as import/export policies, are needed for the service, the commands in the **bgp** context must be configured. When **bgp-evpn** is enabled in a VPLS instance, other families are supported within the same service (bgp-ad and bgp-mh, not bgp-vpls). This **bgp** context configures the common BGP parameters for all the BGP families in the service. The **pw-template-binding** command is ignored for **bgp-evpn**. Even if the general BGP parameters for the service are auto-derived (as in this example), the **bgp** context must be enabled.

```
*A:PE-3>config>service>vpls# bgp ?
  - bgp
  - no bgp

 [no] pw-template-bi* + Configure pw-template bind policy
 [no] route-distingu* - Configure route distinguisher
 [no] route-target    - Configure route target
 [no] vsi-export      - VSI export route policies
 [no] vsi-import      - VSI import route policies
```

- **bgp-evpn evi <1..65535>** — The EVPN instance or EVI is a 2-byte identifier used for the auto-derivation of the service RD, service RT, and for the service-carving algorithm when multi-homing is used. The EVI can be used for both **bgp-evpn vxlan** and **bgp-evpn mpls** when the user needs to auto-derive the RD and RT for the service. The auto-derivation is always based on:
  - RD system-ip:evi
  - RT autonomous-system:evi

  The configured and operating RD/RT values can be checked with the following show command (in this example, the evi value is 1):

```
*A:PE-3# show service id 1 bgp

===============================================================================
BGP Information
===============================================================================
Vsi-Import           : None
Vsi-Export           : None
Route Dist           : None
Oper Route Dist      : 192.0.2.3:1
Oper RD Type         : derivedEvi
Rte-Target Import    : None                  Rte-Target Export: None
Oper RT Imp Origin   : derivedEvi            Oper RT Import   : 64500:1
Oper RT Exp Origin   : derivedEvi            Oper RT Export   : 64500:1
PW-Template Id       : None
-------------------------------------------------------------------------------
===============================================================================
```

Although not required for a basic BGP-EVPN MPLS configuration, some other parameters may be used at the **bgp-evpn** context level, when EVPN-MPLS services are deployed. Some examples are listed here:

- **bgp-evpn>cfm-mac-advertisement** must be enabled when eth-cfm is used across an EVPN-MPLS service among different PEs. If a Maintenance Endpoint (MEP) or Maintenance domain Intermediate Point (MIP) is configured in any of the SAP/SDP bindings in the VPLS and has to exchange eth-cfm packets with a remote MEP/MIP across the EVPN-MPLS core, this command must be enabled. In that way, the MEP/MIP MAC address can be advertised in EVPN (otherwise, the MEP/MIP MAC address would not be learned on remote EVPN-MPLS PEs and eth-cfm would not work correctly).
- **bgp-evpn>mac-advertisement** and **bgp-evpn>mac-duplication** — See the EVPN for VXLAN Tunnels (Layer 2) chapter for a description of these two commands.
- **bgp-evpn>mpls** must be enabled.

When two BGP instances are added to a VPLS service, both BGP-EVPN MPLS and BGP-EVPN VXLAN can be configured at the same time in the service. A maximum of two BGP instances are supported in the same VPLS, such that BGP-EVPN MPLS can use BGP instance 1 or 2, and EBGP-EVPN VXLAN can use BGP instance 1 only. In this chapter, only one BGP instance will be used: BGP-EVPN MPLS uses the default BGP instance (**bgp-instance 1**).

After the relevant **VPLS** parameters, **BGP** and **BGP-EVPN** attributes are added, the specific commands for **bgp-evpn mpls** can be configured as follows:

```
*A:PE-3>config>service>vpls>bgp-evpn>mpls# info
----------------------------------------------
                    ingress-replication-bum-label
                    ecmp 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
----------------------------------------------
```

- **ingress-replication-bum-label** controls whether the system will advertise different service labels for unicast and BUM traffic. If no EVPN multi-homing is configured in the network, this command can be disabled (**no ingress-replication-bum-label**) and the same MPLS label will be advertised for the unicast and BUM traffic for the VPLS instance. If EVPN multi-homing is configured in the PE, this command is strongly recommended to avoid potential transient issues. See the EVPN-MPLS Multi-Homing section.
- **ecmp** controls the number of remote PEs to which the local PE can load balance the unicast traffic. See the EVPN multi-homing section.
- **auto-bind-tunnel** controls the resolution of EVPN destinations to MPLS transport tunnels. This command is also in VPRN services and works in the same way.

- If the **auto-bind-tunnel resolution any** is configured, as in the example, EVPN destinations in the service are resolved based on the best tunnel in the Tunnel Table Manager (TTM). For instance, the following command shows the existing EVPN destinations for VPLS 1 in PE-3. The EVPN-MPLS destination (Termination Endpoint (TEP) 192.0.2.2, label 262140) is resolved to an LDP transport tunnel because the (best) LDP tunnel to 192.0.2.2 shown in the **show router tunnel-table** is LDP. If there was more than one tunnel type in the TTM to 192.0.2.2, the system would pick the lowest **Pref** (preference) tunnel.

```
*A:PE-3# show service id 1 evpn-mpls

===============================================================================
BGP EVPN-MPLS Dest
===============================================================================
TEP Address     Egr Label     Num. MACs   Mcast           Last Change
                 Transport
-------------------------------------------------------------------------------
192.0.2.2       262140        0           Yes             05/04/2017 08:09:05
                ldp
192.0.2.4       262140        0           Yes             05/04/2017 08:09:05
                ldp
192.0.2.5       262140        0           Yes             05/04/2017 08:09:05
                ldp
-------------------------------------------------------------------------------
Number of entries : 3
-------------------------------------------------------------------------------
===============================================================================
---snip---

*A:PE-3# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.2/32      ldp       MPLS  65540     9        192.168.23.1   10
192.0.2.4/32      ldp       MPLS  65537     9        192.168.34.2   10
192.0.2.5/32      ldp       MPLS  65539     9        192.168.35.2   10
192.0.2.6/32      ldp       MPLS  65538     9        192.168.34.2   20
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
```

- If resolution is set to **any**, the following tunnel types are selected in order of preference: RSVP, LDP, Segment Routing, and BGP. The user can configure the preference of the segment-routing tunnel type in the TTM for a specific IGP instance.
- If one or more explicit tunnel types are specified using the resolution-filter option, then only these tunnel types will be selected again following the TTM preference.

- The user must set the resolution to filter to activate the list of tunnel-types configured under resolution-filter.

Although not shown in the **bgp-evpn mpls** basic configuration for PE-3, there are other parameters that can be modified:

```
*A:PE-3# configure service vpls 1 bgp-evpn mpls
  - mpls

      auto-bind-tunn* + Configure BGP EVPN mpls auto-bind-tunnel
      bgp-instance    - Configure BGP instance
 [no] control-word    - Enable/disable setting the CW bit in the label message
      ecmp            - Configure maximum ECMP routes information
 [no] entropy-label   - Enable/disable use of entropy-label
 [no] force-vlan-vc-* - Forces vlan-vc-type forwarding in the data-path
 [no] ingress-replic* - Use the same label as the one advertised for unicast traffic
 [no] restrict-prote* - Enable/disable protected src MAC restriction
 [no] send-evpn-encap - Configure encapsulation for this service
 [no] shutdown        - Administratively Enable/Disable BGP-EVPN mpls
 [no] split-horizon-* - Configure a split-horizon-group
```

- **bgp-instance** defines the BGP instance: default **bgp** or **bgp 1** can be used for either BGP-EVPN MPLS or BGP-EPVN VXLAN; **bgp 2** can only be used for BGP-EVPN MPLS.
- **control-word** enables/disables the insertion of the control-word in the data path. The control-word is disabled by default and is not signaled in EVPN (based on RFC 7432) and has to be consistently configured in all the PEs in the network. The use of the **control-word** prevents packet-reordering from happening in P routers that misinterpret the first nibble of the payload in the packets they receive. In some third-party EVPN vendors, the control-word is enabled by default, so it is recommended to enable it when interoperating with other vendors.
- **entropy-label** enables the use of entropy labels. This is beyond the scope of this chapter.
- **force-vlan-vc-forwarding** allows the system to preserve the vlan-id and pbits of the service-delimiting qtag in a new tag added in the customer frame before sending it to the EVPN core. This command may be used with the **sap ingress vlan-translation** command: the configured translated vlan-id will be sent to the EVPN binds, as opposed to the service-delimiting tag vlan-id. If the ingress SAP/SDP-binding is null encapsulated, the output vlan-id and pbits will be zero.
- **restrict-protected-src** is by default disabled. When enabled, all packets entering the object will be verified not to contain a protected source MAC address. In combination with the parameter **discard-frame**, the packets that contain a protected MAC address will be discarded and an alarm is generated.

- **send-evpn-encap** configures the encapsulation to be advertised with the EVPN routes for the service. The encapsulation is encoded in RFC5512-based tunnel encapsulation extended communities. When configured in the **bgp-evpn>mpls** context, the supported options are none (no send-evpn-encap), mpls, mplsoudp, or both.
- **shutdown** enables/disables the use of MPLS for EVPN. When **mpls no shutdown** is issued, a BGP route-refresh message is sent for the EVPN family.
- **split-horizon-group** <*group-name*> configures an explicit split-horizon-group (SHG) for all the EVPN destinations that can be shared with other SAP/SDP-bindings. See the VPLS to EVPN-MPLS migration and integration section.

After **bgp-evpn mpls** is configured in the service, and **no shutdown**, an inclusive multicast route is sent to the RR. The remote PEs receiving and importing that route will create an EVPN destination to the sending PE. An EVPN destination is identified by a TEP and MPLS label. Use the following show commands to view the service and the EVPN destinations created:

```
show service evpn-mpls
show service id 1 evpn-mpls
show service id 1 bgp-evpn
```

An example of the output is shown for PE-2 when there is no traffic in the network. Therefore, only inclusive multicast routes have been exchanged among the four PEs.

```
*A:PE-2# show service evpn-mpls

===============================================================================
EVPN MPLS Tunnel Endpoints
===============================================================================
EvpnMplsTEP Address EVPN-MPLS Dest      ES Dest             ES BMac Dest
-------------------------------------------------------------------------------
192.0.2.3           1                   0                   0
192.0.2.4           1                   0                   0
192.0.2.5           1                   0                   0
-------------------------------------------------------------------------------
Number of EvpnMpls Tunnel Endpoints: 3
-------------------------------------------------------------------------------
===============================================================================


*A:PE-2# show service id 1 evpn-mpls

===============================================================================
BGP EVPN-MPLS Dest
===============================================================================
TEP Address     Egr Label     Num. MACs   Mcast           Last Change
                 Transport
-------------------------------------------------------------------------------
192.0.2.3       262140        0           Yes             05/04/2017 08:09:05
                ldp
192.0.2.4       262140        0           Yes             05/04/2017 08:09:05
                ldp
```

```
192.0.2.5       262140        0           Yes             05/04/2017 08:09:05
                ldp
-------------------------------------------------------------------------------
Number of entries : 3
-------------------------------------------------------------------------------
===============================================================================
===============================================================================
BGP EVPN-MPLS Ethernet Segment Dest
===============================================================================
Eth SegId                       Num. Macs               Last Change
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================
===============================================================================
BGP EVPN-MPLS ES BMAC Dest
===============================================================================
ES BMAC Addr                            Last Change
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================
*A:PE-2#

*A:PE-2# show service id 1 bgp-evpn

===============================================================================
BGP EVPN Table
===============================================================================
MAC Advertisement  : Enabled            Unknown MAC Route  : Disabled
CFM MAC Advertise  : Disabled
VXLAN Admin Status : Disabled           Creation Origin    : manual
MAC Dup Detn Moves : 5                  MAC Dup Detn Window: 3
MAC Dup Detn Retry : 9                  Number of Dup MACs : 0
MAC Dup Detn BH    : Disabled
IP Route Advert    : Disabled

EVI                : 1
Ing Rep Inc McastAd: Enabled
Accept IVPLS Flush : Disabled
Send EVPN Encap    : Enabled

-------------------------------------------------------------------------------
Detected Duplicate MAC Addresses             Time Detected
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
===============================================================================
===============================================================================
BGP EVPN MPLS Information
===============================================================================
Admin Status       : Enabled
Force Vlan Fwding  : Disabled           Control Word       : Disabled
Split Horizon Group: (Not Specified)
Ingress Rep BUM Lbl: Enabled            Max Ecmp Routes    : 2
Ingress Ucast Lbl  : 262141             Ingress Mcast Lbl  : 262140
Entropy Label      : Disabled
RestProtSrcMacAct  : none
Evpn Mpls Encap    : Enabled            Evpn MplsoUdp      : Disabled
===============================================================================

===============================================================================
BGP EVPN MPLS Auto Bind Tunnel Information
===============================================================================
Resolution         : any
Filter Tunnel Types: (Not Specified)
===============================================================================
```

When traffic is generated, the PEs will start learning MAC addresses and advertising them in BGP so that the remote PEs learn those MAC addresses against EVPN destinations. For instance, when CE-13 sends traffic, PE-3 learns its MAC address and advertises it. The remote PEs (for instance, PE-2) will learn the MAC address and associate it with their EVPN destination to PE-3 (192.0.2.3:262141 in this example):

```
*A:PE-2# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         00:00:11:11:11:11 sap:lag-1:1              L/106    05/05/17 01:50:54
1         00:00:13:13:13:13 eMpls:                   Evpn     05/05/17 01:50:54
                            192.0.2.3:262141
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
```

When the **ingress-replication-bum-label** is enabled in the PEs, the advertisement of MAC addresses will create new EVPN destinations, because the label is different from the one previously sent by the inclusive multicast route that created an EVPN destination. In the preceding example, when PE-3 advertises the CE-13 MAC address, PE-2 will create a new binding (see in the following output in bold) that shows one MAC address that is not Mcast (multicast) capable:

```
*A:PE-2# show service id 1 evpn-mpls

===============================================================================
BGP EVPN-MPLS Dest
===============================================================================
TEP Address     Egr Label     Num. MACs   Mcast           Last Change
                 Transport
-------------------------------------------------------------------------------
192.0.2.3       262140        0           Yes             05/05/2017 01:50:54
                ldp
192.0.2.3       262141        1           No              05/05/2017 02:03:23
                ldp
192.0.2.4       262140        0           Yes             05/04/2017 08:09:05
                ldp
192.0.2.5       262140        0           Yes             05/04/2017 08:09:05
                ldp
-------------------------------------------------------------------------------
```

```
Number of entries : 4
-------------------------------------------------------------------------------
===============================================================================
```

When an EVPN-MPLS destination or MAC address is not created/installed correctly, the user may check the BGP-EVPN routes received and the routes kept in the RIB. The routes that the PE receives are shown when **debug router bgp update** is enabled. These routes are shown even before any BGP processing is carried out.

```
*A:PE-2#
4 2017/05/04 11:41:16.27 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 88
    Flag: 0x90 Type: 14 Len: 44 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.3
        Type: EVPN-MAC Len: 33 RD: 192.0.2.3:1 ESI: ESI-0, tag: 0, mac len: 48
                      mac: 00:00:13:13:13:13, IP len: 0, IP: NULL, label1: 4194256
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:64500:1
        bgp-tunnel-encap:MPLS
"


*A:PE-2# show router bgp routes evpn mac
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN MAC Routes
===============================================================================
Flag  Route Dist.         MacAddr           ESI
      Tag                 Mac Mobility      Label1
                          Ip Address
                          NextHop
-------------------------------------------------------------------------------
u*>i  192.0.2.3:1         00:00:13:13:13:13 ESI-0
      0                   Seq:0             LABEL 262141
                          N/A
                          192.0.2.3

-------------------------------------------------------------------------------
Routes : 1
===============================================================================
```

If the route is successfully imported, it can be shown in the RIB (**show router bgp routes** commands). The route shown in the debug and the same route in a show command do not necessarily have the same label value. The reason for this expected mismatch is that the debug command shows the complete 24-bit field value because the route is shown before BGP can decide and decipher whether the label value is an MPLS label (high-order 20-bits of the label field) or a VNI (all 24 bits of the Label field for VXLAN). When the label in the debug command (4194256) is divided by 16 (2^4), the result is the MPLS label (262141), as follows: 4194256:16=262141.

VPLS to EVPN-MPLS Integration

The SR OS EVPN implementation supports draft-ietf-bess-evpn-vpls-seamless-integ-00 so that EVPN-MPLS and VPLS can be integrated into the same network and within the same service.

The following behavior enables the integration of EVPN and SDP-bindings in the same VPLS network:

- Systems with EVPN endpoints and SDP-bindings to the same far-end bring down the SDP-bindings.
  - SR OS will allow the establishment of an EVPN destination and an SDP-binding to the same far-end but the SDP-binding will be kept operationally down. Only the EVPN endpoint will be operationally up. This is true for spoke-SDPs (manual and BGP-AD) and mesh-SDPs. It is also true between VXLAN and SDP-bindings.
  - If there is an EVPN endpoint to a specified far-end and a spoke-SDP establishment is attempted, the spoke-SDP will be set up but kept down with an operational flag indicating that there is an EVPN route to the same far-end.
  - If there is a spoke-SDP and a valid/used EVPN route arrives, the EVPN endpoint will be set up and the spoke-SDP will be brought down with an operational flag indicating that there is an EVPN route to the same far-end.
  - In the case of an SDP-binding and EVPN endpoint to different far-end IPs on the same remote PE, both links will be up. This can happen if the SDP-binding is terminated in an IPv6 address or IPv4 address different from the system address where the EVPN endpoint is terminated.

  The following example illustrates the preceding description. A spoke-SDP is added to the VPLS 1 configuration on PE-2:

```
configure
    service
        sdp 24 mpls create
```

```
            far-end 192.0.2.4
            ldp
            no shutdown
        exit
        vpls 1
            spoke-sdp 24:1 create
            exit
        exit
```

The service configuration on PE-4 is as follows:

```
configure
    service
        sdp 42 mpls create
            far-end 192.0.2.2
            ldp
            no shutdown
        exit
        sdp 46 mpls create
            far-end 192.0.2.6
            ldp
            no shutdown
        exit
        vpls 1 customer 1 create
            bgp
            exit
            bgp-evpn
                evi 1
                mpls
                    ingress-replication-bum-label
                    ecmp 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            spoke-sdp 42:1 create
            exit
            spoke-sdp 46:1 create
            exit
            no shutdown
```

Spoke SDP 24:1 is operationally down, as can be verified as follows:

```
*A:PE-2# show service id 1 base

---snip---

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:lag-1:1                              q-tag        1518    1518    Up   Up
sdp:24:1 S(192.0.2.4)                    Spok         0       8978    Up   Down
===============================================================================
```

Spoke SDP 24:1 is down because of an EVPN route conflict, as indicated by the flags:

```
*A:PE-2# show service id 1 sdp 24 detail | match Flag context all
Flags              : PWPeerFaultStatusBits
                     EvpnRouteConflict
```

- The user can add spoke-SDPs and all the EVPN-MPLS endpoints in the same SHG.
  - A CLI command exists in the **bgp-evpn>mpls** context so that the EVPN-MPLS endpoints can be added to an SHG.
  - The **bgp-evpn mpls split-horizon-group** must reference a user-configured split-horizon-group. User-configured split-horizon-groups can be configured within the service context.
  - The same group-name can be associated with SAPs, spoke-SDPs, PW-templates, PW-template-bindings, and EVPN-MPLS endpoints.
  - If the **split-horizon-group** command in **bgp-evpn>mpls** is not used, the default split-horizon-group (in which all the EVPN endpoints are) is still used, but it will not be possible to refer to it on SAPs/spoke-SDPs.
- The system disables the advertisement of MAC addresses learned on spoke-SDPs/SAPs that are part of an EVPN split-horizon-group.
  - When the SAPs or spoke-SDPs (manual or BGP-AD-discovered) are configured within the same SHG as the EVPN endpoints, MAC addresses will still be learned on them, but will not be advertised in EVPN.
  - The preceding statement is also true if proxy-ARP/ND is enabled and an IP->MAC address pair is learned on a sap/sdp-binding that belongs to the EVPN SHG.
  - The SAPs and/or spoke-SDPs added to an EVPN SHG should not be part of any EVPN multi-homed ES. If that happened, the PE would still advertise the AD per-EVI route for the SAP and/or spoke-SDP, attracting EVPN traffic that could not be forwarded to that SAP and/or SDP-binding.
  - Similar to the preceding statement, an SHG composed of SAPs/SDP-bindings used in a BGP-MH site should not be configured under **bgp-evpn>mpls>split-horizon-group**. This misconfiguration would prevent traffic being forwarded from the EVPN to the BGP-MH site, regardless of the DF/Non-DF state.

An example of a shared SHG configuration on PE-2 is as follows. Because the SAP and EVPN-MPLS are in the same SHG, no MAC addresses learned over SAP 1/2/1:2 will be advertised in EVPN (not even static MACs).

```
configure
    service
        vpls 2 customer 1 create
            split-horizon-group "CORE" create
```

```
exit
bgp
exit
bgp-evpn
    mpls
        split-horizon-group "CORE"
        ingress-replication-bum-label
        ecmp 2
        auto-bind-tunnel
            resolution any
        exit
        no shutdown
    exit
exit
sap 1/2/1:2 split-horizon-group "CORE" create
exit
sap lag-1:2 create
exit
no shutdown
```

EVPN-MPLS Multi-Homing

SR OS supports EVPN multi-homing as per RFC 7432.

The EVPN multi-homing implementation is based on the concept of the ES. An ES is a logical structure that can be defined in one or more PEs and identifies the CE (or access network) multi-homed to the EVPN PEs. An ES is associated with a port, LAG, or SDP object, and is shared by all the services defined on those objects.

Each ES has a unique identifier called ESI (Ethernet Segment Identifier) that is 10 bytes and is manually configured. The ESI is advertised in the control plane to all the PEs in an EVPN network; therefore, it is very important to ensure that the 10-byte ESI value is unique throughout the entire network. Single-homed CEs are assumed to be connected to an ES with ESI = 0 (single-homed ESs are not explicitly configured).

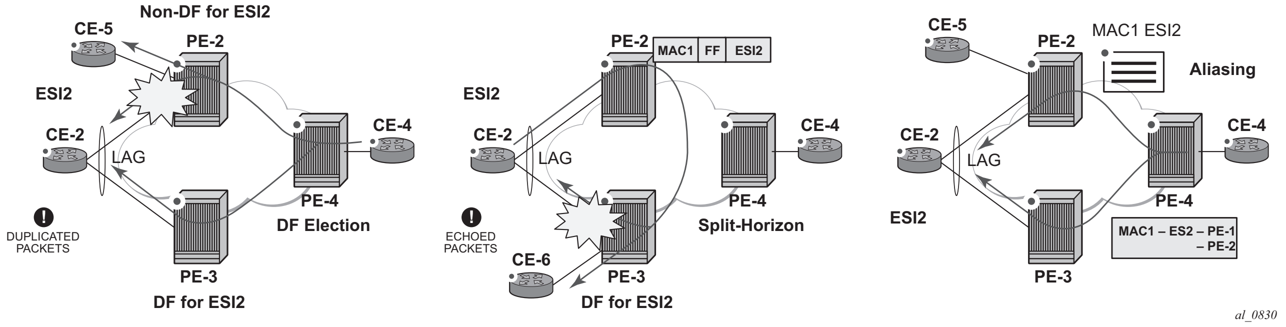
The ES is part of the base BGP-EVPN configuration and is not applied to any EVPN-MPLS service, by default. An ES can be shared by multiple services; the association of a specific SAP or spoke-SDP to an ES is automatically made when the SAP is defined in the same LAG or port configured in the ES, or when the spoke-SDP is defined in the same SDP configured in the ES. The following sections show the configuration of:

- an all-active multi-homing ES with a LAG associated with it
- a single-active multi-homing ES linked to an SDP

All-Active Multi-Homing Concepts

EVPN all-active multi-homing is built around three concepts: DF election, split-horizon (with an ESI-label), and aliasing, as shown in Figure 42, from left to right.

Figure 42 **EVPN-MPLS All-Active Multi-Homing Concepts**



- With DF election, when PE-4 sends BUM traffic to the remote ES (CE-2), only one PE segment sends the BUM packets to the ES (PE-3 is the DF in the preceding example, and is elected to send BUM packets to CE-2). The non-DF, PE-2, removes the LAG SAP from the default multicast list (PE-2 does not bring CE-2 down, because it still needs to send upstream/downstream unicast traffic). PE-2 and PE-3 elect a DF for each service, based on the ES routes and the service-carving algorithm.
- With split-horizon, the PE part of the ES (PE-3 in the preceding example) identifies the BUM packets coming from the PE for the remote (PE-2), but within the same ES (ESI-2), and filters the packets so that they are not sent back to the ES, creating duplication. When PE-2 (non-DF) sends BUM traffic to PE-3 (DF), it uses a special MPLS label in the data path that PE-3 previously advertised for ESI-2 in an AD per-ES route. When PE-3 does an ingress lookup, it recognizes the ESI-label and filters the traffic (PE-3 still sends the BUM traffic to other SAPs/SDP-bindings).
- With aliasing, remote PEs that are not part of the ES can load-balance unicast traffic to all the PEs that are part of the ES, irrespective of from which PE a destination MAC address was learned. PE-4 will create an EVPN destination to ESI-2 that will be resolved to the two next-hops: PE-2 and PE-3. Unicast load-balancing will happen as long as ECMP > 1 is enabled in PE-4.

Nokia recommends the use of **ingress-replication-bum-label** on the PEs that are part of an all-active ES. In an all-active multi-homing scenario, if a specified MAC address (for example, the CE-2 MAC address in the left-hand-side diagram), is not learned yet in a remote PE (for example, PE-4), but is known in the two PEs of the ES (for example, PE-2 and PE-3), the latter PEs might send duplicated packets to the CE.

This issue is solved by the use of **ingress-replication-bum-label** in PE-2 and PE-3. If configured, PE-2/PE-3 will know that the received packet is an unknown unicast packet; therefore, the Non-DF (PE-2) will not send the packet to CE-2 and there will not be duplication.

All-Active Multi-Homing Configuration

The all-active multi-homing configuration example is based on Figure 41.

MTU-1 is connected to the EVPN network using all-active multi-homing. According to RFC 7432, MTU-1 will be able to send traffic to both PEs for VPLS-1. Regular LAG load-balancing is used in MTU-1. Remote PEs such as PE-4 or PE-5 will be able to load-balance the unicast traffic to PE-2 and PE-3. PE-2 and PE-3 will discover that both are part of ESI-12 (due to the exchange of ES routes) and will elect a DF for VPLS-1. The non-DF for VPLS-1, in this case PE-2, will remove lag-1:1 from the VPLS-1 default multicast list. Also, when PE-2 and PE-3 send BUM traffic to each other, they will insert an ESI-label so that they can identify that the source of the BUM packet is ESI-12.

The following output shows the configuration of ESI-12 in PE-2 and PE-3, as well as the LAG interfaces for all-active multi-homing (see Figure 41). The configuration of LAG-1 in MTU-1 is also shown. Per RFC 7432, only a CE/MTU with a LAG can be connected to an all-active multi-homing ES. No other configuration is permitted on the CE for all-active multi-homing.

The configuration of LAG 1 on MTU-1 is as follows:

```
configure
    lag 1
        mode access
        encap-type dot1q
        port 1/1/1
        port 1/1/2
        lacp active administrative-key 32768
        no shutdown
```

LAG 1 is configured as follows on PE-2:

```
configure
    lag 1
        mode access
        encap-type dot1q
        port 1/1/2
        lacp active administrative-key 1 system-id 00:00:00:00:02:03
        no shutdown
```

LAG 1 is configured as follows on PE-3:

```
configure
    lag 1
        mode access
        encap-type dot1q
        port 1/1/1
        lacp active administrative-key 1 system-id 00:00:00:00:02:03
        no shutdown
```

Ethernet segment “ESI-12” is configured in the service **system bgp-evpn** context on PE-2 and PE-3, as follows:

```
configure
    service
        system
            bgp-evpn
                ethernet-segment "ESI-12" create
                    esi 01:00:00:00:00:12:00:00:00:01
                    es-activation-timer 3
                    service-carving
                        mode auto
                    exit
                    multi-homing all-active
                    lag 1
                    no shutdown
                exit
```

When configuring an ES, the following must be considered:

- Any EVPN parameter that is not specific to any particular VPLS service, and is common to all the EVIs, is configured in a base BGP-EVPN instance located at **config>service>system>bgp-evpn**. In this base instance, the following attributes may be configured:
  - **ethernet-segments**
  - the base BGP-EVPN instance **route-distinguisher** that will be used for the ES routes. If this **route-distinguisher** is not configured, by default a type-1 RD will be derived as system-ip:0, as shown in the command help:

```
*A:PE-2>config>service>system>bgp-evpn# route-distinguisher
  - no route-distinguisher
  - route-distinguisher <rd>

 <rd>                 : <ip-addr:comm-val>
                        ip-addr        - a.b.c.d
                        comm-val       - [0..65535]
                        default: system-ip:0
```

- The ES must be configured with a name and can contain the following parameters when configured for all-active multi-homing:

  - **esi** — 10-byte identifier that represents the ES in the BGP control plane. The same ESI must be configured in all the PEs connected to the same CE/MTU (using a unique value that cannot be associated with any other CE/MTU/access network). RFC 7432 defines five different types of ESI. In SR OS, the **type** byte, as well as the other 9 bytes can be arbitrarily configured.
  - **multi-homing all-active** — This command indicates that the ES is in all-active mode.
  - **lag** <*lag-id*> — The LAG connected to the CE/MTU must be added to the ES. In this example, lag-1 is added to ESI-12, on both PE-2 and PE-3. Although a different LAG-id may have been assigned to the same ES on PE-2 and PE-3, PE-2 and PE-3 must have the same configuration on the ES LAG; that is, encap-type. Also, if LACP is added (it is not mandatory), both PEs must have the same admin-key, system-id, and system-priority. MTU-1 will see PE-2 and PE-3 as a single LAG peer. For all-active multi-homing, only the **lag** option is accepted by the system; **port** or **sdp** are not accepted.
  - [**no**] **shutdown** — This command controls the administrative state of the ES.
- The preceding parameters are the minimum necessary so that the ES can be activated. In addition to those parameters, there are a few more that the user can configure if requiring values different from the default ones:
  - **es-activation-timer** [**0..100**] can be configured at **redundancy>bgp-evpn-multi-homing>es-activation-timer** or at **service>system>bgp-evpn>eth-seg>es-activation-timer** level (the most specific value is used).

    The **es-activation-timer** operation is as follows:

    - Upon reception of an ES, AD per-ES/EVI route update/withdrawal for a local ESI, the DF-candidate list of IPs is updated and the DF election algorithm is run without waiting for any timer.
    - If the result of the DF election requires the PE to be promoted from non-DF to DF, the **es-activation-timer** will start, and only after its expiration will the PE add the SAP to the default-multicast list. Transitions from non-DF to non-DF, or from DF to non-DF, are immediate and do not wait for any timer.
    - This use of an **es-activation-timer** value minimizes the risks of loops and packet duplication due to **transient** multiple DFs.
    - The same **es-activation-timer** must be configured in all the PEs that are part of the same ESI. The user must configure either a long timer to minimize the risks of loops/duplication, or **es-activation-timer**=0 to speed up the convergence for NDF→ DF transitions. The default value is 3 seconds.

– **service-carving —** As defined in RFC 7432, service-carving controls the distribution of DF/non-DF roles across the different services defined in an ES.

```
*A:PE-2>config>service>system>bgp-evpn>eth-seg>service-carving# mode
  - mode {manual|auto}

 <manual|auto>         : auto|manual|off


*A:PE-2>config>service>system>bgp-evpn>eth-seg>service-carving# manual
  - manual

 [no] evi             - Configure EVI range (primary for non-preference based DF
                        election and lowest-preference for preference based DF
                        election)
 [no] isid            - Configure ISID range (primary for non-preference based DF
                        election and lowest-preference for preference based DF
                        election)
 [no] preference      + Configure DF preference election information
```

As shown above, **service-carving** has three different modes:

– **service-carving mode auto** (default) — The DF election algorithm will run the function [V(evi) mod N(peers) = i(ordinal)] to know who the DF for a specified service and ESI is. In this example, ESI-12 is configured with mode **auto**; therefore, for VPLS-1 (with EVI-1), PE-3 will be elected as DF because evi(1) mod (2)peers = 1, and the ordinal 1 corresponds to the second lowest IP, PE-3. The algorithm takes the configured **evi** in the service; therefore, the **evi** is mandatory, and for the same service must match in all the PEs that are part of the ES. This guarantees that the election algorithm is consistent across all the PEs of the ESI.

– **service-carving mode manual —** The user can manually decide for which **evi** identifiers the PE is DF or **primary: service-carving mode manual / manual evi <start> [to <*to*>] primary**. The PE will be non-DF for the non-specified EVIs. If **service-carving mode manual** is configured, but no range is defined, all the services are considered to be non-DF. If a range is configured, but the **service-carving** is not **mode manual**, the range has no effect. Only two PEs are supported when **service-carving mode manual** is configured.

– **service-carving mode off** — The lowest originator IP will win the election for a specified service and ES.

– Because the **evi** is used for the service-carving algorithm, it must always be configured in a service with SAPs/SDP bindings created in an ES, irrespective of the service-carving mode (service-carving off, auto, or manual).

Although not configured as part of the ES, the **config>redundancy>bgp-evpn-multi-homing>boot-timer** allows the necessary time for the control plane protocols to come up after the PE has rebooted, and before bringing up the ESs and running the DF algorithm. Some considerations about the boot-timer:

- The **boot-timer** should use a value long enough to allow the IOMs and BGP sessions to come up before exchanging ES routes and run the DF election for each EVI (it is 10 s, by default).
- The **boot-timer** runs per EVI on the ESs in the system. While **system-up-time < boot-timer**, the system will not run the DF election for any EVI. When the boot-timer expires, the DF election for the EVI is run and, if the system is elected DF for the EVI, the **es-activation-timer** will start.
- The system will not advertise ES routes until the boot timer expires. This guarantees that the peer ES PEs do not run the DF election either, until the PE is ready to become the DF, if needed.
- The following show command displays the configured **boot-timer**, as well as the remaining timer if the system is still in boot stage.

```
*A:PE-2# show redundancy bgp-evpn-multi-homing

===============================================================================
Redundancy BGP EVPN Multi-homing Information
===============================================================================
Boot-Timer              : 10 secs
Boot-Timer Remaining    : 0 secs
ES Activation Timer     : 3 secs
===============================================================================
*A:PE-2#
```

After ESI-12 is configured in PE-2 and PE-3, the lag-1 SAPs in both PEs can be added to the VPLS-1 service. Until the ESI-12 is successfully enabled, the LAG SAPs will be kept down with a **StandByForMHProtocol** flag. This is illustrated in the following example for PE-2.

```
*A:PE-2# configure service system bgp-evpn ethernet-segment "ESI-12" shutdown
*A:PE-2# configure service vpls 1 sap lag-1:1 create


*A:PE-2# show service id 1 sap lag-1:1 detail | match "  Oper State"
Admin State        : Up                       Oper State        : Down


*A:PE-2# show service id 1 sap lag-1:1 detail | match Flag
Flags              : StandByForMHProtocol


*A:PE-2# configure service system bgp-evpn ethernet-segment "ESI-12" no shutdown


*A:PE-2# show log log-id 99

===============================================================================
Event Log 99
```

```
===============================================================================
Description : Default System Log
Memory Log contents  [size=500   next event=118  (not wrapped)]

117 2017/05/05 13:52:44.77 UTC MINOR: SVCMGR #2203 Base
"Status of SAP lag-1:1 in service 1 (customer 1) changed to admin=up oper=up flags="
```

All-Active Multi-Homing Operation

To confirm that all-active multi-homing is working correctly for ESI-12, the user can use the following commands:

- **show service system bgp-evpn** — Shows the RD is used for the ES route.
- **show service system bgp-evpn ethernet-segment** — Shows all the ESs configured in the PE and their admin/operational status.
- **show service system bgp-evpn ethernet-segment name ESI-12 evi 1** — Shows the DF candidate PEs for EVI 1 and whether the system is DF for EVI.
- **show service system bgp-evpn ethernet-segment name ESI-12 all** — Shows all the information related to a specific ESI.

The base BGP-EVPN information includes the RD:

```
*A:PE-2# show service system bgp-evpn

===============================================================================
System BGP EVPN Information
===============================================================================
Eth Seg Route Dist.               : <none>
Eth Seg Oper Route Dist.          : 192.0.2.2:0
Eth Seg Oper Route Dist Type      : default
Ad Per ES Route Target            : evi-rt
Leaf Label                        : 0
===============================================================================
*A:PE-2#
```

The following command shows the configured ESs in the PE and their status:

```
*A:PE-2# show service system bgp-evpn ethernet-segment

===============================================================================
Service Ethernet Segment
===============================================================================
Name                             ESI                           Admin     Oper
-------------------------------------------------------------------------------
ESI-12                           01:00:00:00:00:12:00:00:00:01 Enabled   Up
-------------------------------------------------------------------------------
Entries found: 1
===============================================================================
*A:PE-2#
```

The following command shows that PE-2 is not the DF and the DF candidate PEs for EVI 1 are PE-2 and PE-3:

```
*A:PE-2# show service system bgp-evpn ethernet-segment name "ESI-12" evi 1

===============================================================================
EVI DF and Candidate List
===============================================================================
EVI           SvcId         Actv Timer Rem      DF  DF Last Change
-------------------------------------------------------------------------------
1             1             0                   no  05/05/2017 13:52:45
===============================================================================

===============================================================================
DF Candidates                           Time Added
-------------------------------------------------------------------------------
192.0.2.2                               05/05/2017 13:52:45
192.0.2.3                               05/05/2017 13:52:47
-------------------------------------------------------------------------------
Number of entries: 2
===============================================================================
*A:PE-2#
```

The following command shows all information related to ESI-12 on PE-2:

```
*A:PE-2# show service system bgp-evpn ethernet-segment name "ESI-12" all

===============================================================================
Service Ethernet Segment
===============================================================================
Name                    : ESI-12
Eth Seg Type            : None
Admin State             : Enabled            Oper State         : Up
ESI                     : 01:00:00:00:00:12:00:00:00:01
Multi-homing            : allActive          Oper Multi-homing  : allActive
ES SHG Label            : 262142
Source BMAC LSB         : <none>
Lag Id                  : 1
ES Activation Timer     : 3 secs
Svc Carving             : auto               Oper Svc Carving   : auto
Cfg Range Type          : primary
===============================================================================

===============================================================================
EVI Information
===============================================================================
EVI                 SvcId               Actv Timer Rem      DF
-------------------------------------------------------------------------------
1                   1                   0                   no
-------------------------------------------------------------------------------
Number of entries: 1
===============================================================================

-------------------------------------------------------------------------------
DF Candidate list
-------------------------------------------------------------------------------
EVI                                     DF Address
-------------------------------------------------------------------------------
```

```
1                                       192.0.2.2
1                                       192.0.2.3
-------------------------------------------------------------------------------
Number of entries: 2
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
---snip---
```

The following command shows all information related to ESI-12 on PE-3:

```
*A:PE-3# show service system bgp-evpn ethernet-segment name "ESI-12" all

===============================================================================
Service Ethernet Segment
===============================================================================
Name                    : ESI-12
Admin State             : Enabled            Oper State         : Up
ESI                     : 01:00:00:00:00:12:00:00:00:01
Multi-homing            : allActive          Oper Multi-homing  : allActive
ES SHG Label            : 262142
Source BMAC LSB         : <none>
Lag Id                  : 1
ES Activation Timer     : 3 secs
Svc Carving             : auto               Oper Svc Carving   : auto
Cfg Range Type          : primary
===============================================================================
===============================================================================
EVI Information
===============================================================================
EVI                 SvcId               Actv Timer Rem      DF
-------------------------------------------------------------------------------
1                   1                   0                   yes
-------------------------------------------------------------------------------
Number of entries: 1
===============================================================================
-------------------------------------------------------------------------------
DF Candidate list
-------------------------------------------------------------------------------
EVI                                     DF Address
-------------------------------------------------------------------------------
1                                       192.0.2.2
1                                       192.0.2.3
-------------------------------------------------------------------------------
Number of entries: 2
-------------------------------------------------------------------------------
---snip---
```

The preceding commands show the ESI-12 configuration on both PEs and the result of the DF election for EVI 1.

The following output shows the ES route received on PE-2:

```
*A:PE-2#
1 2017/05/05 13:42:54.81 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 70
    Flag: 0x90 Type: 14 Len: 34 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.3
        Type: EVPN-Eth-Seg Len: 23 RD: 192.0.2.3:0
              ESI: 01:00:00:00:00:12:00:00:00:01, IP-Len: 4 Orig-IP-Addr: 192.0.2.3

    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:00:00:00:00:12:00
"
```

The ES RT as shown as target:00:00:00:00:12:00 in the extended community is auto-derived from the ESI bytes 2 to 7 (with the type byte being byte 1). Only PE-2 and PE-3 generate this RT and therefore import each other's ES route.

The following message in log 99 on PE-3 shows the result of the DF election for EVI 1.

```
4 2017/05/05 13:42:54.81 UTC MINOR: SVCMGR #2094 Base
"Ethernet Segment:ESI-12, EVI:1, Designated Forwarding state changed to:true"
```

The **show service system bgp-evpn ethernet-segment name ESI-12 all** command shows the ESI-label allocated to the PE: **ES SHG Label 262142** in the CLI output for PE-3. In this example, this label is allocated by PE-3 for ESI-12 (a different one is allocated per ESI) and advertised in the AD per-ES route for ESI-12. The following output shows the AD per-ES and AD per-EVI (for evi 1) routes sent by PE-3 and received by PE-2.

- The AD per-ES route can be identified by the **MAX-ET** in the ethernet-tag field (as per RFC 7432) and carries the ESI-label as well as the multi-homing mode (all-active in this case) in the ESI-label extended community (see Figure 40).

  Prior to release 14.0.R1, a separate AD per-ES route is sent per EVI. In release 14.0.R1, or later, the user can enable the aggregation of AD per-ES routes by using the following command: **config>service>system>bgp-evpn>ad-per-es-route-target evi-rt-set route-distinguisher** *ip-address*. If enabled, a single AD per-ES route with the associated RD and a set of EVI route-targets will be advertised (to a maximum of 128). When there are more than 128 EVIs defined in the ethernet-segment, more than one route will be sent by the system.
- The AD per-EVI route will have an eth-tag 0 and will carry the service label in the NLRI.

```
*A:PE-2#
2 2017/05/05 13:52:46.80 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 80
    Flag: 0x90 Type: 14 Len: 36 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.3
        Type: EVPN-AD Len: 25 RD: 192.0.2.3:1 ESI: 01:00:00:00:00:12:00:00:00:01,
                      tag: MAX-ET Label: 0

    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:64500:1
        esi-label:262142/All-Active
"


3 2017/05/05 13:52:46.80 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 80
    Flag: 0x90 Type: 14 Len: 36 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.3
        Type: EVPN-AD Len: 25 RD: 192.0.2.3:1 ESI: 01:00:00:00:00:12:00:00:00:01
                      tag: 0 Label: 4194256

    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:64500:1
        bgp-tunnel-encap:MPLS
"


*A:PE-2# show router bgp routes evpn auto-disc esi 01:00:00:00:00:12:00:00:00:01
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN Auto-Disc Routes
===============================================================================
Flag  Route Dist.         ESI                           NextHop
      Tag                                               Label
-------------------------------------------------------------------------------
u*>i  192.0.2.3:1         01:00:00:00:00:12:00:00:00:01 192.0.2.3
      0                                                 LABEL 262141

u*>i  192.0.2.3:1         01:00:00:00:00:12:00:00:00:01 192.0.2.3
      MAX-ET                                            LABEL 0

-------------------------------------------------------------------------------
Routes : 2
```

```
===============================================================================

*A:PE-2# show router bgp routes evpn auto-disc esi 01:00:00:00:00:12:00:00:00:01
hunt
---snip---
===============================================================================
BGP EVPN Auto-Disc Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network         : N/A
Nexthop         : 192.0.2.3
From            : 192.0.2.3
Res. Nexthop    : 192.168.23.2
---snip---
Community       : target:64500:1 bgp-tunnel-encap:MPLS
---snip---
EVPN type       : AUTO-DISC
ESI             : 01:00:00:00:00:12:00:00:00:01
Tag             : 0
Route Dist.     : 192.0.2.3:1
MPLS Label      : LABEL 262141

---snip---

Network         : N/A
Nexthop         : 192.0.2.3
From            : 192.0.2.3
Res. Nexthop    : 192.168.23.2
---snip---
Community       : target:64500:1 esi-label:262142/All-Active
---snip---
EVPN type       : AUTO-DISC
ESI             : 01:00:00:00:00:12:00:00:00:01
Tag             : MAX-ET
Route Dist.     : 192.0.2.3:1
MPLS Label      : LABEL 0
---snip---
```

From a service perspective, as soon as CE-11 sends some traffic, the PE learning the CE-11 MAC address will advertise it to the network. The remote PEs (PE-4 and PE-5) will create a new EVPN-MPLS ES destination to ESI-12, with two next-hops: PE-2 and PE-3. The following outputs show the following information:

- PE-4 has learned AD per-EVI/ES routes for ESI-12 from PE-2 and PE-3, as well as the CE-11 MAC address from PE-3 (because MTU-1 picked up its link to PE-3 to send CE-11 frames).

```
*A:PE-4# show router bgp routes evpn auto-disc esi 01:00:00:00:00:12:00:00:00:01
===============================================================================
 BGP Router ID:192.0.2.4        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
```

```
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN Auto-Disc Routes
===============================================================================
Flag  Route Dist.         ESI                           NextHop
      Tag                                               Label
-------------------------------------------------------------------------------
u*>i  192.0.2.2:1         01:00:00:00:00:12:00:00:00:01 192.0.2.2
      0                                                 LABEL 262141

u*>i  192.0.2.2:1         01:00:00:00:00:12:00:00:00:01 192.0.2.2
      MAX-ET                                            LABEL 0

u*>i  192.0.2.3:1         01:00:00:00:00:12:00:00:00:01 192.0.2.3
      0                                                 LABEL 262141

u*>i  192.0.2.3:1         01:00:00:00:00:12:00:00:00:01 192.0.2.3
      MAX-ET                                            LABEL 0
-------------------------------------------------------------------------------
Routes : 4
===============================================================================
```

PE-4 has learned MAC address 00:00:11:11:11:11 of CE-11 in ESI-12. The BGP EVPN MAC route has PE-3 as next hop:

```
*A:PE-4# show router bgp routes evpn mac
===============================================================================
 BGP Router ID:192.0.2.4        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN MAC Routes
===============================================================================
Flag  Route Dist.         MacAddr           ESI
      Tag                 Mac Mobility      Label1
                          Ip Address
                          NextHop
-------------------------------------------------------------------------------
u*>i  192.0.2.3:1         00:00:11:11:11:11 01:00:00:00:00:12:00:00:00:01
      0                   Seq:0             LABEL 262141
                          N/A
                          192.0.2.3
---snip---
```

- In the FDB for VPLS-1, PE-4 has learned the CE-11 MAC address associated with a newly created EVPN-MPLS ES destination:

```
*A:PE-4# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC                Source-Identifier        Type      Last Change
                                                      Age
-------------------------------------------------------------------------------
1         00:00:11:11:11:11 eES:                      Evpn      05/05/17 08:57:00
                            01:00:00:00:00:12:00:00:00:01
---snip---
```

- Due to the aliasing function, the newly created EVPN-MPLS ES destination to ESI-12 has two next-hops (PE-2 and PE-3), to which PE-4 can load-balance the unicast traffic because **ecmp 2** is configured in the VPLS-1 of PE-4.

```
*A:PE-4# show service id 1 evpn-mpls

===============================================================================
BGP EVPN-MPLS Dest
===============================================================================
TEP Address     Egr Label     Num. MACs   Mcast           Last Change
                 Transport
-------------------------------------------------------------------------------
192.0.2.2       262140        0           Yes             05/05/2017 08:30:01
                ldp
192.0.2.3       262140        0           Yes             05/05/2017 08:30:29
                ldp
192.0.2.5       262140        0           Yes             05/05/2017 08:10:40
                ldp
-------------------------------------------------------------------------------
Number of entries : 3
-------------------------------------------------------------------------------
===============================================================================

===============================================================================
BGP EVPN-MPLS Ethernet Segment Dest
===============================================================================
Eth SegId                       Num. Macs               Last Change
-------------------------------------------------------------------------------
01:00:00:00:00:12:00:00:00:01   1                       05/05/2017 10:57:00
---snip---
```

The **show service id 1 evpn-mpls esi 01:00:00:00:00:12:00:00:00:01** command shows the next-hops that the EVPN-MPLS ES destination is resolved to.

```
*A:PE-4# show service id 1 evpn-mpls esi 01:00:00:00:00:12:00:00:00:01

===============================================================================
BGP EVPN-MPLS Ethernet Segment Dest
===============================================================================
Eth SegId                       Num. Macs               Last Change
-------------------------------------------------------------------------------
01:00:00:00:00:12:00:00:00:01   1                       05/05/2017 10:57:00
===============================================================================

===============================================================================
BGP EVPN-MPLS Dest TEP Info
===============================================================================
TEP Address              Egr Label                Last Change
                         Transport
-------------------------------------------------------------------------------
```

```
192.0.2.2                262141                   05/05/2017 11:47:00
                         ldp
192.0.2.3                262141                   05/05/2017 11:47:00
                         ldp
-------------------------------------------------------------------------------
Number of entries : 2
-------------------------------------------------------------------------------
===============================================================================
```

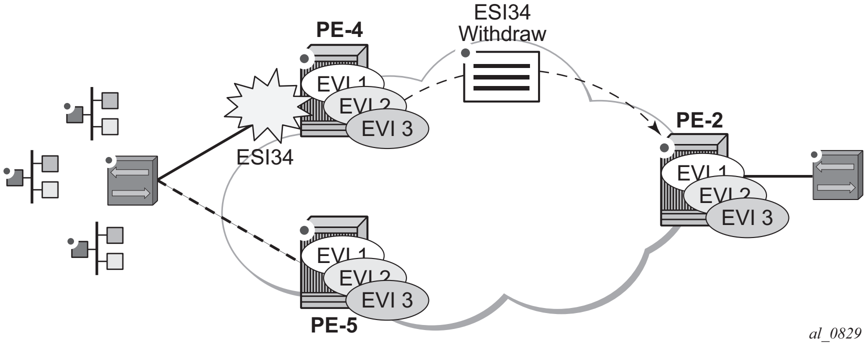
- PE-3 will show the CE-11 MAC address as learned locally in SAP lag-1:1 (because the data plane learning of the CE-11 MAC address happened in PE-3). For PE-2, even though it learned the MAC address from EVPN, it will install it as associated with SAP lag-1:1 because the EVPN route came with ESI-12, which is a local ESI. Because of this, whenever PE-2 receives a frame with MAC DA equal to the CE-11 MAC address, it will be able to forward the frame locally to the SAP lag-1:1. The following output shows the CE-11 MAC address as it is installed in PE-2 and PE-3:

```
*A:PE-2# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         00:00:11:11:11:11 sap:lag-1:1              Evpn     05/05/17 08:27:59
---snip---

*A:PE-3# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         00:00:11:11:11:11 sap:lag-1:1              L/30     05/05/17 08:57:59
---snip---
```

Single-Active Multi-Homing Concepts

Figure 43 illustrates two concepts in EVPN single-active multi-homing: mass-withdraw and backup path.

Figure 43 **EVPN-MPLS Single-Active Multi-Homing: Mass-Withdraw, Backup Path**



- With mass-withdraw, when ESI-34 goes down, PE-2 does not have to wait for all the MAC routes to be withdrawn to converge all the services. Instead, PE-4 will withdraw the AD per-ES routes (also the AD per-EVI and MAC routes) and that will be used at PE-2 as a notification to stop sending traffic to PE-4 for any MAC address associated with ESI-34.
- With backup path, when PE-2 is notified of the ESI-34 failure due to the withdrawn AD routes, it will not flush any MAC address associated with ESI-34. Instead, it will change the next-hop of the EVPN-MPLS ES destination to the remaining PE in the ESI-34. Backup path only works when there are two PEs in the same ES. If there were more than two PEs in ESI-34, PE-2 would flush all the MAC addresses upon receiving a mass-withdraw notification, because it would not know who the new active PE is.

Single-Active Multi-Homing Configuration

The single-active multi-homing configuration example is based on Figure 41:

MTU-6 is connected to the EVPN network using single-active multi-homing. With the MTU-6 configuration, a VPLS service with active-standby spoke-sdp to PE-4 and PE-5 is configured. In PE-4 and PE-5, the SDP connected to MTU-6 is linked to ESI-34. Both will run the DF election algorithm for EVI 1, and the non-DF PE (PE-4 in this example) will bring down the spoke-SDP and notify MTU-6.

The following output shows the configuration of ESI-34 in PE-4 and PE-5, as well as the SDPs. The configuration of MTU-6 is also shown for completeness. It is important to keep the default **no ignore-standby-signaling** command on MTU-6 spoke-SDPs because the PW switchover in MTU-6 will be triggered based on the PW status bits sent by PE-4 and PE-5.

SDP 46 with far-end MTU-6 is configured on PE-4:

```
configure
```

```
    service
        sdp 46 mpls create
            far-end 192.0.2.6
            ldp
            no shutdown
        exit
```

Ethernet segment “ESI-34” is configured on PE-4 as follows:

```
configure
    service
        system
            bgp-evpn
                ethernet-segment "ESI-34" create
                    esi 01:00:00:00:00:34:00:00:00:01
                    es-activation-timer 3
                    service-carving
                        mode auto
                    exit
                    multi-homing single-active
                    sdp 46
                    no shutdown
                exit
```

On PE-5, SDP 56 is configured as follows:

```
configure
    service
        sdp 56 mpls create
            far-end 192.0.2.6
            ldp
            no shutdown
        exit
```

Ethernet segment “ESI-34” is configured as follows on PE-5:

```
configure
    service
        system
            bgp-evpn
                ethernet-segment "ESI-34" create
                    esi 01:00:00:00:00:34:00:00:00:01
                    es-activation-timer 3
                    service-carving
                        mode auto
                    exit
                    multi-homing single-active
                    sdp 56
                    no shutdown
                exit
```

On MTU-6, the service configuration is as follows:

```
configure
    service
        sdp 64 mpls create
```

```
                far-end 192.0.2.4
                ldp
                no shutdown
            exit
            sdp 65 mpls create
                far-end 192.0.2.5
                ldp
                no shutdown
            exit
            vpls 1 customer 1 create
                endpoint "CORE" create
                exit
                sap 1/2/1:1 create
                exit
                spoke-sdp 64:1 endpoint "CORE" create
                exit
                spoke-sdp 65:1 endpoint "CORE" create
                exit
                no shutdown
```

For a detailed description of the base BGP-EVPN instance and **ethernet-segment** configuration, see the All-Active Multi-Homing Configuration section. The **es-activation-timer**, **esi**, **service-carving**, **boot-timer**, and **shutdown** commands are used in the same way as for all-active multi-homing. Only the differences compared to all-active multi-homing are described here:

- **multi-homing single-active** must be configured so that the ES acts as single-active. Optionally, the **no-esi-label** attribute can be added to the **multi-homing single-active** command. This attribute controls the use of the ESI-label for single-active multi-homing. Although the ESI-label is always used in all-active multi-homing when sending BUM traffic between the PEs in the ES, it is configurable for single-active. However, Nokia recommends to use the default option (using ESI-label) to avoid potential transient issues when there is a DF switchover.
- **sdp** <*sdp-id*> is configured so that the ES can be associated with the SDP connected to MTU-6. Although all-active multi-homing only allows LAG associations to the ES, single-active allows LAG, port, and SDP. In this example, SDP is the option, because the access network is MPLS-based.

Similar to the all-active multi-homing case, when configuring the service in PE-4 and PE-5, the service objects are automatically associated with the ESI-34, because they are defined in the SDPs linked to the ESI. The configuration for VPLS 1 on PE-5 is as follows:

```
configure
    service
        vpls 1 customer 1 create
            bgp
            exit
            bgp-evpn
                evi 1
                mpls
```

```
                    ingress-replication-bum-label
                    ecmp 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            spoke-sdp 56:1 create
                no shutdown
            exit
            no shutdown
```

In all-active multi-homing, the non-DF does not bring down the service SAP associated with the ES (it only removes it from the default-multicast-list). However, in single-active multi-homing, the service spoke-SDP (or SAP, if that was the object associated) is brought operationally down. The following output shows the spoke-SDP state in PE-4 (non-DF), as operationally down with the **StandbyForMHProtocol** flag and the **Local Pw Bits** that are signaled to MTU-6:

```
*A:PE-4# show service system bgp-evpn ethernet-segment name "ESI-34" evi 1

===============================================================================
EVI DF and Candidate List
===============================================================================
EVI           SvcId         Actv Timer Rem      DF  DF Last Change
-------------------------------------------------------------------------------
1             1             0                   no  05/05/2017 11:16:14
===============================================================================
===============================================================================
DF Candidates                           Time Added
-------------------------------------------------------------------------------
192.0.2.4                               05/05/2017 11:16:40
192.0.2.5                               05/05/2017 11:16:40
-------------------------------------------------------------------------------
Number of entries: 2
===============================================================================
```

Spoke-SDP 46:1 is operationally down on PE-4:

```
*A:PE-4# show service id 1 base

---snip---
-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sdp:46:1 S(192.0.2.6)                    Spok         0       8974    Up   Down
===============================================================================
```

Spoke-SDP 46:1 is operationally down with the StandbyForMHProtocol flag:

```
*A:PE-4# show service id 1 sdp 46:1 detail | match Flag
Flags              : StandbyForMHProtocol
```

The local PW bits (**pwFwdingStandby**) are sent to MTU-6:

```
*A:PE-4# show service id 1 sdp 46:1 detail | match Pw
Local Pw Bits      : pwFwdingStandby
Peer Pw Bits       : None
```

Single-Active Multi-Homing Operation

The same commands used in the All-Active Multi-Homing Operation section can be used for single-active; see that section.

The **show service system bgp-evpn ethernet-segment name ESI-34** command shows an Ethernet-segment **Oper Multi-homing** in addition to the configured **Multi-homing** mode. This occurs because, in spite of configuring the ES as all-active, it may operate as single-active if there is a mismatch between the modes advertised by PE-4 and PE-5 in the AD per-ES routes (per RFC 7432). In this example, the configured and the operational value are the same:

```
*A:PE-4# show service system bgp-evpn ethernet-segment name "ESI-34"

================================================================================
Service Ethernet Segment
================================================================================
Name                    : ESI-34
Eth Seg Type            : None
Admin State             : Enabled            Oper State         : Up
ESI                     : 01:00:00:00:00:34:00:00:00:01
Multi-homing            : singleActive       Oper Multi-homing  : singleActive
ES SHG Label            : 262142
Source BMAC LSB         : <none>
Sdp Id                  : 46
ES Activation Timer     : 3 secs
Svc Carving             : auto               Oper Svc Carving   : auto
Cfg Range Type          : primary
================================================================================
```

As soon as CE-16 sends some traffic, the DF PE (PE-5) will learn the CE-16 MAC address and will advertise it to the network. The remote PEs (PE-2 and PE-3) will create a new EVPN-MPLS ES destination to ESI-34, but this time with only one next-hop, PE-5, because this is single-active multi-homing. The following outputs show the following information:

- PE-2 has learned AD per-EVI/ES routes for ESI-34 from PE-4 and PE-5, as well as the CE-16 MAC address from an ES EVPN-MPLS destination, which is resolved to PE-5 (the DF for ESI-34).

```
*A:PE-2# show router bgp routes evpn auto-disc esi 01:00:00:00:00:34:00:00:00:01
================================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
```

```
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN Auto-Disc Routes
===============================================================================
Flag  Route Dist.         ESI                           NextHop
      Tag                                               Label
-------------------------------------------------------------------------------
u*>i  192.0.2.4:1         01:00:00:00:00:34:00:00:00:01 192.0.2.4
      0                                                 LABEL 262141

u*>i  192.0.2.4:1         01:00:00:00:00:34:00:00:00:01 192.0.2.4
      MAX-ET                                            LABEL 0

u*>i  192.0.2.5:1         01:00:00:00:00:34:00:00:00:01 192.0.2.5
      0                                                 LABEL 262141

u*>i  192.0.2.5:1         01:00:00:00:00:34:00:00:00:01 192.0.2.5
      MAX-ET                                            LABEL 0

-------------------------------------------------------------------------------
Routes : 4
===============================================================================
```

PE-2 has learned the CE-16 MAC address from an ES EVPN-MPLS destination:

```
*A:PE-2# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         00:00:11:11:11:11 sap:lag-1:1              Evpn     05/05/17 12:22:41
1         00:00:16:16:16:16 eES:                     Evpn     05/05/17 12:22:41
                            01:00:00:00:00:34:00:00:00:01
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
```

On PE-2, the ES EVPN-MPLS destination is resolved to DF PE-5:

```
*A:PE-2# show service id 1 evpn-mpls esi 01:00:00:00:00:34:00:00:00:01

===============================================================================
BGP EVPN-MPLS Ethernet Segment Dest
===============================================================================
Eth SegId                       Num. Macs               Last Change
-------------------------------------------------------------------------------
01:00:00:00:00:34:00:00:00:01   1                       05/05/2017 12:22:41
===============================================================================
===============================================================================
BGP EVPN-MPLS Dest TEP Info
===============================================================================
TEP Address              Egr Label                Last Change
                         Transport
-------------------------------------------------------------------------------
192.0.2.5                262141                   05/05/2017 12:22:41
                         ldp
-------------------------------------------------------------------------------
Number of entries : 1
-------------------------------------------------------------------------------
===============================================================================
```

- In this case, the local PEs, PE-4 and PE-5, will learn the CE MAC address from an EVPN-MPLS destination and a local spoke-SDP, respectively.

```
*A:PE-4# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         00:00:16:16:16:16 eES:                     Evpn     05/05/17 12:30:14
                            01:00:00:00:00:34:00:00:00:01
---snip---
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
```

The ES EVPN-MPLS destination is resolved to DF PE-5:

```
*A:PE-4# show service id 1 evpn-mpls esi 01:00:00:00:00:34:00:00:00:01

===============================================================================
BGP EVPN-MPLS Ethernet Segment Dest
===============================================================================
Eth SegId                       Num. Macs               Last Change
-------------------------------------------------------------------------------
01:00:00:00:00:34:00:00:00:01   1                       05/05/2017 12:30:14
===============================================================================

===============================================================================
BGP EVPN-MPLS Dest TEP Info
===============================================================================
TEP Address              Egr Label                Last Change
                         Transport
-------------------------------------------------------------------------------
192.0.2.5                262141                   05/05/2017 12:30:14
                         ldp
-------------------------------------------------------------------------------
Number of entries : 1
-------------------------------------------------------------------------------
===============================================================================
```

DF PE-5 learns the CE-16 MAC address from a local spoke SDP:

```
*A:PE-5# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         00:00:16:16:16:16 sdp:56:1                 L/60     05/05/17 12:37:26
---snip---
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
```

Ethernet-Segment Failures

If either ES fails, a DF re-election will happen and the corresponding AD per-ES/EVI routes will be withdrawn, causing the remote PEs to modify the list of next-hops for the EVPN-MPLS ES destination. The following example illustrates a failure on the SDP between MTU-6 and PE-5 (the DF).

Step 1. A failure occurs in the LSP between MTU-6 and PE-5. This can be any event that brings the SDP down.

```
*A:PE-5#
7 2017/05/05 13:24:54.60 UTC MINOR: SVCMGR #2303 Base
"Status of SDP 56 changed to admin=up oper=down"
```

Step 2. Immediately, PE-5 gives up the DF role and withdraws the ES route, as well as the AD routes and MAC routes. As soon as PE-4 receives any ES or AD withdraw, it will re-run the DF algorithm and, when the es-activation-timer expires, it will become the DF and activate its spoke-SDP.

```
*A:PE-5#
9 2017/05/05 13:24:54.60 UTC MINOR: SVCMGR #2094 Base
"Ethernet Segment:ESI-34, EVI:1, Designated Forwarding state changed to:false"
```

The ES in PE-5 is operational down:

```
*A:PE-5# show service system bgp-evpn ethernet-segment name "ESI-34"

===============================================================================
Service Ethernet Segment
===============================================================================
Name                    : ESI-34
Eth Seg Type            : None
Admin State             : Enabled            Oper State         : Down
ESI                     : 01:00:00:00:00:34:00:00:00:01
Multi-homing            : singleActive       Oper Multi-homing  : singleActive
ES SHG Label            : 262142
Source BMAC LSB         : <none>
Sdp Id                  : 56
ES Activation Timer     : 3 secs
```

```
Svc Carving             : auto               Oper Svc Carving   : auto
Cfg Range Type          : primary
===============================================================================
```

PE-5 is no longer the DF and the only DF candidate is PE-4:

```
*A:PE-5# show service system bgp-evpn ethernet-segment name "ESI-34" evi 1

===============================================================================
EVI DF and Candidate List
===============================================================================
EVI           SvcId         Actv Timer Rem      DF  DF Last Change
-------------------------------------------------------------------------------
1             1             0                   no  05/05/2017 13:24:55
===============================================================================

===============================================================================
DF Candidates                           Time Added
-------------------------------------------------------------------------------
192.0.2.4                               05/05/2017 08:16:40
-------------------------------------------------------------------------------
Number of entries: 1
===============================================================================
```

PE-4 becomes the DF and the spoke-SDP 46:1 is brought up.

```
*A:PE-4#
7 2017/05/05 13:24:57.57 UTC MINOR: SVCMGR #2094 Base
"Ethernet Segment:ESI-34, EVI:1, Designated Forwarding state changed to:true"

8 2017/05/05 13:24:57.57 UTC MINOR: SVCMGR #2326 Base
"Status of SDP Bind 46:1 in service 1 (customer 1) local PW status bits changed
to none"

9 2017/05/05 13:24:57.57 UTC MINOR: SVCMGR #2306 Base
"Status of SDP Bind 46:1 in service 1 (customer 1) changed to admin=up oper=up flags="
```

The ES is up in PE-4:

```
*A:PE-4# show service system bgp-evpn ethernet-segment name "ESI-34"

===============================================================================
Service Ethernet Segment
===============================================================================
Name                    : ESI-34
Eth Seg Type            : None
Admin State             : Enabled            Oper State         : Up
ESI                     : 01:00:00:00:00:34:00:00:00:01
Multi-homing            : singleActive       Oper Multi-homing  : singleActive
ES SHG Label            : 262142
Source BMAC LSB         : <none>
Sdp Id                  : 46
ES Activation Timer     : 3 secs
Svc Carving             : auto               Oper Svc Carving   : auto
Cfg Range Type          : primary
===============================================================================
```

PE-4 is the DF and there are no other DF candidates:

```
*A:PE-4# show service system bgp-evpn ethernet-segment name "ESI-34" evi 1

===============================================================================
EVI DF and Candidate List
===============================================================================
EVI           SvcId         Actv Timer Rem      DF  DF Last Change
-------------------------------------------------------------------------------
1             1             0                   yes 05/05/2017 13:24:58
===============================================================================

===============================================================================
DF Candidates                           Time Added
-------------------------------------------------------------------------------
192.0.2.4                               05/05/2017 08:16:40
-------------------------------------------------------------------------------
Number of entries: 1
===============================================================================
```

Step 3. The remote PEs, PE-2 and PE-3, receive the AD routes withdrawal and modify the next-hop for the EVPN-MPLS ES destination.

```
52 2017/05/05 13:17:21.60 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.5
"Peer 1: 192.0.2.5: UPDATE
Peer 1: 192.0.2.5 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 86
    Flag: 0x90 Type: 15 Len: 82 Multiprotocol Unreachable NLRI:
        Address Family EVPN
        Type: EVPN-AD Len: 25 RD: 192.0.2.5:1 ESI: 01:00:00:00:00:34:00:00:00:01,
                      tag: 0 Label: 0

        Type: EVPN-AD Len: 25 RD: 192.0.2.5:1 ESI: 01:00:00:00:00:34:00:00:00:01,
                      tag: MAX-ET Label: 0
---snip---
```

The ES EVPN-MPLS destination is resolved to the DF PE-4:

```
*A:PE-2# show service id 1 evpn-mpls esi 01:00:00:00:00:34:00:00:00:01

===============================================================================
BGP EVPN-MPLS Ethernet Segment Dest
===============================================================================
Eth SegId                       Num. Macs               Last Change
-------------------------------------------------------------------------------
01:00:00:00:00:34:00:00:00:01   1                       05/05/2017 13:17:22
===============================================================================
===============================================================================
BGP EVPN-MPLS Dest TEP Info
===============================================================================
TEP Address              Egr Label                Last Change
                         Transport
-------------------------------------------------------------------------------
192.0.2.4                262141                   05/05/2017 13:17:22
                         ldp
-------------------------------------------------------------------------------
Number of entries : 1
-------------------------------------------------------------------------------
===============================================================================
```

The following must be considered:

- The DF election procedure is revertive, that is, when the failed SDP comes back up, PE-5 will take over again as DF and the network will re-converge.
- The DF election is triggered by the following events:
  - **configure service system bgp-evpn ethernet-segment ESI-34 no shutdown** triggers the DF election for all the services in the ES.
  - A new update/withdrawal of an ES route (containing an ESI configured locally) triggers the DF election for all the services in the ESI.
  - A new update/withdrawal of an AD per-ES route (containing an ESI configured locally) triggers the DF election for all the services associated with the list of RTs received along with the route.
  - A new update of an AD per-ES route with a change in the ESI-label extended community (single-active bit or MPLS label) triggers the DF election for all the services associated with the list of RTs received along with the route.
  - A new update/withdrawal of an AD route per-EVI (containing an ESI configured locally) triggers the DF election for that service.

BGP-EVPN Route Selection in EVPN Networks

The selection of the best route for a MAC address is as follows:

- If a PE receives more than one route for the same MAC address, the best MAC route is chosen:
  - If the route key is equal in two or more routes (that is, the mac, mac-length, ip, ip-length, RD, eth-tag), then regular BGP selection applies:
    - If local-pref, AS-path, origin, and MED are equal, the lowest IGP distance to the BGP next-hop is chosen (unless **ignore-nh-metric** is configured). If the BGP next-hop is resolved by an LSP, the cost from the tunnel-table is used.
    - As a last resort tie-breaker, the route with the lowest originator ID, or received from the peer with the lowest BGP Identifier, is chosen (unless **ignore-router-id** is configured and the routes being compared are EBGP routes).
  - If the mac-length, mac, ip-length, ip, eth-tag are equal, and the RD is different, the EVPN selection process is applied in the following order:
    - Conditional static macs (local protected macs)
    - EVPN static macs (remote protected macs)
    - Data plane learned MACs (regular learning on SAPs/SDP-bindings)

  - EVPN macs with higher SEQ number
  - Lowest IP (next-hop IP of the EVPN NLRI)
  - Lowest eth-tag (will be normally zero)
  - Lowest RD
- After a MAC route is selected, the system checks for an associated ES.
  - If it has an ES, the system uses the MAC address as the EVPN-MPLS ES destination. The ES destination is constructed based on the AD per-EVI routes received for that ES (regardless of MAC address priorities with the ES).
  - The system selects the first ECMP number of AD per-EVI routes arranged by the IP address of PEs (lower IPs are selected first).
  - If the same PE has advertised multiple RDs, the system selects the route with the lowest RD for that PE.

In the example of Figure 41, PE-4 resolves the next-hops for ESI-12 as described in the second choice above, that is, because ECMP=2, the two available next-hops are chosen. If ECMP is changed to 1, PE-4 will pick up the lower IP (in the BGP next-hop). This is illustrated in the following output:

```
*A:PE-4# show service id 1 evpn-mpls esi 01:00:00:00:00:12:00:00:00:01

===============================================================================
BGP EVPN-MPLS Ethernet Segment Dest
===============================================================================
Eth SegId                       Num. Macs               Last Change
-------------------------------------------------------------------------------
01:00:00:00:00:12:00:00:00:01   1                       05/05/2017 13:15:32
===============================================================================
===============================================================================
BGP EVPN-MPLS Dest TEP Info
===============================================================================
TEP Address              Egr Label                Last Change
                         Transport
-------------------------------------------------------------------------------
192.0.2.2                262141                   05/05/2017 13:13:16
                         ldp
192.0.2.3                262141                   05/05/2017 13:15:32
                         ldp
-------------------------------------------------------------------------------
Number of entries : 2
-------------------------------------------------------------------------------
===============================================================================
```

When ECMP equals 1, only the BGP next hop with the lower IP is chosen:

```
*A:PE-4# configure service vpls 1 bgp-evpn mpls ecmp 1
```

```
*A:PE-4# show service id 1 evpn-mpls esi 01:00:00:00:00:12:00:00:00:01
===============================================================================
BGP EVPN-MPLS Ethernet Segment Dest
===============================================================================
Eth SegId                       Num. Macs               Last Change
-------------------------------------------------------------------------------
01:00:00:00:00:12:00:00:00:01   1                       05/05/2017 14:00:24
===============================================================================
===============================================================================
BGP EVPN-MPLS Dest TEP Info
===============================================================================
TEP Address              Egr Label                Last Change
                         Transport
-------------------------------------------------------------------------------
192.0.2.2                262141                   05/05/2017 13:13:16
                         ldp
-------------------------------------------------------------------------------
Number of entries : 1
-------------------------------------------------------------------------------
===============================================================================
```

Comparing EVPN Multi-homing and BGP Multi-homing

EVPN-MPLS services support EVPN-MH (EVPN multi-homing) and also BGP-MH as in chapter BGP Multi-Homing for VPLS Networks. While EVPN-MH is the standard way of providing access resiliency in RFC 7432, BGP-MH is also a standard mechanism supported in VPLS or EVPN networks. The following table provides some comparison between both technologies.

Table 4 **Comparing EVPN Multi-homing and BGP Multi-homing**

VPN Requirements	EVPN-MH	BGP-MH	Comments
All-active MH (flow-based load-balancing)	Yes	No	EVPN-MH provides better bandwidth utilization
Single-active MH (service-based load-balancing)	Yes	Yes	
DF PE election - automatic service balancing	Yes Service-carving	No Requires vsi policies and LP manipulation	EVPN-MH provides better automation
DF PE election – manual configuration per service	Yes	No	EVPN-MH allows for manual DF config for EVIs and ISIDs (2 PEs)
Split-horizon indication in the data plane	Yes ESI-label	No	Prevents transient loops when dual-active DFs show up
DF indication in the control plane	No	Yes	BGP MH guarantees one DF at a time. EVPN relies on Timers to ensure one DF at a time

Table 4 **Comparing EVPN Multi-homing and BGP Multi-homing (Continued)**

VPN Requirements	EVPN-MH	BGP-MH	Comments
Allows multiple SAPs or SDP-bindings per service on the same site	No	Yes Through the use of SHGs	
Boot timer and site(es)-activation-timers	Yes	Yes	BGP-MH supports more granular configuration (service level)
Support for oper-groups	No	Yes	
Non-DF notification to the CE (MPLS and CFM)	Yes	Yes	Avoids blackholing

In addition to the preceding comparison, the following configuration excerpt compares EVPN-MH with BGP-MH on a bgp-evpn VPLS service and shows that, while EVPN-MH does not have any configuration at service level, BGP-MH is configured within the VPLS context, which gives a more granular control over the redundancy provided. See the BGP Multi-Homing for VPLS Networks chapter for more information about BGP-MH.

```
config>service>system>bgp-evpn# info
----------------------------------------
  ethernet-segment "ESI-34" create
    esi 01:00:00:00:00:34:00:00:00:01
    es-activation-timer 3
    service-carving
      mode auto
      exit
    multi-homing single-active
    sdp 46
    no shutdown
    exit
```

```
config>service>vpls# info
----------------------------------------
  bgp
  exit
  bgp-evpn
    evi 1
    vxlan
      shutdown
    exit
    mpls
      ingress-replication-bum-label
      ecmp 2
      auto-bind-tunnel
        resolution any
      exit
      no shutdown
    exit
  exit
  spoke-sdp 46:1 create
    no shutdown
  exit
  no shutdown
--------------------------------------
```

For BGP multi-homing, site "site-1" is configured, as follows. The RD needs to be configured in the bgp context.

```
config>service>vpls# info
---------------------------------------
  bgp
    route-distinguisher 192.0.2.4:1
  exit
  bgp-evpn
    evi 1
    vxlan
      shutdown
    exit
    mpls
      ingress-replication-bum-label
      ecmp 2
      auto-bind-tunnel
        resolution any
      exit
      no shutdown
    exit
  exit
  site "site-1" create
    site-id 1
    spoke-sdp 46:1
    site-activation-timer 3
    no shutdown
  exit
  spoke-sdp 46:1 create
    no shutdown
  exit
  no shutdown
--------------------------------------
```

Proxy-ARP/ND Configuration for EVPN-MPLS Networks

Although not strictly a BGP-EVPN configuration, **vpls>proxy-arp** and **vpls>proxy-nd** functions are typically enabled along with EVPN-MPLS in order to reduce the amount of flooding in the network. The proxy-ARP/ND agent in the VPLS service will snoop ARP-requests and/or Neighbor Solicitation messages and will reply to those messages locally (if the information is known) without having to flood the requests to the network.

The configuration options for proxy-ARP are the following:

```
*A:PE-2# configure service vpls 1 proxy-arp ?
```

```
  - no proxy-arp
  - proxy-arp

 [no] age-time        - Configure aging timer for proxy ARP entries
      dup-detect      - Configure anti-spoofing MAC address information
 [no] dynamic         + Configure dynamic entry information
 [no] dynamic-arp-po* - Configure population of dynamic proxy ARP entries
 [no] garp-flood-evpn - Configure to flood GARP request/replys into EVPN
 [no] send-refresh    - Configure send refresh time
 [no] shutdown        - Administratively enable/disable proxy ARP configuration
 [no] static          - Configure static IP address to MAC address associations
      table-size      - Configure the maximum number of entries in the proxy ARP table
 [no] unknown-arp-re* - Configure to flood unknown ARP request
```

The configuration options for proxy-ND are the following:

```
*A:PE-2# configure service vpls 1 proxy-nd  ?
  - no proxy-nd
  - proxy-nd

 [no] age-time        - Configure aging timer for proxy ND entries
      dup-detect      - Configure anti-spoofing MAC address information
 [no] dynamic         + Configure dynamic entry information
 [no] dynamic-nd-pop* - Configure population of dynamic proxy ND entries
      evpn-nd-advert* - Configure EVPN Neighbor Discovery advertisements
 [no] host-unsolicit* - Configure whether to flood evpn with host neighbor
                        advertisement
 [no] router-unsolic* - Configure whether to flood evpn with router neighbor
                        advertisement
 [no] send-refresh    - Configure send refresh time
 [no] shutdown        - Administratively enable/disable proxy ND configuration
 [no] static          - Configure static IP address to MAC address associations
      table-size      - Configure the maximum number of entries in the proxy ND table
 [no] unknown-ns-flo* - Configure to flood unknown ND solicitation
```

When proxy-ARP/ND is enabled, the following configuration guidelines must be followed:

- **dynamic-arp-populate** or **dynamic-nd-populate** should be used only in networks with a consistent configuration of this command in all PEs.
- When using **dynamic-arp-populate**/**dynamic-nd-populate**, the **age-time** value should be configured to a value equal to three times the **send-refresh** value. This will help reduce the EVPN withdrawals and re-advertisements in the network.
- With large **age-time** values, it would be sufficient to configure the **send-refresh** value to half of the **proxy-ARP**/**ND age-time** or **FDB age-time**.
- In scaled environments (with thousands of services), it is not recommended to set the send-refresh value to less than 300 s. In such scenarios, Nokia recommends using a minimum proxy-ARP/ND **age-time** and FDB **age** of 900 s.

- The use of the following commands reduces or suppresses the ARP/ND flooding in an EVPN network, because EVPN MAC routes replace the function of the regular data plane ARP/ND messages:
  - **no garp-flood-evpn**
  - **no unknown-arp-request-flood-evpn**
  - **no unknown-ns-flood-evpn**
  - **no host-unsolicited-na-flood-evpn**
  - **no router-unsolicited-na-flood-evpn**
- Nokia recommends using the preceding commands only in EVPN networks where the CEs are routers directly connected to an SR OS node acting as the PE. Networks using aggregation switches between the host/routers and the PEs should flood GARP/ND messages in EVPN to make sure the remote caches are updated and BGP does not miss the advertisement of these entries.
- When the **anti-spoof-mac** is used with proxy-ARP/ND, ingress filters (in the access SAPs/SDP-bindings) should be configured to drop all traffic with destination anti-spoof-mac. The same MAC address should be configured in all PEs where dup-detect is active.
- When proxy-ND is used, the configuration of the following commands should be consistent in all the PEs in the network:
  - **router-unsolicited-na-flood-evpn**
  - **host-unsolicited-na-flood-evpn**
  - **evpn-nd-advertise**
- Because EVPN does not propagate the **router** flag in IPv6--> MAC address advertisements, in a mixed network with hosts and routers where **evpn-nd-advertise router** is configured, unsolicited host NA messages should be flooded so that the entire network gets to learn all of the host and router ND entries. In the same way, **evpn-nd-advertise host** should be configured so that unsolicited router NA messages are flooded.

Finally, along with proxy-ARP/ND, **vpls>discard-unknown** may be used in some EVPN-MPLS deployments where all the CEs are routers and they announce themselves to the network by sending GARPs or NAs (Neighbor Solicitation messages). According to RFC 7432, whether or not to flood packets to unknown destination MAC addresses should be an administrative choice, depending on how learning happens between CEs and PEs. **Discard-unknown** provides that administrative choice in case all the MAC addresses in an EVI can be learned even before any traffic is exchanged.

Proxy-ARP/ND along with **discard-unknown** helps reduce the BUM traffic in an EVPN network significantly; however, their use must be analyzed and considered, depending on the type of CEs in the EVI.

An example of proxy-ARP configuration is as follows. This configuration should be added to all PEs. When a new ARP message is received on any of the PEs, they will learn the IP-MAC address pair and will advertise it to the network.

```
configure
    service
        vpls 1
            proxy-arp
                age-time 900
                send-refresh 300
                dynamic-arp-populate
                no shutdown
            exit
```

Enabling proxy-ARP increases the number of MAC/IP routes being sent by the PEs. This is due to the following reasons:

- An additional MAC/IP route will be advertised per new learned IP-MAC address pair, regardless of having advertised the same MAC address already.
- A MAC per VPLS service will be advertised with a system MAC address. That MAC address will be used as MAC SA for proxy-ARP confirm messages when an IP moves to a different PE.

The following output shows the MAC/IP routes on PE-2 when proxy-ARP is enabled in the network.

```
*A:PE-2# show router bgp routes evpn mac
===============================================================================
 BGP Router ID:192.0.2.2       AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN MAC Routes
===============================================================================
Flag  Route Dist.         MacAddr           ESI
      Tag                 Mac Mobility      Label1
                          Ip Address
                          NextHop
-------------------------------------------------------------------------------
u*>i  192.0.2.3:1         16:0b:ff:00:03:3a ESI-0
      0                   Static            LABEL 262141
                          N/A
                          192.0.2.3

u*>i  192.0.2.4:1         16:0c:ff:00:03:3a ESI-0
      0                   Static            LABEL 262141
                          N/A
                          192.0.2.4

u*>i  192.0.2.5:1         00:00:16:16:16:16 01:00:00:00:00:34:00:00:00:01
```

```
      0                 Seq:0           LABEL 262141
                        N/A
                        192.0.2.5

u*>i  192.0.2.5:1       16:4f:ff:00:03:3a ESI-0
      0                 Static          LABEL 262141
                        N/A
                        192.0.2.5
---snip---
===============================================================================
```

Troubleshooting and Debug Commands

When troubleshooting an EVPN-MPLS network, the following show commands and debug commands are recommended, as already discussed throughout this chapter:

- **show redundancy bgp-evpn-multi-homing**
- **show router bgp routes evpn (and filters)**
- **show service evpn-mpls [<TEP ip-address>]**
- **show service id bgp-evpn**
- **show service id evpn-mpls (and modifiers)**
- **show service id fdb (and modifiers)**
- **show service system bgp-evpn**
- **show service system bgp-evpn ethernet-segment (and modifiers)**
- **debug router bgp update**
- **log-id 99**

In addition to the preceding commands, the following tools dump commands may also help:

- **tools dump service evpn usage** — This command shows the amount of EVPN-MPLS (and EVPN-VXLAN) destinations consumed in the system.
- **tools dump service system bgp-evpn ethernet-segment** <*name*> **evi <[1..65535]> df** — This command computes the DF election for a specific ESI and EVI. Note: The **show service system bgp-evpn ethernet-segment** commands shows whether the local PE is DF or non-DF for a specific EVI, but it does not show who the DF is if it is not the local PE. In case of more than 2 PEs in the ES, this command may be especially useful.

Some examples are provided below for PE-2. PE-2 is showing seven EVPN-MPLS destinations due to the following:

- Each remote PE consumes one EVPN-MPLS destination for unicast (if they advertise MAC/IP routes to PE-2 and the ingress-replication-bum-label is configured in all the PEs). PE-2 has three remote unicast EVPN-MPLS destinations.
- Each remote PE consumes one EVPN-MPLS destination for multicast (if they advertise inclusive multicast routes to PE-2). PE-2 has three remote multicast EVPN-MPLS destinations.
- Each remote ES consumes one EVPN-MPLS destination (it is only one per ES, regardless of the multi-homing mode and the number of PEs in the ES). PE-2 has one remote ES (ESI-34).

```
*A:PE-2# tools dump service evpn usage
EVPN usage statistics at 002 06:29:38.940:

MPLS-TEP                                           :           3
VXLAN-TEP                                          :           0
Total-TEP                                          :     3/ 16383

Mpls Dests (TEP, Egress Label + ES + ES-BMAC)      :           7
Mpls Etree Leaf Dests                              :           0
Vxlan Dests (TEP, Egress VNI)                      :           0
Total-Dest                                         :     7/196607

Sdp Bind +  Evpn Dests                             :     8/245759
ES L2/L3 PBR                                       :     0/ 32767
Evpn Etree Remote BUM Leaf Labels                  :           0
```

To compute the DF election for EVI 1:

```
*A:PE-2# tools dump service system bgp-evpn ethernet-segment "ESI-12" evi 1 df

[05/05/2017 11:40:28] Computed DF: 192.0.2.3 (Remote) (Boot Timer Expired: Yes)
```

Conclusion

SR OS has a full RFC 7432 EVPN-MPLS implementation including single-active and all-active multi-homing. This example has shown how to configure and operate EVPN-MPLS for a simple non multi-homing configuration as well as a multi-homing configuration. Other topics, such as the integration of VPLS objects with EVPN-MPLS and proxy-ARP/ND, have also been discussed.

EVPN for MPLS Tunnels in Epipe Services (EVPN-VPWS)

This chapter provides information about EVPN for MPLS Tunnels in Epipe Services (EVPN-VPWS).

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SR OS Release 14.0.R4, but the CLI in the current edition is based on SR OS release 15.0.R2. Ethernet Virtual Private Network - Virtual Private Wire Service (EVPN-VPWS) is supported in SR OS Release 14.0.R1, and later. EVPN-VPWS in multi-homing scenarios is supported in SR OS Release 14.0.R4, and later.

Chapter EVPN for MPLS Tunnels is prerequisite reading.

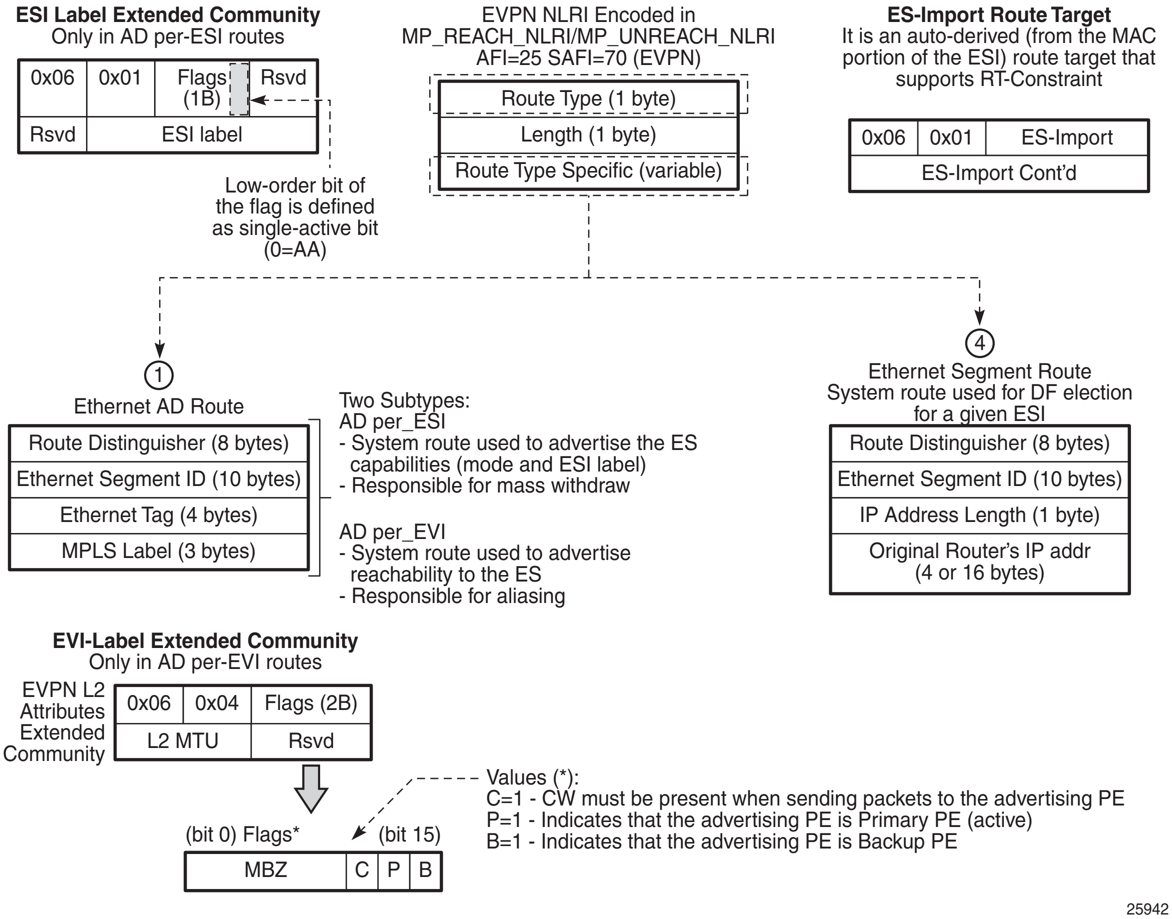
Overview

Service providers prefer an optimized, standardized, and unified control plane for VPNs. EVPN-VPWS is supported in MPLS networks that also run EVPN-MPLS in VPLS services. From a control plane perspective, EVPN-VPWS is a simplified point-to-point version of *RFC7432 - BGP MPLS-Based Ethernet VPN*, because there is no need to advertise MAC routes in VPWS. EVPN-VPWS is described in *draft-ietf-bess-evpn-vpws*.

EVPN-VPWS supports all-active multi-homing (per-flow load-balancing multi-homing) as well as single-active multi-homing (per-service load-balancing multi-homing), using the same Ethernet Segments (ESs) used for EVPN-MPLS VPLS services. EVPN-VPWS uses route-type 1 and route-type 4; it does not use route-types 2, 3, or 5, because MAC/IP routes, Inclusive Multicast, or IP-prefix routes are not required.

Figure 44 shows the encoding of the required extensions for the route-types 1 and 4 for EVPN-VPWS.

Figure 44 **Route Types and NLRIs for EVPN-VPWS**


ESI Label Extended Community
Only in AD per-ESI routes
0x06
0x01
Flags (1B)
Rsvd
Rsvd
ESI label
Low-order bit of the flag is defined as single-active bit (0=AA)
EVPN NLRI Encoded in MP_REACH_NLRI/MP_UNREACH_NLRI AFI=25 SAFI=70 (EVPN)
Route Type (1 byte)
Length (1 byte)
Route Type Specific (variable)
ES-Import Route Target
It is an auto-derived (from the MAC portion of the ESI) route target that supports RT-Constraint
0x06
0x01
ES-Import
ES-Import Cont'd
1
Ethernet AD Route
Route Distinguisher (8 bytes)
Ethernet Segment ID (10 bytes)
Ethernet Tag (4 bytes)
MPLS Label (3 bytes)
Two Subtypes:
AD per_ESI
- System route used to advertise the ES capabilities (mode and ESI label)
- Responsible for mass withdraw
AD per_EVI
- System route used to advertise reachability to the ES
- Responsible for aliasing
4
Ethernet Segment Route
System route used for DF election for a given ESI
Route Distinguisher (8 bytes)
Ethernet Segment ID (10 bytes)
IP Address Length (1 byte)
Original Router's IP addr (4 or 16 bytes)
EVI-Label Extended Community
Only in AD per-EVI routes
EVPN L2 Attributes Extended Community
0x06
0x04
Flags (2B)
L2 MTU
Rsvd
(bit 0) Flags*
(bit 15)
MBZ
C
P
B
Values (*):
C=1 - CW must be present when sending packets to the advertising PE
P=1 - Indicates that the advertising PE is Primary PE (active)
B=1 - Indicates that the advertising PE is Backup PE
25942

Two sub-types are defined for route-type 1. Route-type 4 has no sub-types. The route types used for EVPN-VPWS have the following purposes:

- Route-type 1 - Auto-discovery per EVPN instance (AD per-EVI). This route type is used in all EVPN-VPWS scenarios, with or without multi-homing. For EVPN-VPWS, the Ethernet Tag field is encoded with the local Attachment Circuit (AC) of the advertising PE. This value is configured using the **service>bgp-evpn>local-ac-name>eth-tag** <*value*> command. The Route Distinguisher (RD), MPLS label, and the Ethernet Segment ID (ESI) are encoded as for EVPN-MPLS. The MPLS label field is used as service label. In case of multi-homing, AD per-EVI routes containing the same ESI are used to provide aliasing and a backup path to the PEs part of the ES. The L2 MTU is encoded with the service MTU configured in the Epipe. The flags used for EVPN-VPWS are:

  – Flag C will be set if a control word is configured in the service.
  – Flag P will be set if the advertising PE is primary PE.
    - If no multi-homing is used, there is no primary PE (P=0).
    - In all-active multi-homing, all PEs in the ES are primary (P=1).
    - In single-active multi-homing, only one PE per-EVI in the ES is primary (P=1).
  – Flag B will be set if the advertising PE is backup PE.
    - The B-flag is only set in case of single-active multi-homing and only for one PE, even if more than two PEs are present in the same single-active ES. The backup PE is the winner of the second Designated Forwarder (DF) election (excluding the DF). The remaining non-DF PEs send B=0.

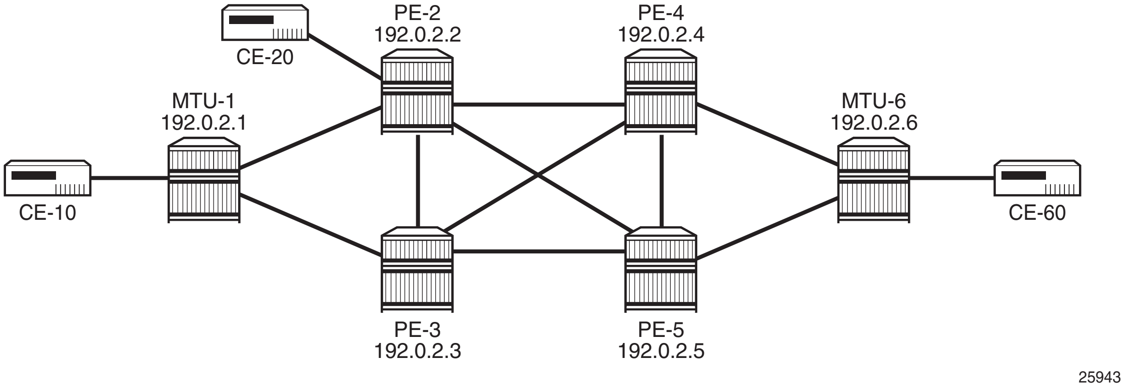
  If there is no multi-homing, the ESI, flag P, and flag B will be zero.

- Route-type 1 - AD per Ethernet segment (AD per-ES). Same encoding as for EVPN-MPLS. AD per-ES is only used in multi-homing scenarios where it is advertised per ES from the PE. It carries the ESI label (used for split-horizon, but only for VPLS services and not for Epipe services) and can affect procedures such as the DF election, as well as the aliasing on remote PEs.
- Route-type 4 - ES route. Same encoding as for EVPN-MPLS. Route-type 4 is only used in multi-homing scenarios. This route advertises a local configured ES. The exchange of this route can discover remote PEs that are part of the same ES and the DF election algorithm among them.

Configuration

Figure 45 shows the example topology that will be used throughout this chapter.

Figure 45 **EVPN-VPWS Example Topology**



The example topology consists of six 7750 SR routers with the following initial configuration:

- Network (or hybrid) ports interconnect the core PEs with configured router interfaces.
- MTU-1 is a pure Ethernet aggregator. The ports toward the core PEs are access ports. Likewise, the ports on PE-2 and PE-3 toward MTU-1 are access ports.
- Core PEs and MTU-6 run IS-IS on all interfaces. Point-to-point adjacencies are established for the exchange of system IP addresses.
- Link LDP is configured between all PEs, and toward/from MTU-6.
- EVPN uses BGP for exchanging reachability at service level. Therefore, BGP peering sessions must be established among the core PEs for the EVPN family. Although typically a separate router is used, in this chapter, PE-2 is used as route reflector with the following BGP configuration:

```
configure
    router
        autonomous-system 64500
        bgp
            vpn-apply-import
            vpn-apply-export
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            split-horizon
            rapid-update evpn
            group "internal"
                family evpn
                cluster 1.1.1.1
                peer-as 64500
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
                neighbor 192.0.2.5
                exit
            exit
        exit
```

The BGP configuration on the other PEs is as follows:

```
configure
    router
        autonomous-system 64500
        bgp
            vpn-apply-import
            vpn-apply-export
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            split-horizon
            rapid-update evpn
            group "internal"
                family evpn
                peer-as 64500
                neighbor 192.0.2.2
                exit
            exit
        exit
```

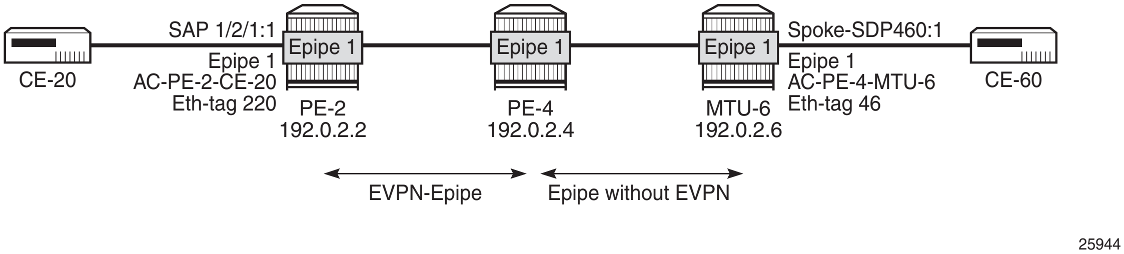
The following EVPN-VPWS scenarios are described in this section:

- EVPN for MPLS tunnels in Epipe services without multi-homing
- EVPN for MPLS tunnels in Epipe services with all-active multi-homing
- EVPN for MPLS tunnels in Epipe services with single-active multi-homing

EVPN for MPLS Tunnels in Epipe Services without Multi-Homing

BGP-EVPN can be enabled in Epipe services with either SAPs or spoke-SDPs at the access, as shown in Figure 46.

Figure 46 **Example Topology for EVPN-VPWS without Multi-Homing**



On PE-2, Epipe 1 is configured as follows:

```
configure
    service
        epipe 1 customer 1 create
            bgp
            exit
            bgp-evpn
                local-ac-name AC-PE-2-CE-20
                    eth-tag 220
                exit
                remote-ac-name AC-PE-4-MTU-6
                    eth-tag 46
                exit
                evi 1
                mpls
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            sap 1/2/1:1 create
            exit
            no shutdown
```

On PE-4, the service configuration is as follows:

```
configure
    service
        sdp 460 create
            far-end 192.0.2.6
            no shutdown
        exit
        epipe 1 customer 1 create
            bgp
            exit
            bgp-evpn
                local-ac-name AC-PE-4-MTU-6
        eth-tag 46
    exit
    remote-ac-name AC-PE-2-CE-20
        eth-tag 220
    exit
    evi 1
    mpls
        auto-bind-tunnel
            resolution any
        exit
        no shutdown
    exit
exit
spoke-sdp 460:1 create
exit
no shutdown
```

Where the following commands are relevant for the EVPN-VPWS configuration:

- **bgp** enables the context for the BGP configuration relevant to the service. The bgp context configures the common BGP parameters for all BGP families in the service, such as route distinguisher and route target. Even if the general BGP parameters for the service are auto-derived, the **bgp** context must be enabled.

```
*A:PE-2# configure service epipe 1 bgp
  - bgp
  - no bgp

 [no] pw-template-bi* + Configure pw-template bind policy
 [no] route-distingu* - Configure route distinguisher
 [no] route-target    - Configure route target
```

Note: No pw-template binding is supported for EVPN. Although it is included in the configuration, pw-template binding cannot be configured if the service is EVPN.

- The following parameters can be configured in the bgp-evpn context:

```
*A:PE-2# configure service epipe 1 bgp-evpn
  - bgp-evpn
  - no bgp-evpn

 [no] evi             - EVPN Identifier
 [no] local-ac-name   + Configure local attachment-circuit name
      mpls            + Configure BGP EVPN mpls
 [no] remote-ac-name  + Configure remote attachment-circuit name
```

  – The **evi** is a 2-byte identifier used for auto-deriving the service RD, service RT, and for the **service-carving** when multi-homing is used. The auto-derivation of RD and RT is as follows:

    - RD system-ip:evi
    - RT autonomous-system:evi

The EVI values must be unique in the system, regardless of the type of service they are assigned to (Epipe or VPLS).

- The **local-ac-name** and r**emote-ac-name** identify the two attachment circuits connected by the EVPN-VPWS service. The configured Ethernet tag for the local AC is advertised in the Ethernet Tag field of the AD per-EVI route for the Epipe, along with the corresponding RD, RT, and MPLS label. Both local and remote Ethernet tags are mandatory to bring up the Epipe service. If the received Ethernet tag for the Epipe service matches the configured remote AC **eth-tag**, it will create an EVPN-MPLS destination to the next hop.

  The local **eth-tag** cannot be modified without shutting down bgp-evpn mpls in the Epipe, as shown in the following output:

```
*A:PE-2# configure service epipe 1 bgp-evpn local-ac-name AC-PE-2-CE-20 eth-tag 221
MINOR: SVCMGR #8025 Evpn mpls is enabled
```

  Unlike local eth-tags, remote eth-tags can be modified without shutdown.

- The following configuration options are available for Epipes in the **bgp-evpn>mpls** context:

```
*A:PE-2# configure service epipe 1 bgp-evpn mpls
  - mpls

      auto-bind-tunn* + Configure BGP EVPN mpls auto-bind-tunnel
 [no] control-word    - Enable/disable setting the CW bit in the label message
      ecmp            - Configure maximum ECMP routes information
 [no] entropy-label   - Enable/disable use of entropy-label
 [no] force-vlan-vc-* - Forces vlan-vc-type forwarding in the data-path
 [no] send-evpn-encap - Configure encapsulation for this service
 [no] shutdown        - Administratively Enable/Disable BGP-EVPN mpls
```

  This is a subset of the options for VPLS services; see chapter EVPN for MPLS Tunnels.

When the local AC (SAP 1/2/1:1) is up, PE-2 sends a BGP EVPN AD per-EVI route that contains Ethernet tag 220 for the local AC:

```
4 2017/05/08 05:48:54.89 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.4
"Peer 1: 192.0.2.4: UPDATE
Peer 1: 192.0.2.4 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 88
    Flag: 0x90 Type: 14 Len: 36 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.2
        Type: EVPN-AD Len: 25 RD: 192.0.2.2:1 ESI: ESI-0, tag: 220 Label: 4194208

    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:1
        l2-attribute:MTU: 1514 C: 0 P: 0 B: 0
        bgp-tunnel-encap:MPLS
"
```

The auto-derived RD is 192.0.2.1:1 and the RT is 64500:1.

When the remote AC on PE-4 (spoke-SDP 460:1) is up, PE-2 receives the following BGP update from PE-4:

```
5 2017/05/08 05:49:31.98 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.4
"Peer 1: 192.0.2.4: UPDATE
Peer 1: 192.0.2.4 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 88
    Flag: 0x90 Type: 14 Len: 36 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.4
        Type: EVPN-AD Len: 25 RD: 192.0.2.4:1 ESI: ESI-0, tag: 46 Label: 4194208

    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:1
        l2-attribute:MTU: 1514 C: 0 P: 0 B: 0
        bgp-tunnel-encap:MPLS
"
```

When the received RT matches and the received Ethernet Tag matches the configured remote AC, the EVPN-MPLS destination (comprised of a Termination Endpoint (TEP) and Egress Label) will be created on PE-2 and PE-4:

```
*A:PE-2# show service id 1 evpn-mpls

===============================================================================
BGP EVPN-MPLS Dest
===============================================================================
TEP Address               Egr Label               Last Change
                           Transport
-------------------------------------------------------------------------------
192.0.2.4                 262138                  05/08/2017 06:32:10
                          ldp
-------------------------------------------------------------------------------
Number of entries : 1
-------------------------------------------------------------------------------
===============================================================================
---snip---
```

The MPLS label in the debug message is not the same as in the service, because the router will strip the extra four lowest bits to get the 20-bit MPLS label. The egress label for the EVPN-MPLS destination on PE-4 is 262138. The 24-bit label value in the BGP update debug is 16 (2^4) times as high: 262138*16 = 4194208. This is because the debug message is shown before the router can parse the label field and see if it corresponds to an MPLS label (20 bits) or a VXLAN VNI (24 bits).

The BGP AD per-EVI routes for Ethernet Tag 46 can be shown with the following command:

```
*A:PE-2# show router bgp routes evpn auto-disc tag 46
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN Auto-Disc Routes
===============================================================================
Flag  Route Dist.         ESI                           NextHop
      Tag                                               Label
-------------------------------------------------------------------------------
u*>i  192.0.2.4:1         ESI-0                         192.0.2.4
      46                                                LABEL 262138

-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-2#
```

BGP EVPN information for Epipe 1 can be shown with the following command:

```
*A:PE-2# show service id 1 bgp-evpn

===============================================================================
BGP EVPN Table
===============================================================================
EVI                : 1                 Creation Origin    : manual
===============================================================================
BGP EVPN MPLS Information
===============================================================================
Admin Status       : Enabled
Force Vlan Fwding  : Disabled          Control Word       : Disabled
Max Ecmp Routes    : 0
Local AC Name      : AC-PE-2-CE-20     Eth Tag            : 220
Remote AC Name     : AC-PE-4-MTU-6     Eth Tag            : 46
Entropy Label      : Disabled
===============================================================================

===============================================================================
BGP EVPN MPLS Auto Bind Tunnel Information
===============================================================================
Resolution         : any
```

```
Filter Tunnel Types: (Not Specified)
===============================================================================
*A:PE-2#
```

➡ **Note:** Each PE will send its service MTU into the L2 MTU field in the L2-attribute in the AD per-EVI route for the Epipe service. The received L2 MTU will be checked. In case of a mismatch between the received MTU and the configured service MTU, the router will not set up the EVPN destination and, therefore, the service will not come up.

EVPN for MPLS Tunnels in Epipe Services with Multi-Homing

SR OS supports EVPN multi-homing as per *draft-ietf-bess-evpn-vpws*.

The EVPN multi-homing implementation is based on the concept of the Ethernet Segment (ES). An ES is a logical structure that can be defined in one or more PEs and identifies the CE (or access network) multi-homed to the EVPN PEs. An ES is associated with a port, LAG, or SDP object, and is shared by all the services defined on those objects. It can also be shared between Epipe and VPLS services.

Each ES has a unique Ethernet Segment Identifier (ESI) that is 10 bytes and is manually configured. The ESI is advertised in the control plane to all the PEs in an EVPN network; therefore, it is very important to ensure that the 10-byte ESI value is unique throughout the entire network. Single-homed CEs are assumed to be connected to an ES with ESI = 0 (single-homed ESs are not explicitly configured).

The ES is part of the base BGP-EVPN configuration and is not applied to any EVPN-MPLS service, by default. An ES can be shared by multiple services; the association of a specific SAP or spoke-SDP to an ES is automatically made when the SAP is defined in the same LAG or port configured in the ES, or when the spoke-SDP is defined in the same SDP configured in the ES.

Regardless of the multi-homing mode, the local Ethernet Tag values must match on all the PEs that are part of the same ES. The PEs in the ES will use the AD per-EVI routes from the peer PEs to validate the PEs as DF election candidates for an EVI. The DF election is only relevant for single-active multi-homing ESs. For Epipes defined in an all-active multi-homing ES, there is no DF election required, because all PEs are forwarding traffic and all traffic is treated as unicast.

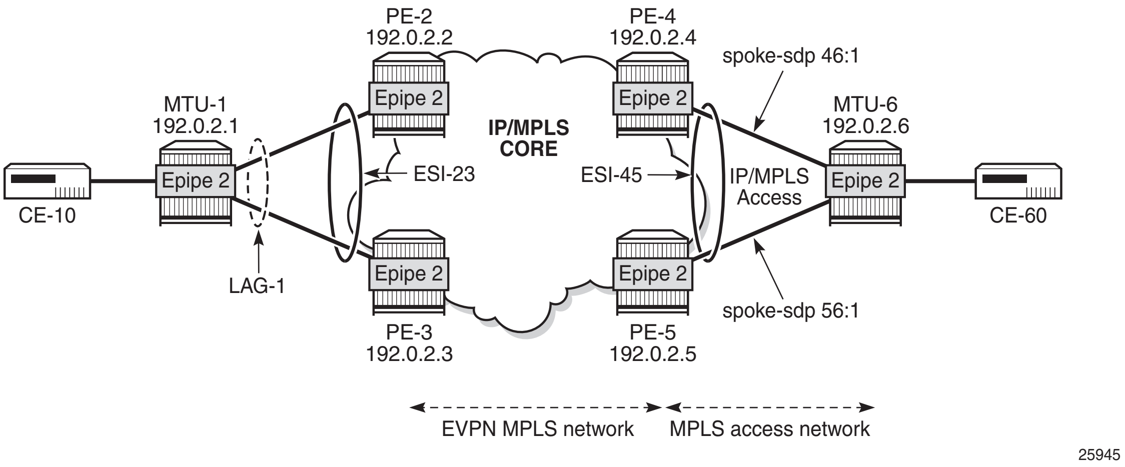
Aliasing is supported when sending traffic to an ES destination. Assuming ECMP is enabled on the ingress PE (and shared queuing or ingress policing), per-flow load-balancing will be performed among all the PEs that advertised P=1. PEs advertising P=0 are not considered as next hops for an ES destination.

The following sections show the configuration of:

- an all-active multi-homing ES with a LAG associated with it
- a single-active multi-homing ES linked to an SDP

Figure 47 shows the example topology has an all-active multi-homing ES "ESI-23" with a LAG associated to it in PE-2 and PE-3. A single-active multi-homing ES "ESI-45" with an SDP associated to it is configured in PE-4 and PE-5.

Figure 47 **Example Topology EVPN-VPWS with Multi-Homing**



EVPN for MPLS Tunnels in Epipe Services with All-Active Multi-Homing

All-active multi-homing allows for per-flow load-balancing. Unlike EVPN-MPLS in VPLS services, EVPN-VPWS has no DF election in all-active multi-homing. All PEs in the ES are active and the remote PE will do per-flow load-balancing. ESI-23 is configured on PE-2 and PE-3 in all-active multi-homing with LAG 1 associated to it. This LAG is used as a SAP in Epipe 2 on both PE-2 and PE-3. The configuration of the ES and Epipe 2 is identical on PE-2 and PE-3, including the local AC and remote AC names and eth-tags:

```
configure
    service
        system
            bgp-evpn
                ethernet-segment "ESI-23" create
```

```
                    esi 01:00:00:00:00:23:00:00:00:01
                    es-activation-timer 3
                    service-carving
                        mode auto
                    exit
                    multi-homing all-active
                    lag 1
                    no shutdown
                exit
            exit
        exit
        epipe 2 customer 1 create
            bgp
            exit
            bgp-evpn
                local-ac-name AC-ESI-23-MTU-1
                    eth-tag 231
                exit
                remote-ac-name AC-ESI-45-MTU-6
                    eth-tag 456
                exit
                evi 2
                mpls
                    ecmp 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            sap lag-1:1 create
            exit
            no shutdown
```

See chapter EVPN for MPLS Tunnels for a detailed explanation of the configuration parameters of the ES.

In EVPN-VPWS multi-homing scenarios, three route types are exchanged: AD per-EVI, AD per-ES, and ES routes. The following ES route (route-type 4) for ESI 01:00:00:00:00:23:00:00:00:01 sent by PE-2 is imported at PE-3:

```
77 2017/05/08 10:48:20.75 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 70
    Flag: 0x90 Type: 14 Len: 34 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.2
        Type: EVPN-Eth-Seg Len: 23 RD: 192.0.2.2:0
              ESI: 01:00:00:00:00:23:00:00:00:01, IP-Len: 4 Orig-IP-Addr: 192.0.2.2

    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:00:00:00:00:23:00
"
```

The target 00:00:00:00:23:00 in the extended community is derived from the ESI (bytes 2 to 7) and is only imported by the PEs that are part of the same ES; that is, PE-2 and PE-3 in this example.

At the same time, the following AD per-ES route (route-type 1) with Maximum Ethernet Tag (MAX-ET, all Fs) and label 0 is sent by RR PE-2 and imported by the rest of the PEs. The following two BGP updates with MAX-ET are received by PE-4:

```
70 2017/05/08 10:48:00.93 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 80
    Flag: 0x90 Type: 14 Len: 36 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.2
        Type: EVPN-AD Len: 25 RD: 192.0.2.2:2 ESI: 01:00:00:00:00:23:00:00:00:01,
                      tag: MAX-ET Label: 0

    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:64500:2
        esi-label:262137/All-Active
"
```

```
78 2017/05/08 10:48:20.75 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 94
    Flag: 0x90 Type: 14 Len: 36 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.3
        Type: EVPN-AD Len: 25 RD: 192.0.2.3:2
                      ESI: 01:00:00:00:00:23:00:00:00:01, tag: MAX-ET Label: 0

    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.3
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        1.1.1.1
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:64500:2
        esi-label:262138/All-Active
"
```

The ESI label is in the extended community, as well as the indication that the multi-homing is all-active. Epipe services do not require ESI labels because BUM traffic is not recognized as such in EVPN-VPWS services. However, because the ES can be shared by Epipe and VPLS services, the AD per-ES route still includes a non-zero ESI label.

The following AD per-EVI routes (route-type 1) with Ethernet Tag 231 sent by RR PE-2 are received and imported on PE-4:

```
81 2017/05/08 10:48:20.75 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 88
    Flag: 0x90 Type: 14 Len: 36 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.2
        Type: EVPN-AD Len: 25 RD: 192.0.2.2:2
                      ESI: 01:00:00:00:00:23:00:00:00:01, tag: 231 Label: 4194176

    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:2
        l2-attribute:MTU: 1514 C: 0 P: 1 B: 0
        bgp-tunnel-encap:MPLS
"

92 2017/05/08 08:48:52.94 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 102
    Flag: 0x90 Type: 14 Len: 36 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.3
        Type: EVPN-AD Len: 25 RD: 192.0.2.3:2 ESI: 01:00:00:00:00:23:00:00:00:01,
                      tag: 231 Label: 4194192

    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.3
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        1.1.1.1
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:2
        l2-attribute:MTU: 1514 C: 0 P: 1 B: 0
        bgp-tunnel-encap:MPLS
"
```

This route contains the flags for control word (C), primary (P), and backup (B). In all-active multi-homing, all nodes are primary (P=1).

PE-4 has learned AD per-EVI/ES routes for ESI-23 from PE-2 and PE-3, as shown in the following output:

```
*A:PE-4# show router bgp routes evpn auto-disc esi 01:00:00:00:00:23:00:00:00:01
===============================================================================
 BGP Router ID:192.0.2.4        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN Auto-Disc Routes
===============================================================================
Flag  Route Dist.         ESI                           NextHop
      Tag                                               Label
-------------------------------------------------------------------------------
u*>i  192.0.2.2:2         01:00:00:00:00:23:00:00:00:01 192.0.2.2
      231                                               LABEL 262136
u*>i  192.0.2.2:2         01:00:00:00:00:23:00:00:00:01 192.0.2.2
      MAX-ET                                            LABEL 0
u*>i  192.0.2.3:2         01:00:00:00:00:23:00:00:00:01 192.0.2.3
      231                                               LABEL 262137
u*>i  192.0.2.3:2         01:00:00:00:00:23:00:00:00:01 192.0.2.3
      MAX-ET                                            LABEL 0
-------------------------------------------------------------------------------
Routes : 4
===============================================================================
*A:PE-4#
```

For Epipe 2 on PE-4, the EVPN MPLS destination is not pointing at a specific TEP, but ESI-23, as shown in the following output:

```
*A:PE-4# show service id 2 evpn-mpls

===============================================================================
BGP EVPN-MPLS Dest
===============================================================================
TEP Address               Egr Label               Last Change
                           Transport
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================

===============================================================================
BGP EVPN-MPLS Ethernet Segment Dest
===============================================================================
Eth SegId                               Last Change
-------------------------------------------------------------------------------
01:00:00:00:00:23:00:00:00:01           05/08/2017 10:48:21
-------------------------------------------------------------------------------
Number of entries: 1
-------------------------------------------------------------------------------
```

```
================================================================================
*A:PE-4#
```

When ECMP > 1 on the ingress PE, multiple TEPs can correspond to a specific ESI (aliasing). In this case, ECMP=2 and PE-4 and PE-5 have two TEP addresses and Egress labels for ESI 01:00:00:00:00:23:00:00:00:01, as shown for PE-4:

```
*A:PE-4# show service id 2 evpn-mpls esi 01:00:00:00:00:23:00:00:00:01

========================================================
BGP EVPN-MPLS Ethernet Segment Dest
========================================================
Eth SegId                       Last Change
--------------------------------------------------------
01:00:00:00:00:23:00:00:00:01   05/08/2017 10:48:21
========================================================

================================================================================
BGP EVPN-MPLS Dest TEP Info
================================================================================
TEP Address              Egr Label                Last Change
                         Transport
--------------------------------------------------------------------------------
192.0.2.2                262136                   05/08/2017 10:48:21
                         ldp
192.0.2.3                262137                   05/08/2017 10:48:21
                         ldp
--------------------------------------------------------------------------------
Number of entries : 2
--------------------------------------------------------------------------------
================================================================================
*A:PE-4#
```

Note: Even if ECMP is configured, the ingress router will not load-balance the traffic unless shared queuing or ingress policing is configured. This is not specific to EVPN but generic to the way Epipes forward traffic.

In all-active multi-homing for EVPN-VPWS, there is no DF election and all PEs in the ES are active. For ESI-23, both PE-2 and PE-3 are active/primary/DF, but there are no DF candidates, because there is no DF election:

```
* A:PE-2# show service system bgp-evpn ethernet-segment name "ESI-23" evi 2

================================================================================
EVI DF and Candidate List
================================================================================
EVI           SvcId         Actv Timer Rem      DF  DF Last Change
--------------------------------------------------------------------------------
2             2             0                   yes 05/09/2017 06:32:49
================================================================================
================================================================================
DF Candidates                           Time Added
--------------------------------------------------------------------------------
```

```
No entries found
===============================================================================
*A:PE-2#
```

Similarly, on PE-3:

```
*A:PE-3# show service system bgp-evpn ethernet-segment name "ESI-23" evi 2

===============================================================================
EVI DF and Candidate List
===============================================================================
EVI           SvcId         Actv Timer Rem      DF  DF Last Change
-------------------------------------------------------------------------------
2             2             0                   yes 05/09/2017 06:33:24
===============================================================================
===============================================================================
DF Candidates                           Time Added
-------------------------------------------------------------------------------
No entries found
===============================================================================
*A:PE-3#
```

To confirm that all-active multi-homing is working correctly, the following command shows all information related to a specific ESI; in this case, ESI-23 on PE-2:

```
* A:PE-2# show service system bgp-evpn ethernet-segment name "ESI-23" all

===============================================================================
Service Ethernet Segment
===============================================================================
Name                    : ESI-23
Eth Seg Type            : None
Admin State             : Enabled           Oper State         : Up
ESI                     : 01:00:00:00:00:23:00:00:00:01
Multi-homing            : allActive         Oper Multi-homing  : allActive
ES SHG Label            : 262137
Source BMAC LSB         : <none>
Lag Id                  : 1
ES Activation Timer     : 3 secs
Svc Carving             : auto              Oper Svc Carving   : auto
Cfg Range Type          : primary
===============================================================================

===============================================================================
EVI Information
===============================================================================
EVI                 SvcId               Actv Timer Rem      DF
-------------------------------------------------------------------------------
2                   2                   0                   yes
-------------------------------------------------------------------------------
Number of entries: 1
---snip---
```

EVPN for MPLS Tunnels in Epipe Services with Single-Active Multi-Homing

Single-active multi-homing allows for per-service load-balancing. Single-active multi-homing is configured on PE-4 and PE-5 with ES "ESI-45". Both PEs have an SDP to MTU-6, which is associated with the ES and to the Epipe service. The configuration of the local and remote AC names and Ethernet tags is identical on PE-4 and PE-5.

On PE-4, the service configuration is as follows:

```
configure
    service
        sdp 46 mpls create
            far-end 192.0.2.6
            ldp
            no shutdown
        exit
        system
            bgp-evpn
                ethernet-segment "ESI-45" create
                    esi 01:00:00:00:00:45:00:00:00:01
                    es-activation-timer 3
                    service-carving
                        mode auto
                    exit
                    multi-homing single-active
                    sdp 46
                    no shutdown
                exit
            exit
        exit
        epipe 2 customer 1 create
            bgp
            exit
            bgp-evpn
                local-ac-name "AC-ESI-45-MTU-6"
                    eth-tag 456
                exit
                remote-ac-name "AC-ESI-23-MTU-1"
                    eth-tag 231
                exit
                evi 2
                mpls
                    ecmp 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            spoke-sdp 46:2 create
            exit
            no shutdown
```

On PE-5, the configuration is similar, but with a different SDP:

```
configure
    service
        sdp 56 mpls create
            far-end 192.0.2.6
            ldp
            no shutdown
        exit
        system
            bgp-evpn
                ethernet-segment "ESI-45" create
                    esi 01:00:00:00:00:45:00:00:00:01
                    es-activation-timer 3
                    service-carving
                        mode auto
                    exit
                    multi-homing single-active
                    sdp 56
                    no shutdown
                exit
            exit
        exit
        epipe 2 customer 1 create
            bgp
            exit
            bgp-evpn
                local-ac-name "AC-ESI-45-MTU-6"
                    eth-tag 456
                exit
                remote-ac-name "AC-ESI-23-MTU-1"
                    eth-tag 231
                exit
                evi 2
                mpls
                    ecmp 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            spoke-sdp 56:2 create
            exit
            no shutdown
```

Three route types will be exchanged between the core PEs: AD per-EVI, AD per-ES, and ES routes.

PE-4 and PE-5 advertise ES routes that are only imported by them. As an example, the following is the ES route with originator PE-4 sent by RR PE-2 to PE-5. It contains a target 00:00:00:00:45:00 in the extended community that is derived from the ESI:

```
247 2017/05/08 10:49:17.85 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.5
"Peer 1: 192.0.2.5: UPDATE
Peer 1: 192.0.2.5 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 84
    Flag: 0x90 Type: 14 Len: 34 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.4
        Type: EVPN-Eth-Seg Len: 23 RD: 192.0.2.4:0
              ESI: 01:00:00:00:00:45:00:00:00:01, IP-Len: 4 Orig-IP-Addr: 192.0.2.4

    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.4
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        1.1.1.1
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:00:00:00:00:45:00
"
```

The AD per-ES route has a Maximum Ethernet Tag (MAX-TAG) and an ESI label in the extended community. The multi-homing mode is single-active. As in the case of all-active multi-homing, the ESI label is not used in Epipe services. The following BGP update with originator PE-4 is sent by RR PE-2 to its client PE-5:

```
250 2017/05/08 10:49:17.84 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.5
"Peer 1: 192.0.2.5: UPDATE
Peer 1: 192.0.2.5 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 94
    Flag: 0x90 Type: 14 Len: 36 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.4
        Type: EVPN-AD Len: 25 RD: 192.0.2.4:2
                      ESI: 01:00:00:00:00:45:00:00:00:01, tag: MAX-ET Label: 0

    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.4
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        1.1.1.1
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:64500:2
        esi-label:262136/Single-Active
"
```

The AD per-EVI route contains flags for primary and backup, which will be different for routes received from PE-4 and PE-5. In this case, PE-4 is primary in the single-active multi-homing ES (P=1):

```
256 2017/05/08 10:49:17.84 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.5
"Peer 1: 192.0.2.5: UPDATE
Peer 1: 192.0.2.5 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 102
    Flag: 0x90 Type: 14 Len: 36 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.4
        Type: EVPN-AD Len: 25 RD: 192.0.2.4:2
                      ESI: 01:00:00:00:00:45:00:00:00:01, tag: 456 Label: 4194160

    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.4
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        1.1.1.1
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:2
        l2-attribute:MTU: 1514 C: 0 P: 1 B: 0
        bgp-tunnel-encap:MPLS
"
```

PE-5 is backup in the single-active multi-homing ES (B=1):

```
258 2017/05/08 10:49:20.78 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.5
"Peer 1: 192.0.2.5: UPDATE
Peer 1: 192.0.2.5 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 88
    Flag: 0x90 Type: 14 Len: 36 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.5
        Type: EVPN-AD Len: 25 RD: 192.0.2.5:2
                      ESI: 01:00:00:00:00:45:00:00:00:01, tag: 456 Label: 4194192

    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:2
        l2-attribute:MTU: 1514 C: 0 P: 0 B: 1
        bgp-tunnel-encap:MPLS
"
```

The BGP EVPN AD routes can be shown with the following command:

```
*A:PE-2# show router bgp routes evpn auto-disc esi 01:00:00:00:00:45:00:00:00:01
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN Auto-Disc Routes
===============================================================================
Flag  Route Dist.         ESI                           NextHop
      Tag                                               Label
-------------------------------------------------------------------------------
u*>i  192.0.2.4:2         01:00:00:00:00:45:00:00:00:01 192.0.2.4
```

```
      456                                               LABEL 262135
u*>i  192.0.2.4:2         01:00:00:00:00:45:00:00:00:01 192.0.2.4
      MAX-ET                                            LABEL 0
u*>i  192.0.2.5:2         01:00:00:00:00:45:00:00:00:01 192.0.2.5
      456                                               LABEL 262137
u*>i  192.0.2.5:2         01:00:00:00:00:45:00:00:00:01 192.0.2.5
      MAX-ET                                            LABEL 0

-------------------------------------------------------------------------------
Routes : 4
===============================================================================
*A:PE-2#
```

For each PE in the single-active ES, there are two AD routes: the routes with MAX-ET are AD per-ES routes and the routes with a configured Ethernet Tag are AD per-EVI routes.

The EVPN MPLS destination for Epipe 2 on PE-2 is ESI-45, as shown in the following output:

```
*A:PE-2# show service id 2 evpn-mpls

===============================================================================
BGP EVPN-MPLS Dest
===============================================================================
TEP Address               Egr Label               Last Change
                           Transport
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================

===============================================================================
BGP EVPN-MPLS Ethernet Segment Dest
===============================================================================
Eth SegId                               Last Change
-------------------------------------------------------------------------------
01:00:00:00:00:45:00:00:00:01           05/08/2017 10:48:09
-------------------------------------------------------------------------------
Number of entries: 1
-------------------------------------------------------------------------------
===============================================================================
*A:PE-2#
```

The ESI is resolved to the TEP address of the primary (DF) PE-4, as follows:

```
*A:PE-2# show service id 2 evpn-mpls esi 01:00:00:00:00:45:00:00:00:01
========================================================
BGP EVPN-MPLS Ethernet Segment Dest
========================================================
Eth SegId                       Last Change
--------------------------------------------------------
01:00:00:00:00:45:00:00:00:01   05/08/2017 10:48:09
========================================================
===============================================================================
BGP EVPN-MPLS Dest TEP Info
===============================================================================
TEP Address              Egr Label                Last Change
                         Transport
-------------------------------------------------------------------------------
192.0.2.4                262135                   05/08/2017 10:48:09
                         ldp
-------------------------------------------------------------------------------
Number of entries : 1
-------------------------------------------------------------------------------
===============================================================================
*A:PE-2#
```

The DF election is key for the forwarding and backup functions in single-active multi-homing ESs. The PE elected as DF will be the primary for the ES in the Epipe and will unblock the SAP/spoke-SDP for upstream and downstream traffic. The rest of the PEs in the ES will bring their ES SAPs or spoke-SDPs operationally down.

PE-5 is a non-DF, as follows:

```
*A:PE-5# show service system bgp-evpn ethernet-segment name "ESI-45" evi 2

===============================================================================
EVI DF and Candidate List
===============================================================================
EVI           SvcId         Actv Timer Rem      DF  DF Last Change
-------------------------------------------------------------------------------
2             2             0                   no  05/09/2017 06:34:11
===============================================================================

===============================================================================
DF Candidates                           Time Added
-------------------------------------------------------------------------------
192.0.2.4                               05/09/2017 06:42:10
192.0.2.5                               05/09/2017 06:34:24
-------------------------------------------------------------------------------
Number of entries: 2
===============================================================================
*A:PE-5#
```

In single-active multi-homing, the service spoke-SDP (or SAP) is brought operationally down on the non-DF, as shown in the following output:

```
*A:PE-5# show service id 2 base
---snip---
-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sdp:56:2 S(192.0.2.6)                    Spok         0       8978    Up   Down
===============================================================================
*A:PE-5#
```

The spoke-SDP 56:2 is operationally down with a StandbyForMHProtocol flag:

```
*A:PE-5# show service id 2 sdp 56:2 detail | match Flag
```

```
Flags              : StandbyForMHProtocol
```

Two consecutive DF elections take place: the first DF election includes all PEs in the ES for that Epipe and determines which PE is the primary PE (flags P=1, B=0). The second DF election excludes this DF and determines which PE is the backup (P=0, B=1). All other PEs signal flags P=0 and B=0.

When the primary PE fails, AD per-ES/EVI withdrawal messages are sent to the remote PE, which will update its next hop to the backup. The backup PE takes over immediately without waiting for the **es-activation-timer** to bring up its SAP/spoke-SDP.

Ethernet Segment Failures

When the SDP toward the primary (DF) fails, the backup PE needs to take over. An SDP failure is emulated and log 99 on PE-4 shows that SDP 46 is operational down and PE-4 is no longer the DF:

```
89 2017/05/09 10:54:35.10 UTC MINOR: SVCMGR #2303 Base
"Status of SDP 46 changed to admin=up oper=down"

92 2017/05/09 10:54:35.10 UTC MINOR: SVCMGR #2094 Base
"Ethernet Segment:ESI-45, EVI:2, Designated Forwarding state changed to:false"
```

Remote PEs receive route withdrawal updates (unreachable NLRI) from former DF PE-4, for example on PE-2:

```
109 2017/05/09 10:54:24.24 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.4
"Peer 1: 192.0.2.4: UPDATE
Peer 1: 192.0.2.4 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 86
    Flag: 0x90 Type: 15 Len: 82 Multiprotocol Unreachable NLRI:
        Address Family EVPN
        Type: EVPN-AD Len: 25 RD: 192.0.2.4:2 ESI: 01:00:00:00:00:45:00:00:00:01,
              tag: MAX-ET Label: 0
        Type: EVPN-Eth-Seg Len: 23 RD: 192.0.2.4:0
              ESI: 01:00:00:00:00:45:00:00:00:01, IP-Len: 4 Orig-IP-Addr: 192.0.2.4
        Type: EVPN-AD Len: 25 RD: 192.0.2.4:2 ESI: 01:00:00:00:00:45:00:00:00:01,
              tag: 456 Label: 0
"
```

The backup PE-5 is promoted to primary (P=1, B=0) and sends BGP updates accordingly. The following AD per-EVI is received on PE-2:

```
112 2017/05/09 10:54:24.25 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.5
"Peer 1: 192.0.2.5: UPDATE
Peer 1: 192.0.2.5 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 88
    Flag: 0x90 Type: 14 Len: 36 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.5
        Type: EVPN-AD Len: 25 RD: 192.0.2.5:2 ESI: 01:00:00:00:00:45:00:00:00:01,
              tag: 456 Label: 4194192
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:2
        l2-attribute:MTU: 1514 C: 0 P: 1 B: 0
        bgp-tunnel-encap:MPLS
"
```

PE-5 brings up its spoke-SDP without waiting for the es-activation-timer and takes over immediately. It is now the only DF candidate, and therefore the DF, as follows:

```
*A:PE-5# show service system bgp-evpn ethernet-segment name "ESI-45" evi 2
===============================================================================
EVI DF and Candidate List
===============================================================================
EVI           SvcId         Actv Timer Rem      DF  DF Last Change
-------------------------------------------------------------------------------
2             2             0                   yes 05/09/2017 10:50:53
===============================================================================

===============================================================================
DF Candidates                           Time Added
-------------------------------------------------------------------------------
192.0.2.5                               05/09/2017 10:51:07
-------------------------------------------------------------------------------
Number of entries: 1
===============================================================================
*A:PE-5#
```

BGP updates are exchanged and the remote PEs will resolve the ESI to the TEP address 192.0.2.5. For example, on PE-2:

```
*A:PE-2# show service id 2 evpn-mpls esi 01:00:00:00:00:45:00:00:00:01

========================================================
BGP EVPN-MPLS Ethernet Segment Dest
========================================================
Eth SegId                       Last Change
--------------------------------------------------------
01:00:00:00:00:45:00:00:00:01   05/09/2017 10:54:24
========================================================

===============================================================================
BGP EVPN-MPLS Dest TEP Info
===============================================================================
TEP Address              Egr Label                 Last Change
                         Transport
-------------------------------------------------------------------------------
192.0.2.5                262137                    05/09/2017 10:54:24
                         ldp
```

```
-------------------------------------------------------------------------------
Number of entries : 1
-------------------------------------------------------------------------------
===============================================================================
*A:PE-2#
```

This process is always revertive; as soon as the SDP 46 is operationally up again, a new DF election is triggered with two DF candidates and PE-4 will be elected as DF.

Troubleshooting and Debugging

The following show and debug commands can be used in EVPN-VPWS:

- show redundancy bgp-evpn-multi-homing
- show router bgp routes evpn (and filters)
- show service evpn-mpls [<TEP ip-address>]
- show service id bgp-evpn
- show service id evpn-mpls (and modifiers)
- show service system bgp-evpn
- show service system bgp-evpn ethernet-segment (and modifiers)
- debug router bgp update
- log-id 99

Most of these commands have been shown in the preceding sections; some commands are shown in this section.

Information about the configured boot timers (before DF election) and ES activation timer (after the system has been elected DF) can be shown as follows:

```
*A:PE-2# show redundancy bgp-evpn-multi-homing

===============================================================================
Redundancy BGP EVPN Multi-homing Information
===============================================================================
Boot-Timer              : 10 secs
Boot-Timer Remaining    : 0 secs
ES Activation Timer     : 3 secs
===============================================================================
*A:PE-2#
```

See chapter EVPN for MPLS Tunnels for a description of these timers.

The following command shows that the BGP route-type 4 (ES route) messages are only imported by the PEs in the same ES; for example, on PE-3:

```
*A:PE-3# show router bgp routes evpn eth-seg
===============================================================================
 BGP Router ID:192.0.2.3        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN Eth-Seg Routes
===============================================================================
Flag  Route Dist.         ESI                           NextHop
      OrigAddr
-------------------------------------------------------------------------------
u*>i  192.0.2.2:0         01:00:00:00:00:23:00:00:00:01 192.0.2.2
      192.0.2.2
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-3#
```

On PE-4:

```
* A:PE-4# show router bgp routes evpn eth-seg
===============================================================================
 BGP Router ID:192.0.2.4        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN Eth-Seg Routes
===============================================================================
Flag  Route Dist.         ESI                           NextHop
      OrigAddr
-------------------------------------------------------------------------------
u*>i  192.0.2.5:0         01:00:00:00:00:45:00:00:00:01 192.0.2.5
      192.0.2.5

-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-4#
```

The following command shows all the EVPN MPLS destinations toward TEP 192.0.2.4. Epipe 1 has an EVPN MPLS destination toward TEP 192.0.2.4 directly and Epipe 2 has an EVPN MPLS destination to ESI-45, which can be resolved to TEP 192.0.2.4. This is shown in the following output:

```
*A:PE-2# show service evpn-mpls 192.0.2.4

===============================================================================
BGP EVPN-MPLS Dest
```

```
===============================================================================
Service Id                              Egr Label
-------------------------------------------------------------------------------
1                                       262138
-------------------------------------------------------------------------------
===============================================================================

===============================================================================
BGP EVPN-MPLS Ethernet Segment Dest
===============================================================================
Service Id          Eth Seg Id                              Egr Label
-------------------------------------------------------------------------------
2                   01:00:00:00:00:45:00:00:00:01           262135
-------------------------------------------------------------------------------
===============================================================================

===============================================================================
BGP EVPN-MPLS ES BMac Dest
===============================================================================
Service Id               ES BMac                  Egr Label
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================
*A:PE-2#
```

The following command lists all configured ESs on the system:

```
*A:PE-2# show service system bgp-evpn ethernet-segment
===============================================================================
Service Ethernet Segment
===============================================================================
Name                             ESI                           Admin     Oper
-------------------------------------------------------------------------------
ESI-23                           01:00:00:00:00:23:00:00:00:01 Enabled   Up
-------------------------------------------------------------------------------
Entries found: 1
===============================================================================
*A:PE-2#
```

In addition to the preceding commands, the following tools dump commands may be useful:

- **tools dump service evpn usage -** This command shows the number of EVPN-MPLS (and EVPN-VXLAN) destinations in the system.
- **tools dump service system bgp-evpn ethernet-segment <name> evi <[1..65535]> df -** This command computes the DF election for a specific ESI and EVI. For all-active, there is no DF election and all PEs forward traffic. For single-active, one PE will be active for a service while another PE will be backup. This command shows the DF (primary), even if it is not the local PE.

The usage of EVPN resources can be shown as follows:

```
*A:PE-2# tools dump service evpn usage
EVPN usage statistics at 05/08/2017 13:18:41:
```

```
MPLS-TEP                                        :            1
VXLAN-TEP                                       :            0
Total-TEP                                       :     1/ 16383

Mpls Dests (TEP, Egress Label + ES + ES-BMAC)   :            2
Mpls Etree Leaf Dests                           :            0
Vxlan Dests (TEP, Egress VNI)                   :            0
Total-Dest                                      :     2/196607

Sdp Bind +  Evpn Dests                          :     2/245759
ES L2/L3 PBR                                    :     0/ 32767
Evpn Etree Remote BUM Leaf Labels               :            0
```

On PE-2, there is one MPLS-TEP (192.0.2.4 in Epipe 1 and Epipe 2) and there are two MPLS destinations: 192.0.2.4 and ESI 01:00:00:00:00:45:00:00:00:01. PE-5 is not an MPLS-TEP for PE-2, because it is not a primary and, therefore, not forwarding any traffic.

In all-active multi-homing, the DF election is not applicable:

```
*A:PE-2# tools dump service system bgp-evpn ethernet-segment "ESI-23" evi 2 df

[05/08/2017 13:32:21] All Active VPWS - DF N/A
```

In single-active multi-homing, the following command shows which PE is the DF:

```
*A:PE-5# tools dump service system bgp-evpn ethernet-segment "ESI-45" evi 2 df
[05/08/2017 13:33:38] Computed DF: 192.0.2.4 (Remote) (Boot Timer Expired: Yes)
[05/08/2017 13:33:38] Computed Backup: 192.0.2.5 (This Node)
```

The command is launched on PE-5, which is a backup. The computed DF is PE-4 and the boot timer has expired, meaning there is no DF re-election pending.

Conclusion

EVPN-VPWS is a simplified point-to-point version of *RFC7432 - BGP MPLS-Based Ethernet VPN*. When used for Epipe and VPLS services, EVPN provides a unified control plane mechanism that simplifies the network deployment and operation. Single-active and all-active multi-homing can be used in Epipes; EVPN-VPWS is a differentiator of EVPN compared to traditional TLDP or BGP Epipe redundancy mechanisms. The Ethernet Segments used for multi-homing can be shared between EVPN VPLS and EVPN Epipes.

EVPN for MPLS Tunnels in Routed VPLS

This chapter provides information about EVPN for MPLS Tunnels in Routed VPLS.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

The information and configuration in this chapter are based on SR OS Release 15.0.R4. EVPN-MPLS and IP-prefix advertisement in routed VPLS (R-VPLS) without Multi-homing (MH) is supported in SR OS Release 14.0.R1, and later. EVPN-MPLS and IP-prefix advertisement in R-VPLS with all-active and single-active MH is supported in SR OS Release 14.0.R4, and later. Virtual Router Redundancy Protocol (VRRP) in passive mode is also supported in SR OS Release 14.0.R4, and later.

Chapter EVPN for VXLAN Tunnels (Layer 3) is prerequisite reading.

Overview

The EVPN-MPLS in R-VPLS feature matches the EVPN-VXLAN in R-VPLS feature, which is described in chapter EVPN for VXLAN Tunnels (Layer 3). The following capabilities are supported in an R-VPLS service where **bgp-evpn mpls** is enabled:

- R-VPLS with Virtual Router Redundancy Protocol (VRRP) support on the VPRN interfaces
- R-VPLS support including **ip-route-advertisement** (IP prefix routes - BGP-EVPN route type 5) with regular interfaces
- R-VPLS support including **ip-route-advertisement** with **evpn-tunnel** interfaces
- R-VPLS with IPv6 support on the VPRN IP interface

All-active and single-active MH Ethernet segments (ESs) are supported in R-VPLS. When Ethernet Segments (ESs) are used along with R-VPLS services in two or more PEs, Passive VRRP provides an "anycast default gateway" that optimizes inter-subnet forwarding for hosts in the R-VPLS. Passive VRRP is described in the following section.

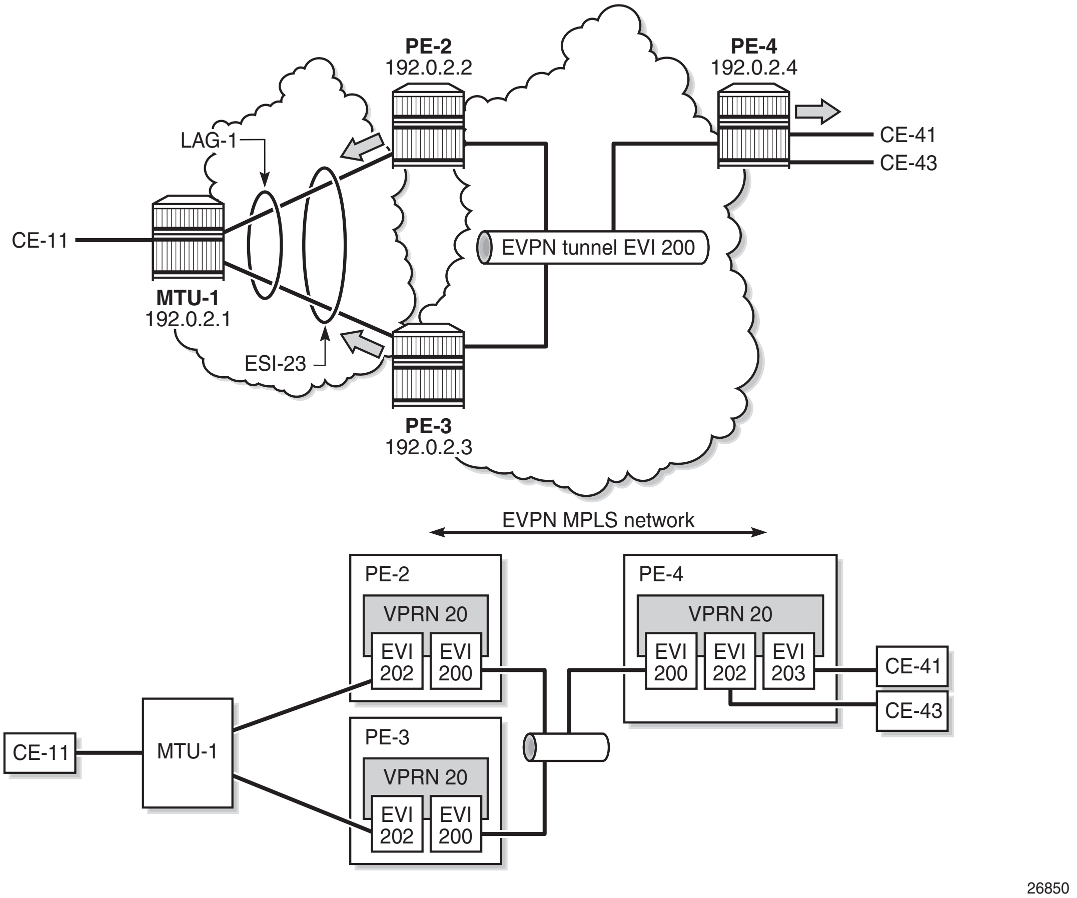
Passive VRRP

VRRP can be configured in passive mode, which suppresses the transmission and reception of keepalive messages. Passive mode can be enabled by adding the keyword **passive** at creation time. Passive mode cannot be enabled or disabled on the fly. Passive VRRP can be configured in the base router, in an IES, or in a VPRN, using the following commands:

```
*A:PE-2# tree flat detail | match vrrp | match passive
configure router interface ipv6 vrrp <virtual-router-id> [owner] [passive]
configure router interface vrrp <virtual-router-id> [owner] [passive]
configure service ies interface ipv6 vrrp <virtual-router-id> [owner] [passive]
configure service ies interface vrrp <virtual-router-id> [owner] [passive]
configure service vprn interface ipv6 vrrp <virtual-router-id> [owner] [passive]
configure service vprn interface vrrp <virtual-router-id> [owner] [passive]
```

All PEs configured with passive VRRP become VRRP master and take ownership of the virtual IP and MAC addresses. Figure 48 shows the use of passive VRRP where the VRID and default gateway (GW) are identical for all nodes, and therefore, the vMAC/vIP are identical. Each PE sends Gratuitous Address Resolution Protocol (GARP) messages with the same vMAC/vIP.

Figure 48 **Passive VRRP - vMAC/vIP Advertised By GARP**



Ethernet VPN instance (EVI) 202 is configured on all PEs as an R-VPLS with passive VRRP. Each individual R-VPLS interface has a unique MAC/IP, but they all have the same vMAC/vIP because they share the same VRID and backup IP. The vMAC is auto-derived out of 00:00:5e:00:00:<VRID>, as per RFC3768.

The behavior is as follows:

- PEs advertise their real MAC/IP and their vMAC/vIP in EVPN for EVI 202.
- All hosts in EVI 202 have a unique configured default GW.
- When a CE sends upstream traffic to a remote subnet, the packets are routed by the closest PE because the vMAC is local on each PE.

- In case of ES failure, or in case of single-active MH if the traffic arrives at the non-Designated Forwarder (NDF) PE, the traffic will not be discarded at the peer ES PE. Virtual MACs bypass the R-VPLS interface protection, so traffic can be forwarded between the PEs without being dropped. Note that if Passive VRRP was not used in this case and the same regular interface anycast MAC/IP was used instead, the peer PE would discard the traffic due to the MAC Source Address.

Passive VRRP provides an efficient anycast default gateway solution, with the following advantages compared to regular VRRP:

- No need for multiple VRRP instances to achieve default GW load-balancing. Only one VRRP instance is in the R-VPLS, so only one default GW is needed for all hosts.
- Fast convergence because all the nodes in the VRID are master.
- Better scalability because there is no need for keepalive messages or BFD to detect failures.

Passive VRRP provides the following advantages compared to using the same anycast MAC/IP in all the Integrated Routing Bridging (IRB) interfaces:

- VRRP vMAC source address (SA) bypasses the protection in the receiving R-VPLS service; therefore, frames with MAC SA matching the local vMAC are not discarded, and VRRP vMAC SAs can be used in combination with EVPN multi-homing.
- PEs will not show traps claiming duplicate IP addresses.
- vMACs are auto-derived from the VRID, so no need to configure the same MAC address in all the IRB interfaces.
- PEs can still use their real (unique) IRB IP addresses when sending ICMP packets for troubleshooting purposes.

Configuration

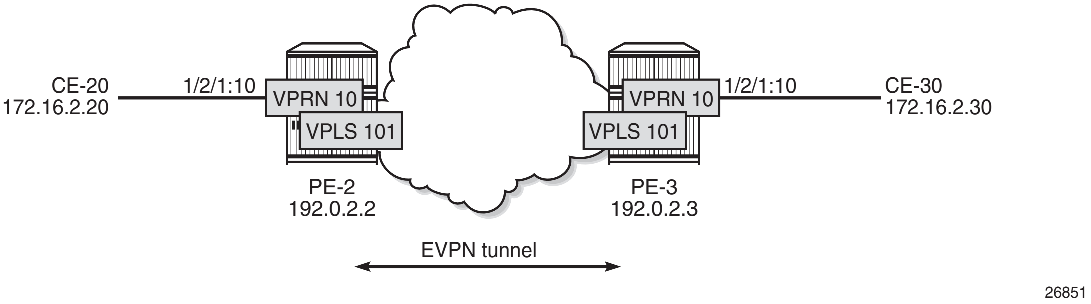
In this section, the following use cases will be described:

- EVPN-MPLS R-VPLS without multi-homing
- EVPN-MPLS R-VPLS with all-active multi-homing ES
- EVPN-MPLS R-VPLS with single-active multi-homing ES

EVPN-MPLS R-VPLS without Multi-homing

The first scenario describes R-VPLS support including IP route advertisement (BGP-EVPN route type 5) with EVPN tunnel interfaces, without multi-homing. VPLS 101 does not have any connected host, but the linked VPRN has SAP 1/2/1:10. Figure 49 shows the example topology used for R-VPLS with EVPN tunnel but without multi-homing. IP prefixes are advertised.

Figure 49 **R-VPLS with EVPN Tunnel, without Multi-homing**



The initial configuration includes the following:

- Cards, MDAs, ports
- Router interface between PE-2 and PE-3
- IS-IS (or OSPF)
- LDP enabled on the router interface between PE-2 and PE-3

BGP is configured for address family EVPN on PE-2 and PE-3. The BGP configuration on PE-2 is as follows. The BGP configuration on PE-3 is similar.

```
configure
    router
        autonomous-system 64500
        bgp
            family evpn
            vpn-apply-import
            vpn-apply-export
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            rapid-update evpn
            group "internal"
                peer-as 64500
                neighbor 192.0.2.3
                exit
            exit
        exit
```

The CEs are connected to SAP 1/2/1:10 in VPRN 10. R-VPLS 101 is bound to VPRN 10 and VPRN 10 has a dedicated interface "int-evi-100" for the EVPN tunnel. In general, if only one route-target (RT) is used for import and export in the EVPN-VPLS, it is good to add the EVI and have the route distinguisher (RD) and RT auto-derived from the EVI. It is simpler and avoids configuration mistakes. The service configuration on PE-2 is as follows:

```
configure
    service
        vprn 10 customer 1 create
            route-distinguisher 192.0.2.2:10
            vrf-target target:64500:10
            interface "int-PE-2-CE-20" create
                address 172.16.2.1/24
                sap 1/2/1:10 create
                exit
            exit
            interface "int-evi-101" create
                vpls "evi-101"
                    evpn-tunnel
                exit
            exit
            no shutdown
        exit
        vpls 101 name "evi-101" customer 1 create
            allow-ip-int-bind
            exit
            bgp# RD and RT are not manually configured in BGP context
            exit
            bgp-evpn
                ip-route-advertisement
                evi 101# RD and RT will be auto-derived from the EVI
                mpls
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            service-name "evi-101"
            no shutdown
        exit
```

- The **allow-ip-int-binding** command is required so that R-VPLS 101 can be bound to VPRN 10.
- The **service-name** command is required and the configured name "evi-101" must match the name in the VPRN 10 VPLS interface.
- The VPRN 10 VPLS interface is configured with the keyword **evpn-tunnel**. This configuration has the advantage of not having to allocate IP addresses to the R-VPLS interfaces, however it cannot be used when the R-VPLS has local SAPs.

The configuration is similar on PE-3. It is important that the RD is different on PE-2 and PE-3, but it is automatically the case when the RD is auto-derived from the configured EVI, as in the example. The RD on PE-2 is 192.0.2.2:101; on PE-3, the RD is 192.0.2.3:101.

PE-3 receives the following BGP-EVPN IP prefix route for prefix 172.16.2.0/24 from PE-2:

```
34 2017/07/13 12:21:38.10 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 97
    Flag: 0x90 Type: 14 Len: 45 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.2
        Type: EVPN-IP-Prefix Len: 34 RD: 192.0.2.2:101, tag: 0,
                             ip_prefix: 172.16.2.0/24 gw_ip 0.0.0.0 Label: 4194240
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:101
        mac-nh:16:0a:ff:ff:ff:a2
        bgp-tunnel-encap:MPLS
"
```

GW IP 0.0.0.0 is an indication that an EVPN tunnel is in use. With EVPN tunnels, no IRB IP address needs to be configured in the VPRN. EVPN tunnels make provisioning easier to automate and save IP addresses from the tenant IP space.

The BGP tunnel encapsulation is MPLS, but the MPLS label in the debug message is not the same as in the service, because the router will strip the extra four lowest bits to get the 20-bit MPLS label. In the debug message, the label is 4194240. This is because the debug message is shown before the router can parse the label field and see if it corresponds to an MPLS label (20 bits) or a VXLAN VNI (24 bits). The MPLS label is calculated by dividing the label value by 24 (16), as follows: 4194240/16 = 262140.

The MAC next-hop extended community 16:0a:ff:ff:ff:a2 is the MAC address of the interface "int-evi-101" in VPRN 10 on PE-2, as follows:

```
*A:PE-2# show service id 10 interface "int-evi-101" detail | match MAC
MAC Address      : 16:0a:ff:ff:ff:a2    Mac Accounting    : Disabled
```

The routing table for VPRN 10 on PE-3 contains the route for prefix 172.16.2.0/24 as the BGP-EVPN route with next-hop "int-evi-101" and interface name "ET-16:0a:ff:ff:ff:a2" (ET stands for EVPN Tunnel), as follows:

```
*A:PE-3# show router 10 route-table
```

```
===============================================================================
Route Table (Service: 10)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.2.0/24                                 Remote  BGP EVPN  00h06m45s  169
       int-evi-101 (ET-16:0a:ff:ff:ff:a2)                            0
172.16.3.0/24                                 Local   Local     00h06m48s  0
       int-PE-3-CE-30                                                0
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-3#
```

The forwarding database (FDB) for VPLS 101 on PE-3 shows an entry for MAC address 16:0a:ff:ff:ff:a2 that is learned via EVPN. The MAC address is static (S) and protected (P). The MPLS label is 262140.

```
*A:PE-3# show service id 101 fdb detail

===============================================================================
Forwarding Database, Service 101
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
101       16:0a:ff:ff:ff:a2 eMpls:                   EvpnS    07/13/17 12:55:59
                                                     P

                            192.0.2.2:262140
101       16:0b:ff:ff:ff:a2 cpm                      Intf     07/13/17 12:55:57
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-3#
```

When the CEs have IPv6 addresses, the VPRN configuration is similar on the PEs, but the **ipv6** context must be enabled in the EVPN tunnel interface, so that the router can advertise and process routes type 5 with IPv6 prefixes. The configuration of the VPLS is identical for IPv4 and IPv6.

```
configure
    service
        vprn 16 customer 1 create
            route-distinguisher 192.0.2.2:16
            vrf-target target:64500:16
            interface "int-PE-2-CE-26" create
                ipv6
                    address 2001:db8:16::2:1/120
                exit
                sap 1/2/1:16 create
                exit
            exit
            interface "int-evi-106" create
                ipv6
                exit
                vpls "evi-106"
                    evpn-tunnel
                exit
            exit
            no shutdown
        exit
        vpls 106 name "evi-106" customer 1 create
            allow-ip-int-bind
            exit
            bgp
            exit
            bgp-evpn
                ip-route-advertisement
                evi 106
                mpls
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            service-name "evi-106"
            no shutdown
        exit
```

When advertising IPv6 prefixes, the GW IP field in the route type 5 is always populated with the IPv6 address of the R-VPLS interface. In this example, because no specific IPv6 global address is configured, the GW IP will be populated with the auto-created link local address. The following BGP update is received by PE-3 for IPv6 prefix 2001:db8:16::2:0/120:

```
36 2017/07/13 12:21:38.10 123 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 113
    Flag: 0x90 Type: 14 Len: 69 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.2
        Type: EVPN-IP-Prefix Len: 58 RD: 192.0.2.2:106, tag: 0,
         ip_prefix: 2001:db8:16::2:0/120 gw_ip fe80::140a:1ff:fe02:1 Label: 4194144
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:64500:106
        bgp-tunnel-encap:MPLS
"
```

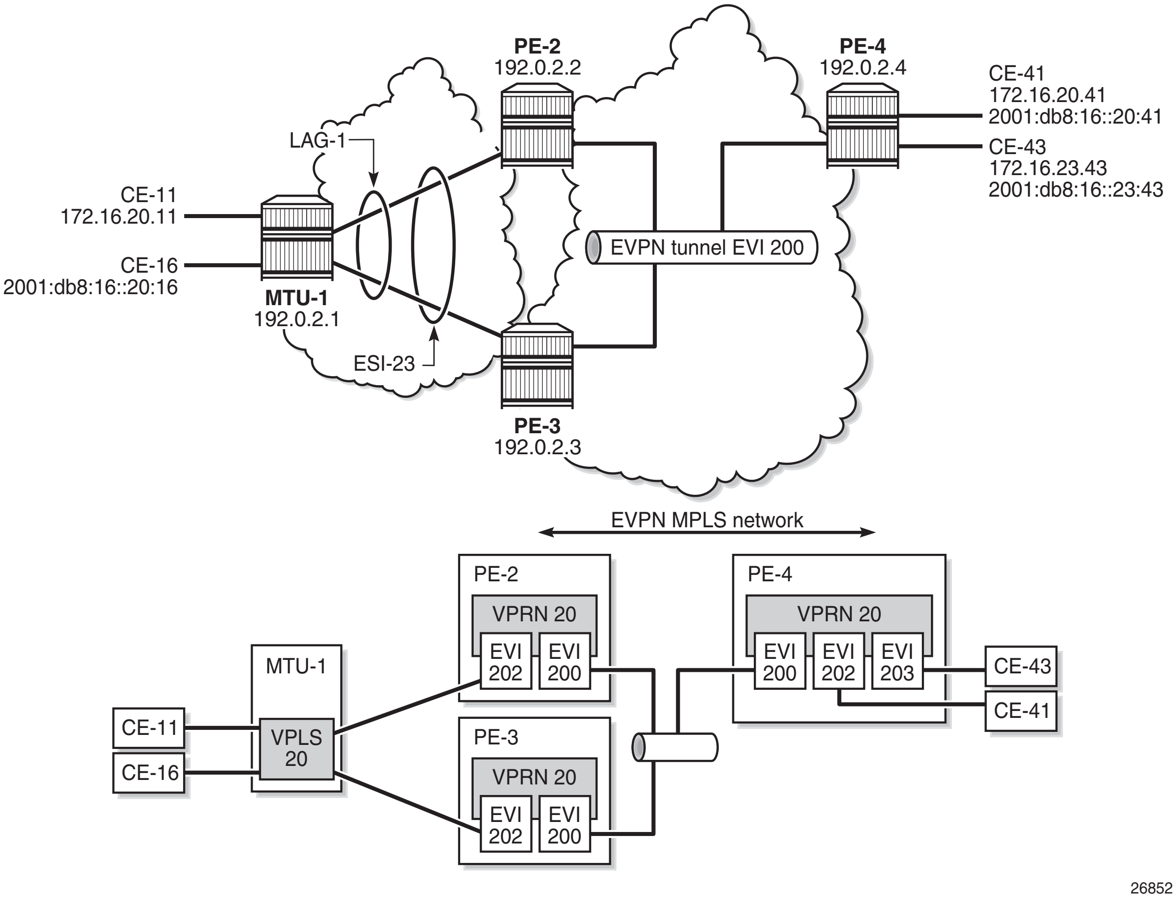
The IPv6 route-table on PE-3 is as follows:

```
*A:PE-3# show router 16 route-table ipv6

===============================================================================
IPv6 Route Table (Service: 16)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
2001:db8:16::2:0/120                          Remote  BGP EVPN  00h17m24s  169
       fe80::140a:1ff:fe02:1-"int-evi-106"                          0
2001:db8:16::3:0/120                          Local   Local     00h17m26s  0
       int-PE-3-CE-36                                               0
-------------------------------------------------------------------------------
No. of Routes: 2
```

EVPN-MPLS R-VPLS with All-active MH

Figure 50 shows the example topology with all-active multi-homing ES "ESI-23".

Figure 50 **EVPN-MPLS R-VPLS with All-Active MH ES**



BGP is configured between PE-2, PE-3, and PE-4 for address family EVPN. The configuration on PE-2 is as follows:

```
configure
    router
        autonomous-system 64500
        bgp
            family evpn
            vpn-apply-import
            vpn-apply-export
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            rapid-update evpn
            group "internal"
                peer-as 64500
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
            exit
        exit
```

All-active multi-homing Ethernet segment "ESI-23" is configured on PE-2 and PE-3, as follows:

```
configure
    service
        system
            bgp-evpn
                ethernet-segment "ESI-23" create
                    esi 01:00:00:00:00:23:00:00:00:01
                    es-activation-timer 3
                    service-carving
                        mode auto
                    exit
                    multi-homing all-active
                    lag 1
                    no shutdown
                exit
```

The following services are configured on the PEs:

- VPRN 20 has interfaces bound to VPLS 200 and VPLS 202. On PE-4, VPRN 20 also has an interface bound to VPLS 203.
- VPLS 200 is configured as an EVPN tunnel that connects the PEs.
- VPLS 202 and VPLS 203 have attachment circuits to CEs.

The services are configured on PE-2 as follows. The configuration on PE-3 and PE-4 is similar.

```
configure
    service
        vprn 20 customer 1 create
            route-distinguisher 192.0.2.2:20
            vrf-target target:64500:20
            interface "int-evi-202" create
                address 172.16.20.2/24
                mac 00:ca:fe:00:02:02
                vrrp 1 passive
                    backup 172.16.20.254
                    ping-reply
                    traceroute-reply
                exit
                ipv6
                    address 2001:db8:16::20:2/120
                    link-local-address fe80::16:20:2 dad-disable
                    vrrp 1 passive
                        backup fe80::16:20:fe
                        ping-reply
                        traceroute-reply
                    exit
                exit
                vpls "evi-202"
                exit
            exit
            interface "int-evi-200" create
                ipv6
                exit
                vpls "evi-200"
                    evpn-tunnel
                exit
            exit
            router-advertisement
                interface "int-evi-202"
                    use-virtual-mac
                    no shutdown
                exit
            exit
            no shutdown
        exit
        vpls 200 customer 1 create
            allow-ip-int-bind
            exit
            bgp
            exit
            bgp-evpn
                ip-route-advertisement
                evi 200
                mpls
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            service-name "evi-200"
            no shutdown
        exit
        vpls 202 customer 1 create
            allow-ip-int-bind
            exit
            bgp
            exit
            bgp-evpn
                evi 202
                mpls
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            service-name "evi-202"
            sap lag-1:20 create
            exit
            no shutdown
        exit
```

The IPv6 VRRP backup address is in the same subnet as the link local address of the interface "int-evi-202". The option **dad-disable** is configured on the link local address to disable Duplicate Address Detection (DAD) and set the IPv6 address as preferred. Also for IPv6, router advertisement must be enabled and configured to use the virtual MAC address.

Passive VRRP

EVI 202 is configured as an R-VPLS with passive VRRP. A passive-VRRP VRID instance suppresses the transmission and reception of keepalive messages. All PEs configured with passive VRRP become VRRP master and take ownership of the virtual IP and MAC address.

Each individual R-VPLS interface has a different MAC/IP on each PE. The MAC/IPs for "int-evi-202" on PE-2 are MAC 00:ca:fe:00:02:02 and IP 172.16.20.2/24 for IPv4 and the same MAC address with IPv6 2001:db8:16::20:2 and fe80::16:20:2. However, the R-VPLS interfaces on all PEs share the same VRID 1 and backup IP address 172.16.20.254, so the same vMAC/vIP 00:00:5e:00:01:01/172.16.20.254 and vMAC/vIP 00:00:5e:00:02:01/ fe80::16:20:fe are advertised by all PEs. PE-2 advertises the following EVPN MAC routes:

```
82 2017/07/13 12:20:38.60 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 292
    Flag: 0x90 Type: 14 Len: 240 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.2
        Type: EVPN-MAC Len: 49 RD: 192.0.2.2:202 ESI: ESI-0, tag: 0, mac len: 48
          mac: 00:00:5e:00:02:01, IP len: 16, IP: fe80::16:20:fe, label1: 4194080
        Type: EVPN-MAC Len: 37 RD: 192.0.2.2:202 ESI: ESI-0, tag: 0, mac len: 48
          mac: 00:00:5e:00:01:01, IP len: 4, IP: 172.16.20.254, label1: 4194080
        Type: EVPN-MAC Len: 49 RD: 192.0.2.2:202 ESI: ESI-0, tag: 0, mac len: 48
          mac: 00:ca:fe:00:02:02, IP len: 16, IP: fe80::16:20:2, label1: 4194080
        Type: EVPN-MAC Len: 49 RD: 192.0.2.2:202 ESI: ESI-0, tag: 0, mac len: 48
          mac: 00:ca:fe:00:02:02, IP len: 16, IP: 2001:db8:16::20:2, label1: 4194080
        Type: EVPN-MAC Len: 37 RD: 192.0.2.2:202 ESI: ESI-0, tag: 0, mac len: 48
          mac: 00:ca:fe:00:02:02, IP len: 4, IP: 172.16.20.2, label1: 4194080
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:202
        bgp-tunnel-encap:MPLS
        mac-mobility:Seq:0/Static
"
```

The three PEs advertise the same (anycast) vMAC/vIP in EVI 202 as protected, but each PE keeps its own MAC entry in the FDB. The following FDB shows that the source identifier for vMAC 00:00:5e:00:01:01 and vMAC 00:00:5e:00:02:01 is the CPM. These two vMAC entries with source identifier CPM are seen on all PEs.

```
*A:PE-2# show service id 202 fdb detail

===============================================================================
Forwarding Database, Service 202
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
202       00:00:01:00:00:11 sap:lag-1:20             L/0      07/13/17 12:00:35
202       00:00:01:00:00:16 sap:lag-1:20             L/0      07/13/17 12:00:36
202       00:00:04:00:00:41 eMpls:                   Evpn     07/13/17 11:57:24
                            192.0.2.4:262135
202       00:00:5e:00:01:01 cpm                      Intf     07/13/17 12:20:19
202       00:00:5e:00:02:01 cpm                      Intf     07/13/17 12:20:19
202       00:ca:fe:00:02:02 cpm                      Intf     07/13/17 11:56:56
202       00:ca:fe:00:02:03 eMpls:                   EvpnS    07/13/17 11:57:12
                                                     P
                            192.0.2.3:262130
202       00:ca:fe:00:02:04 eMpls:                   EvpnS    07/13/17 11:57:23
                                                     P
                            192.0.2.4:262135
-------------------------------------------------------------------------------
No. of MAC Entries: 8
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-2#
```

The interface MAC 00:ca:fe:00:02:02 is local, so it also has the CPM as source identifier. MAC 00:ca:fe:00:02:03 is the PE-3's R-VPLS interface MAC and it is learned via EVPN-MPLS (eMpls) as static (S) and protected (P). MAC address 00:ca:fe:00:02:04 on PE-4 is also static and protected.

PE-4 sends the following IP prefix route (BGP-EVPN route type 5) for prefix 172.16.23.0/24 to the other PEs:

```
35 2017/07/13 12:20:38.60 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 97
    Flag: 0x90 Type: 14 Len: 45 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.4
        Type: EVPN-IP-Prefix Len: 34 RD: 192.0.2.4:200, tag: 0,
                             ip_prefix: 172.16.23.0/24 gw_ip 0.0.0.0 Label: 4194192
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:200
        mac-nh:16:0c:ff:00:00:05
        bgp-tunnel-encap:MPLS
"
```

The IP prefixes are advertised with next-hop equal to the EVPN-tunnel GW MAC "int-evi-200", as follows:

```
*A:PE-4# show router 20 interface "int-evi-200" detail | match MAC
MAC Address      : 16:0c:ff:00:00:05    Mac Accounting    : Disabled
```

The routing table for VPRN 20 on PE-2 contains IP-prefix 172.16.23.0/24 with next-hop 16:0c:ff:00:00:05, as follows:

```
*A:PE-2# show router 20 route-table

===============================================================================
Route Table (Service: 20)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.20.0/24                                Local   Local     00h01m07s  0
       int-evi-202                                                  0
172.16.23.0/24                                Remote  BGP EVPN  00h00m48s  169
       int-evi-200 (ET-16:0c:ff:00:00:05)                           0
-------------------------------------------------------------------------------
No. of Routes: 2
```

The following IPv6 routing table for VPRN 20 on PE-2 contains prefix 2001:db8:16::23:0/120, which has also been advertised by PE-4. The next-hop is again "int-evi-200", only this time the link local ipv6 address is displayed (GW IP) instead of the MAC address. The next-hop is the GW IP value in the route type 5, as long as it is non-zero. When the GW IP is zero, the route type 5 is expected to contain a mac-nh extended community. The MAC encoded in the extended community is used as next-hop in that case.

```
*A:PE-2# show router 20 route-table ipv6

===============================================================================
IPv6 Route Table (Service: 20)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
2001:db8:16::20:0/120                         Local   Local     00h01m06s  0
       int-evi-202                                                  0
2001:db8:16::23:0/120                         Remote  BGP EVPN  00h00m47s  169
       fe80::21:7c98:7803:74d3 -"int-evi-200"                        0
-------------------------------------------------------------------------------
No. of Routes: 2
```

The EVPN tunnel service VPLS 200 has all the MAC addresses of the EVPN interfaces within VPRN 20 as static (S) and protected (P), as follows:

```
*A:PE-2# show service id "evi-200" fdb detail

===============================================================================
Forwarding Database, Service 200
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
200       16:0a:ff:00:00:05 cpm                      Intf     07/13/17 12:20:31
200       16:0b:ff:00:00:05 eMpls:                   EvpnS    07/13/17 12:20:39
                                                     P
                            192.0.2.3:262136
200       16:0c:ff:00:00:05 eMpls:                   EvpnS    07/13/17 12:20:51
                                                     P
                            192.0.2.4:262137
-------------------------------------------------------------------------------
No. of MAC Entries: 3
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-2#
```

The VRRP instance in each PE is master, as follows:

```
*A:PE-2# show router 20 vrrp instance

===============================================================================
VRRP Instances
===============================================================================
Interface Name                   VR Id Own Adm  State       Base Pri   Msg Int
                                 IP        Opr  Pol Id      InUse Pri  Inh Int
-------------------------------------------------------------------------------
int-evi-202                      1     No  Up   Master       100       1
                                 IPv4      Up   n/a         100        No
  Backup Addr: 172.16.20.254
int-evi-202                      1     No  Up   Master       100       1
                                 IPv6      Up   n/a         100        Yes
  Backup Addr: fe80::16:20:fe
-------------------------------------------------------------------------------
Instances : 2
===============================================================================
*A:PE-2#
```

```
*A:PE-3# show router 20 vrrp instance

===============================================================================
VRRP Instances
===============================================================================
Interface Name                   VR Id Own Adm  State       Base Pri   Msg Int
                                 IP        Opr  Pol Id      InUse Pri  Inh Int
-------------------------------------------------------------------------------
int-evi-202                      1     No  Up   Master       100       1
                                 IPv4      Up   n/a         100        No
  Backup Addr: 172.16.20.254
```

```
int-evi-202                      1     No  Up   Master       100       1
                                 IPv6      Up   n/a         100        Yes
  Backup Addr: fe80::16:20:fe
-------------------------------------------------------------------------------
Instances : 2
===============================================================================
*A:PE-3#

*A:PE-4# show router 20 vrrp instance
===============================================================================
VRRP Instances
===============================================================================
Interface Name                   VR Id Own Adm  State       Base Pri   Msg Int
                                 IP        Opr  Pol Id      InUse Pri  Inh Int
-------------------------------------------------------------------------------
int-evi-202                      1     No  Up   Master       100       1
                                 IPv4      Up   n/a         100        No
  Backup Addr: 172.16.20.254
int-evi-203                      2     No  Up   Master       100       1
                                 IPv4      Up   n/a         100        No
  Backup Addr: 172.16.23.254
int-evi-202                      1     No  Up   Master       100       1
                                 IPv6      Up   n/a         100        Yes
  Backup Addr: fe80::16:20:fe
int-evi-203                      2     No  Up   Master       100       1
                                 IPv6      Up   n/a         100        Yes
  Backup Addr: fe80::16:23:fe
-------------------------------------------------------------------------------
Instances : 4
===============================================================================
*A:PE-4#
```

Operation

On PE-4, VPRN 20 has one interface bound to VPLS 202 and another interface bound to VPLS 203. CE-41 is attached to VPLS 202, whereas CE-43 is attached to VPLS 203. When ping messages are sent from CE-41 to CE-43, or vice versa, the messages go via VPRN 20, which has routes to both CEs, as follows:

```
*A:PE-4# show router 20 route-table

===============================================================================
Route Table (Service: 20)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.20.0/24                                Local   Local      04h25m52s  0
       int-evi-202                                                  0
172.16.23.0/24                                Local   Local      04h25m51s  0
       int-evi-203                                                  0
-------------------------------------------------------------------------------
No. of Routes: 2
```

```
*A:PE-4# show router 20 route-table ipv6

===============================================================================
IPv6 Route Table (Service: 20)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
2001:db8:16::20:0/120                         Local   Local     00h00m50s  0
       int-evi-202                                                  0
2001:db8:16::23:0/120                         Local   Local     00h00m50s  0
       int-evi-203                                                  0
-------------------------------------------------------------------------------
No. of Routes: 2
```

When traffic is sent between CE-11 and CE-41, which are both associated with VPLS 202, the forwarding is done by the VPLS and not via the VPRN. The FDB for VPLS 202 on PE-2 is as follows:

```
*A:PE-2# show service id 202 fdb detail

===============================================================================
Forwarding Database, Service 202
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
202       00:00:01:00:00:11 sap:lag-1:20             L/60     07/13/17 12:20:43
202       00:00:01:00:00:16 sap:lag-1:20             L/60     07/13/17 12:20:49
202       00:00:04:00:00:41 eMpls:                   Evpn     07/13/17 12:20:38
                            192.0.2.4:262136
202       00:00:5e:00:01:01 cpm                      Intf     07/13/17 12:20:19
202       00:00:5e:00:02:01 cpm                      Intf     07/13/17 12:20:19
202       00:ca:fe:00:02:02 cpm                      Intf     07/13/17 12:20:18
202       00:ca:fe:00:02:03 eMpls:                   EvpnS    07/13/17 12:20:26
                                                     P
                            192.0.2.3:262133
202       00:ca:fe:00:02:04 eMpls:                   EvpnS    07/13/17 12:20:37
                                                     P
                            192.0.2.4:262136
-------------------------------------------------------------------------------
No. of MAC Entries: 8
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-2#
```

MAC 00:00:01:00:00:11 corresponds to CE-11 and is learned on SAP lag-1:20 on PE-2 and advertised via an EVPN MAC route to the BGP peers. MAC 00:00:04:00:00:41 corresponds to CE-41 and was advertised via an EVPN MAC route from PE-4, where the MAC was learned on SAP 1/2/1:41 of VPLS 202, as shown in the following FDB:

```
*A:PE-4# show service id 202 fdb detail

===============================================================================
Forwarding Database, Service 202
===============================================================================
ServId    MAC                 Source-Identifier        Type     Last Change
                                                       Age
-------------------------------------------------------------------------------
202       00:00:01:00:00:11 eES:                       Evpn     07/13/17 12:20:04
                            01:00:00:00:00:23:00:00:00:01
202       00:00:01:00:00:16 eES:                       Evpn     07/13/17 12:20:10
                            01:00:00:00:00:23:00:00:00:01
202       00:00:04:00:00:41 sap:1/2/1:41               L/60     07/13/17 12:19:59
202       00:00:5e:00:01:01 cpm                        Intf     07/13/17 12:19:58
202       00:00:5e:00:02:01 cpm                        Intf     07/13/17 12:19:58
202       00:ca:fe:00:02:02 eMpls:                     EvpnS    07/13/17 12:19:59
                                                       P
                            192.0.2.2:262133
202       00:ca:fe:00:02:03 eMpls:                     EvpnS    07/13/17 12:19:59
                                                       P
                            192.0.2.3:262133
202       00:ca:fe:00:02:04 cpm                        Intf     07/13/17 12:19:58
-------------------------------------------------------------------------------
No. of MAC Entries: 8
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-4#
```

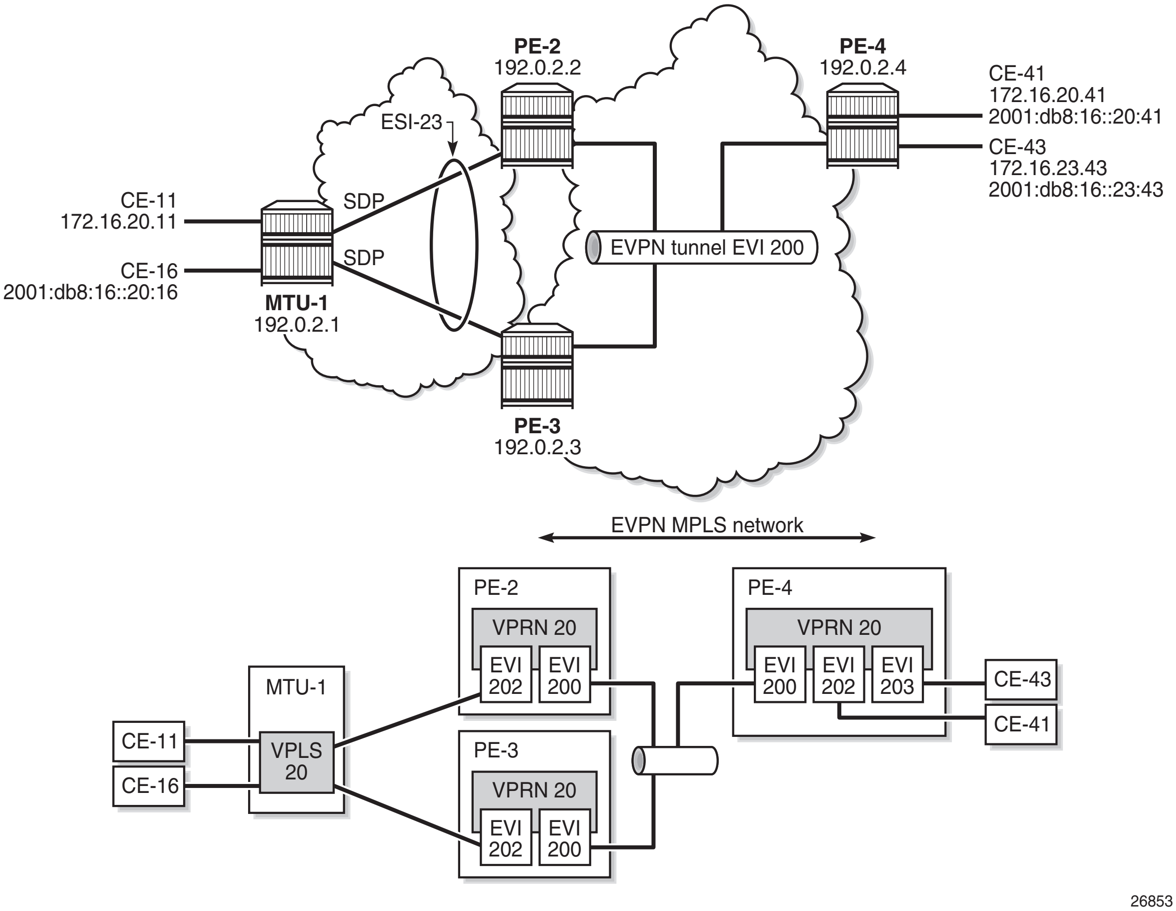
CE-43's MAC address is not present in VPLS 202's FDB. VPLS 203's FDB shows the CE-43's MAC address, but not CE-41's. Traffic between these two VPLS services goes via the VPRN and cannot use Layer 2 forwarding.

```
*A:PE-4# show service id 203 fdb detail

===============================================================================
Forwarding Database, Service 203
===============================================================================
ServId    MAC                 Source-Identifier        Type     Last Change
                                                       Age
-------------------------------------------------------------------------------
203       00:00:04:00:00:43 sap:1/2/1:43               L/60     07/13/17 12:20:32
203       00:00:5e:00:01:02 cpm                        Intf     07/13/17 12:20:16
203       00:00:5e:00:02:02 cpm                        Intf     07/13/17 12:20:16
203       00:ca:fe:00:23:04 cpm                        Intf     07/13/17 12:20:16
-------------------------------------------------------------------------------
No. of MAC Entries: 4
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-4#
```

EVPN-MPLS R-VPLS with Single-active MH

Figure 51 shows the example topology with single-active multi-homing ES "ESI-23". The difference is that the ES is single-active and SDPs are used instead of a LAG.

Figure 51 **EVPN-MPLS R-VPLS with Single-active Multi-Homing**



The configuration is modified as follows:

- LAG 1 is removed from MTU-1, PE-2, and PE-3.
- Network interfaces are configured between MTU-1 and PE-2/PE-3 with IS-IS and LDP enabled.
- SDPs are configured.
- Ethernet segment "ESI-23" is redefined as single-active multi-homing. The SDP is associated with this ES.
- VPLS 202 on PE-2 and PE-3 no longer has a SAP, but a spoke-SDP instead.
- No changes are required on VPRN 20 or VPLS 200.

The service configuration on PE-2 is as follows. The configuration on PE-3 is similar. No changes are required on PE-4.

```
*A:PE-2# configure service
```

```
*A:PE-2>config>service# info
----------------------------------------------
        system
            bgp-evpn
                ethernet-segment "ESI-23"  create
                    esi 01:00:00:00:00:23:00:00:00:01
                    es-activation-timer 3
                    service-carving
                        mode auto
                    exit
                    multi-homing single-active
                    sdp 21
                    no shutdown
                exit
            exit
        exit
---snip---
        sdp 21 mpls create
            far-end 192.0.2.1
            ldp
            keep-alive
                shutdown
            exit
            no shutdown
        exit
---snip---
        vprn 20 customer 1 create
            route-distinguisher 192.0.2.2:20
            vrf-target target:64500:20
            interface "int-evi-202" create
                address 172.16.20.2/24
                mac 00:ca:fe:00:02:02
                vrrp 1 passive
                    backup 172.16.20.254
                    ping-reply
                    traceroute-reply
                exit
                ipv6
                    address 2001:db8:16::20:2/120
                    link-local-address fe80::16:20:2 dad-disable
                    vrrp 1 passive
                        backup fe80::16:20:fe
                        ping-reply
                        traceroute-reply
                    exit
                exit
                vpls "evi-202"
                exit
            exit
            interface "int-evi-200" create
                ipv6
                exit
                vpls "evi-200"
                    evpn-tunnel
                exit
            exit
            router-advertisement
                interface "int-evi-202"
                    use-virtual-mac
                    no shutdown
                exit
            exit
            no shutdown
        exit
        vpls 200 customer 1 create
            allow-ip-int-bind
            exit
            bgp
            exit
            bgp-evpn
                ip-route-advertisement
                evi 200
                vxlan
                    shutdown
                exit
                mpls
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            stp
                shutdown
            exit
            service-name "evi-200"
            no shutdown
        exit
        vpls 202 customer 1 create
            allow-ip-int-bind
            exit
            bgp
            exit
            bgp-evpn
                evi 202
                vxlan
                    shutdown
                exit
                mpls
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            stp
                shutdown
            exit
            service-name "evi-202"
            spoke-sdp 21:20 create
                no shutdown
            exit
            no shutdown
        exit
```

PE-2 is the Designated Forwarder (DF) in the single-active ES, as shown in the following output:

```
*A:PE-2# show service id 202 ethernet-segment
No sap entries
===============================================================================
SDP Ethernet-Segment Information
===============================================================================
SDP                  Eth-Seg                          Status
-------------------------------------------------------------------------------
21:20                ESI-23                           DF
===============================================================================
*A:PE-2#
```

```
*A:PE-3# show service id 202 ethernet-segment
No sap entries
===============================================================================
SDP Ethernet-Segment Information
===============================================================================
SDP                  Eth-Seg                          Status
-------------------------------------------------------------------------------
31:20                ESI-23                           NDF
===============================================================================
*A:PE-3#
```

When traffic has been sent between CE-11 and CE-41, the FDB on PE-2 is as follows. MAC address 00:00:01:00:00:11 corresponds to CE-11 and has been learned on spoke-SDP 21:20; MAC address 00:00:04:00:00:41 corresponds to CE-41 and has been advertised by PE-4 in an EVPN-MAC route.

```
*A:PE-2# show service id 202 fdb detail

===============================================================================
Forwarding Database, Service 202
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
202       00:00:01:00:00:11 sdp:21:20                L/0      07/13/17 12:24:05
202       00:00:01:00:00:16 sdp:21:20                L/0      07/13/17 12:24:10
202       00:00:04:00:00:41 eMpls:                   Evpn     07/13/17 12:20:38
                            192.0.2.4:262136
202       00:00:5e:00:01:01 cpm                      Intf     07/13/17 12:20:19
202       00:00:5e:00:02:01 cpm                      Intf     07/13/17 12:20:19
202       00:ca:fe:00:02:02 cpm                      Intf     07/13/17 12:20:18
202       00:ca:fe:00:02:03 eMpls:                   EvpnS    07/13/17 12:20:26
                                                     P
                            192.0.2.3:262133
202       00:ca:fe:00:02:04 eMpls:                   EvpnS    07/13/17 12:20:37
                                                     P
                            192.0.2.4:262136
-------------------------------------------------------------------------------
No. of MAC Entries: 8
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-2#
```

When the SDP between MTU-1 and DF PE-2 goes down, traffic from CE-41 to CE-11 is forwarded by PE-4 to DF PE-2. PE-2 cannot forward the packets to CE-11 directly, and will forward the packets to its ES peer PE-3. PE-3 will forward to CE-11 even if the MAC SA matches its own vMAC. Virtual MACs bypass the R-VPLS interface protection, so traffic can be forwarded between the PEs without being dropped.

Conclusion

EVPN can be used as the unified control plane VPN technology, not only for providing Layer 2 connectivity, but also Layer 3 (inter-subnet forwarding). EVPN for MPLS tunnels, along with multi-homing and Passive VRRP, provides efficient layer-2/layer-3 connectivity to distributed hosts and routers.

EVPN for PBB over MPLS (PBB-EVPN)

This chapter provides information about EVPN for PBB over MPLS (PBB-EVPN).

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SR OS Release 13.0.R6. The CLI in the current edition is based on SR OS Release 15.0.R3.

Important note: A prerequisite is to read the EVPN for MPLS Tunnels chapter.

Overview

EVPN for Provider Backbone Bridging (PBB) over MPLS (hereafter called PBB-EVPN) is specified in RFC 7623, *Provider Backbone Bridging Combined with Ethernet VPN (PBB-EVPN)*. It provides a simplified version of EVPN-MPLS for cases where the network requires very high scalability and does not need all the advanced features supported by EVPN-MPLS (but still requires single-active and all-active multi-homing capabilities). Table 5 provides a comparison between the capabilities of EVPN and PBB-EVPN in SR OS, and may help to choose between them when designing a VPN service.

Table 5 **EVPN and PBB-EVPN SR OS Feature Comparison**

VPN requirements	EVPN	PBB-EVPN	Comments
All-active Multi-Homing (MH) (flow-based load-balancing)	Yes	Yes	Allows better bandwidth utilization
Single-active MH (service-based load-balancing)	Yes	Yes	

Table 5 **EVPN and PBB-EVPN SR OS Feature Comparison (Continued)**

VPN requirements	EVPN	PBB-EVPN	Comments
Ethernet Local Area Network (E-LAN) and point-to-point E-Line services	Yes	Yes	
Inter-subnet-forwarding	Yes	No	Allows combined Layer 2 / Layer 3 services. EVPN
Proxy-Address Resolution Protocol / Neighbor Discovery (Proxy-ARP/ND) and IP-duplication protection	Yes	No	Allows Broadcast, Unknown unicast and Multicast (BUM) traffic reduction and better security
Customer MAC (CMAC) protection	Yes	No	Allows protecting key static CMACs
Data Center integration	Yes	No	Integration with VXLAN and Nuage Virtualized Services Directory (VSD)
Control plane overhead	Medium	Low	PBB-EVPN only advertises Backbone MACs (BMACs) and no route type 1s
Confinement of CMAC learning	No	Yes	CMACs are only learned on PEs with flows using those CMACs
CMAC summarization	No	Yes	Aggregation of CMACs into BMACs

PBB-EVPN is a combination of 802.1ah PBB and RFC7432, *BGP MPLS-Based Ethernet VPN* (EVPN-MPLS), and reuses the PBB-Virtual Private LAN Service (VPLS) service model, where Border Gateway Protocol BGP-EVPN is enabled in the backbone VPLS (B-VPLS) domain. EVPN is used as the control plane in the B-VPLS domain to control the distribution of BMACs and set up per-backbone service instance identifier (ISID) flooding trees for service instance VPLS (I-VPLS) services. The learning of the CMACs, either on local SAPs/SDP-bindings or associated with remote BMACs, is still performed in the data plane. Only the learning of BMACs in the B-VPLS is performed through BGP.

The SR OS PBB-EVPN implementation supports I-VPLS and PBB-Epipe services, including single-active and all-active multi-homing.

Because PBB-EVPN is based on the same control plane model as EVPN for MPLS, it is recommended to read the EVPN for MPLS Tunnels chapter before configuring PBB-EVPN. PBB-EVPN uses a subset of the BGP-EVPN routes described in EVPN for MPLS Tunnels as shown in Figure 52.

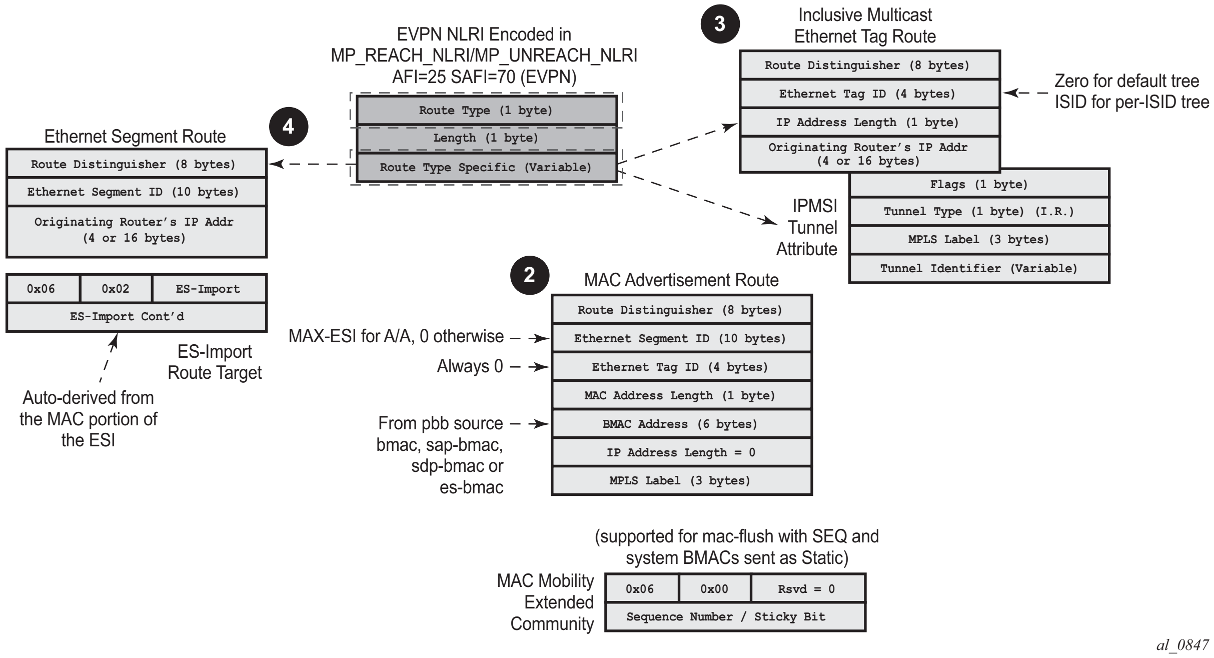

EVPN NLRI Encoded in
MP_REACH_NLRI/MP_UNREACH_NLRI
AFI=25 SAFI=70 (EVPN)
Route Type (1 byte)
Length (1 byte)
Route Type Specific (Variable)
4
Ethernet Segment Route
Route Distinguisher (8 bytes)
Ethernet Segment ID (10 bytes)
Originating Router's IP Addr (4 or 16 bytes)
0x06
0x02
ES-Import
ES-Import Cont'd
ES-Import Route Target
Auto-derived from the MAC portion of the ESI
3
Inclusive Multicast Ethernet Tag Route
Route Distinguisher (8 bytes)
Ethernet Tag ID (4 bytes)
IP Address Length (1 byte)
Originating Router's IP Addr (4 or 16 bytes)
Zero for default tree
ISID for per-ISID tree
IPMSI Tunnel Attribute
Flags (1 byte)
Tunnel Type (1 byte) (I.R.)
MPLS Label (3 bytes)
Tunnel Identifier (Variable)
2
MAC Advertisement Route
Route Distinguisher (8 bytes)
Ethernet Segment ID (10 bytes)
Ethernet Tag ID (4 bytes)
MAC Address Length (1 byte)
BMAC Address (6 bytes)
IP Address Length = 0
MPLS Label (3 bytes)
MAX-ESI for A/A, 0 otherwise
Always 0
From pbb source bmac, sap-bmac, sdp-bmac or es-bmac
(supported for mac-flush with SEQ and system BMACs sent as Static)
MAC Mobility Extended Community
0x06
0x00
Rsvd = 0
Sequence Number / Sticky Bit
al_0847

Figure 52 **EVPN Route Types**

When no EVPN multi-homing is used in the network, only the base routes are used. Route types 2 and 3 are considered the base and mandatory routes:

- Route type 2 — (B) MAC route — In PBB-EVPN, this route type is used for the advertisement of BMACs that will be installed in the remote Forwarding Data Bases (FDBs). There are no IP addresses advertised in PBB-EVPN. The MAC mobility extended community is used for advertising system BMACs as **protected** (with the sticky bit set) and it is also used for CMAC flush in some single-homing scenarios that will be described later.
- Route type 3 — Inclusive Multicast route — This route type is used for the advertisement of the I-VPLS ISIDs (no Epipes) and the desired multicast tree for each of them. The ISIDs are encoded in the Ethernet-tag field of the Network Layer Reachability Information (NLRI). When the B-VPLS is created with **no shutdown**, an Inclusive Multicast route with ISID = 0 is advertised. This is for the creation of the default multicast tree.

When EVPN multi-homing is used in an ISID, route type 4 (Ethernet Segment (ES) route) is used. In PBB-EVPN, there is no route type 1 advertised when multi-homing is used on the ISID services (I-VPLS and Epipes). Only route type 4 is used, and in the same way as it is for EVPN-MPLS. See the EVPN for MPLS Tunnels example for more information about ES routes, how they are formed, and how their RT/RD values are populated.

Configuration

This example describes the basic PBB-EVPN configuration first (without multi-homing) and how the flood containment is handled in PBB-EVPN. Flood containment refers to the efficient distribution of the BUM traffic generated for an ISID.

Networks are not always greenfield, so a smooth migration of PBB-EVPN from PBB-VPLS is required to minimize the effect on existing services. This example also describes this migration, starting from a common PBB-VPLS configuration.

Finally, this example describes the configuration of PBB-EVPN multi-homing.

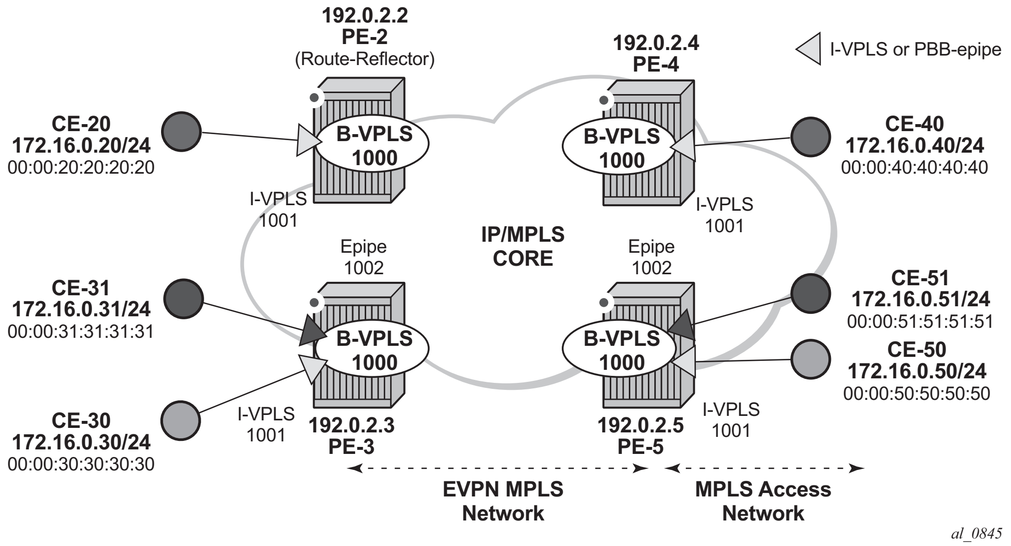
The same setup described in the VPN for MPLS tunnels example is used:

- Four PEs in the core (PE-2, PE-3, PE-4, and PE-5).
- The PEs are interconnected in the same way as explained in EVPN for MPLS Tunnels with the same IP addressing, IS-IS, transport LDP, and BGP peering configuration. There is not any difference with the basic infrastructure. See the EVPN for MPLS Tunnels chapter if more information is required.
- When configuring multi-homing, MTU-1 and MTU-6 are connected to the core.

PBB-EVPN Configuration without Multi-Homing

Figure 53 shows the example topology used in this chapter.

Figure 53 **PBB-EVPN Network without Multi-Homing**



When configuring PBB-EVPN:

- There is no difference at the access side (I-VPLS and Epipe configuration) compared to other PBB technologies supported in SR OS, such as Shortest Path Bridging for MAC (SPBM) or PBB-VPLS.
- The B-VPLS becomes an EVPN-MPLS service, where bgp-evpn mpls is added.

The following output shows an example of a basic configuration in PE-3. B-VPLS 1000 is bgp-evpn enabled and I-VPLS 1001 and Epipe 1002 are linked to B-VPLS 1000.

```
configure
    service
        vpls 1000 customer 1 b-vpls create
            service-mtu 2000
            pbb
                source-bmac 00:00:00:00:00:03
            exit
            bgp
            exit
            bgp-evpn
                evi 1000
                mpls
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            no shutdown
        exit
        vpls 1001 customer 1 i-vpls create
            pbb
                backbone-vpls 1000
                exit
            exit
            sap 1/2/1:1001 create
            exit
            no shutdown
        exit
        epipe 1002 customer 1 create
            pbb
                tunnel 1000 backbone-dest-mac 00:00:00:00:00:05 isid 1002
            exit
            sap 1/2/1:1002 create
            exit
            no shutdown
        exit
```

In the preceding output, there is no new configuration needed for I-VPLS/Epipe services. As for the B-VPLS, the output shows the minimum configuration required. If needed, the following parameters can be modified under **bgp-evpn**:

```
*A:PE-2# configure service vpls 1000 bgp-evpn
  - bgp-evpn
  - no bgp-evpn

 [no] accept-ivpls-e* - Configure to accept non-zero ethernet-tag MAC routes and
                        process for CMAC flushing
 [no] cfm-mac-advert* - Enable/disable the advertisement of MEP, MIP, and VMEP MAC
                        addresses over the BGP EVPN
 [no] evi             - EVPN Identifier
 [no] incl-mcast-ori* - Configure original IP address
 [no] ingress-repl-i* - Configure BGP EVPN IMET-IR route advertisement
 [no] ip-route-adver* - Configure BGP EVPN IP Route Advertisement
      isid-route-tar* + configure ISID route target information
 [no] mac-advertisem* - Configure BGP EVPN MAC Advertisement
      mac-duplication + Configure BGP EVPN MAC Duplication
      mpls            + Configure BGP EVPN mpls
 [no] unknown-mac-ro* - Configure BGP EVPN Unknown MAC Route
      vxlan           + Configure BGP EVPN vxlan


*A:PE-2# configure service vpls 1000 bgp-evpn mpls
  - mpls

      auto-bind-tunn* + Configure BGP EVPN mpls auto-bind-tunnel
      bgp-instance    - Configure BGP instance
 [no] control-word    - Enable/disable setting the CW bit in the label message
      ecmp            - Configure maximum ECMP routes information
 [no] entropy-label   - Enable/disable use of entropy-label
 [no] force-vlan-vc-* - Forces vlan-vc-type forwarding in the data-path
 [no] ingress-replic* - Use the same label as the one advertised for unicast traffic
 [no] restrict-prote* - Enable/disable protected src MAC restriction
 [no] send-evpn-encap - Configure encapsulation for this service
 [no] shutdown        - Administratively Enable/Disable BGP-EVPN mpls
 [no] split-horizon-* - Configure a split-horizon-group
```

A detailed description of these commands is included in the EVPN for MPLS Tunnels chapter. In addition to the preceding commands, the following **service>(b-)vpls>pbb** commands are relevant for PBB-EVPN in the B-VPLS service:

- **force-qtag-forwarding** allows the transparent transport of the customer 802.1p bits across the B-VPLS services.
- **source-bmac** can modify the source BMAC for all the PBB packets containing traffic from non-multi-homed I-VPLS and Epipe services.
- **use-es-bmac** instructs the system to use an ES-specific BMAC for traffic coming from an ES on an I-VPLS or Epipe.
- **use-sap-bmac** instructs the system to use a SAP-specific BMAC for traffic coming from an MC-LAG I-VPLS/Epipe SAP.

Flood Containment for I-VPLS Services

In general, PBB technologies in SR OS support a way to contain flooding for a specified I-VPLS ISID, so that BUM traffic for that ISID only reaches the PEs where the ISID is locally defined. Each PE creates a Multicast Forwarding Information Base (MFIB) per I-VPLS ISID on the B-VPLS instance. That MFIB supports SAP/SDP-binding endpoints that can be populated by:

- Multiple MAC Registration Protocol (MMRP) in regular PBB-VPLS
- IS-IS in SPBM

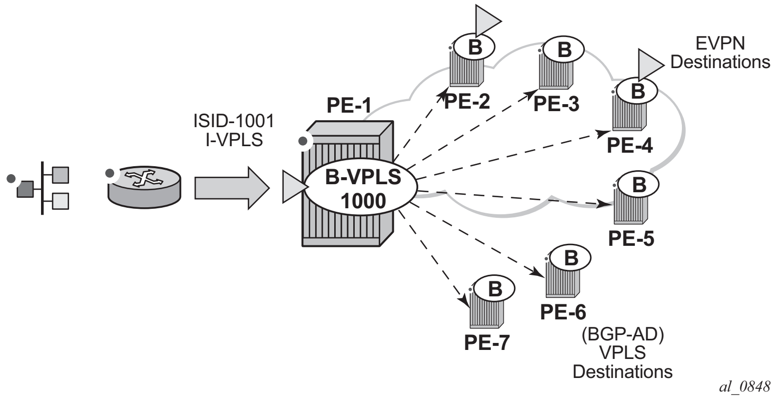
In PBB-EVPN, B-VPLS EVPN destinations can be added to the MFIBs using EVPN Inclusive Multicast Ethernet tag routes when they include the ISID in the Ethernet-tag. By default, when a B-VPLS is successfully enabled (**no shutdown**), the PE advertises:

- An Inclusive Multicast route for ISID = 0 — This allows the remote PEs to add the advertising PE to the default-multicast-list for the B-VPLS.
- An Inclusive Multicast route for each local ISID defined in the system (a local ISID includes configured I-VPLS and static-ISIDs) — This allows the remote PEs to create MFIB entries in the B-VPLS for the received ISIDs.

Because EVPN destinations, B-SAPs, and B-spoke-SDPs can coexist in the same B-VPLS, be aware of the different flooding lists created and how they are used in a B-VPLS. Figure 54 illustrates this concept with an example for B-VPLS 1000 in PE-1. The assumptions are:

- I-VPLS 1001 is created in PE-1, PE-2, and PE-4 only.
- PE-1, PE-2, PE-3, PE-4, and PE-5 support BGP-EVPN in B-VPLS 1000.
- PE-6 and PE-7 only support spoke-SDPs.
- PE-1 is connected to all six PEs.

Figure 54 **PBB-EVPN — Flooding Lists**



In this situation, PE-1 creates two flooding lists in B-VPLS 1000:

- Default-multicast-list — composed of:
  - All the EVPN PEs that advertised ISID = 0 (PE-2, PE-3, PE-4, PE-5).
  - All the B-spoke-SDPs (or B-SAPs) (PE-6, PE-7).
  - All the EVPN PEs that advertised ISID 1001 and no ISID 0 (if an isid-policy is created in PE-1 stating **use-def-mcast** for ISID 1001). Note: third-party PEs may not advertise ISID = 0, but only non-zero ISIDs.
- MFIB for ISID 1001 is composed of:
  - All the EVPN PEs that advertised ISID 1001 (PE-2 and PE-4) unless there is an ISID-policy in PE-1 stating **use-def-mcast** for ISID 1001.
  - Static-ISIDs defined in manual B-spoke-SDPs and B-SAPs (static-ISIDs cannot be created on BGP-AD auto-discovered B-spoke-SDPs).

Based on the above, when BUM traffic is sent to I-VPLS 1001 on PE-1:

- The traffic is encapsulated in PBB with the group BMAC for ISID 1001 and sent (by default) to the MFIB created for ISID 1001 (PE-2 and PE-4).
- If an ISID-policy is added with **use-def-mcast** for ISID 1001, the BUM traffic is encapsulated in PBB with the group BMAC for ISID 1001 and sent to the default-multicast-list, that is, all six remote PEs.

Referring to Figure 53, the following output illustrates the use of the ISID-policy in PBB-EVPN. PE-2 does not have any ISID-policy configured; when it receives BUM traffic from the local I-VPLS 1001, it uses the MFIB for ISID 1001:

```
configure
    service
        vpls 1000 customer 1 b-vpls create
            service-mtu 2000
            pbb
```

```
                source-bmac 00:00:00:00:00:02
            exit
            bgp
            exit
            bgp-evpn
                evi 1000
                mpls
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            no shutdown
```

```
*A:PE-2# show service id 1000 mfib

===============================================================================
Multicast FIB, Service 1000
===============================================================================
Source Address  Group Address         Sap/Sdp Id                   Svc Id   Fwd
                                                                            Blk
-------------------------------------------------------------------------------
*                01:1e:83:00:03:e9    b-eMpls:192.0.2.3:262134     Local    Fwd
                                      b-eMpls:192.0.2.4:262130     Local    Fwd
                                      b-eMpls:192.0.2.5:262132     Local    Fwd
-------------------------------------------------------------------------------
Number of entries: 1
===============================================================================
```

An ISID-policy can be added to modify this behavior and allow PE-2 to use the default-multicast-list. If I-VPLS 1001 exists in all the remote PEs (as in this example), using the default multicast list is as efficient as using the MFIB and saves expensive MFIB resources. In the following output, as soon as the ISID-policy is added, the MFIB entries for ISID 1001 are removed and PE-2 starts using the default multicast list. The **tools dump service id 1000 evpn-mpls default-multicast-list** command shows the EVPN destinations that are part of the default-multicast list:

```
configure
    service
        vpls 1000
            isid-policy
                entry 10 create
                    use-def-mcast
                    range 1001 to 2000
                exit
            exit
```

```
*A:PE-2# tools dump service id 1000 evpn-mpls default-multicast-list
-------------------------------------------------------------------------
TEP Address                              Egr Label
                                         Transport
-------------------------------------------------------------------------
192.0.2.3                                262134
                                         ldp
```

```
192.0.2.4                                 262130
                                          ldp
192.0.2.5                                 262132
                                          ldp
-------------------------------------------------------------------------
```

The MFIB on PE-2 does not contain any entries for ISID 1001 anymore, as follows:

```
*A:PE-2# show service id 1000 mfib

===============================================================================
Multicast FIB, Service 1000
===============================================================================
Source Address  Group Address         Sap/Sdp Id                   Svc Id    Fwd
                                                                             Blk
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Number of entries: 0
===============================================================================
```

PBB-VPLS to PBB-EVPN Migration

The principles required for migrating a PBB-VPLS network to PBB-EVPN are explained in the **VPLS to EVPN-MPLS Integration** section of the EVPN for MPLS Tunnels chapter. Those principles are also applicable to EVPN destinations and spoke-SDPs in the B-VPLS and can be summarized in three points:

- Systems with an EVPN destination and SDP-binding to the same far-end IP bring down the SDP-binding. This avoids loops when both constructs exist in the same network.
- SDP-bindings and EVPN destinations can be placed in the same Split Horizon Group (SHG). When traffic from an SDP-binding/EVPN destination belonging to that SHG is received on a PE, it is never forwarded to another SDP-binding/EVPN destination on the same SHG.
- MAC addresses learned on an SDP-binding or SAP, that belong to an SHG where EVPN destinations are also created, are not advertised in BGP-EVPN.

Based on those principles, this section describes how to migrate a PBB-VPLS network to PBB-EVPN. The network in Figure 53 represents a regular PBB-VPLS network that needs to be migrated to PBB-EVPN.

In that network, the four PEs are running BGP-AD and TLDP for the discovery and setup of the pseudowires in the B-VPLS instance. The advantage of this configuration is that the migration can be done node by node and with minimum impact on customer service.

Initial Configuration

Initially, the network is configured for PBB-VPLS with BGP-AD in B-VPLS 1000. The EVPN family is to be added. At the access, I-VPLS 1001 is connected to the CEs. As an example, the configuration in PE-3 is shown. An equivalent configuration exists in the other three PEs.

Note: The EVPN family is added to the BGP configuration because PBB-EVPN uses this address family. Assuming there are redundant Route Reflectors (RRs), the addition of EVPN can be done without service impact. In this example, the assumption is that the PEs are already configured with the EVPN family.

```
configure
    router
        bgp
            vpn-apply-import
            vpn-apply-export
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            split-horizon
            rapid-update evpn
            group "internal"
                family l2-vpn evpn
                peer-as 64500
                neighbor 192.0.2.2
                exit
            exit
```

```
configure
    service
        pw-template 1 create
            split-horizon-group "CORE"
            exit
        exit
        vpls 1000 customer 1 b-vpls create
            service-mtu 2000
            pbb
                source-bmac 00:00:00:00:00:03
            exit
            bgp
                pw-template-binding 1
                exit
            exit
            bgp-ad
                vpls-id 64500:1000
                no shutdown
            exit
            no shutdown
        exit
        vpls 1001 customer 1 i-vpls create
            pbb
                backbone-vpls 1000
                exit
            exit
            sap 1/2/1:1001 create
            exit
            no shutdown
        exit

*A:PE-3# show service id 1000 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 1000                Vpn Id            : 0
Service Type      : b-VPLS

---snip---

Oper Backbone Src : 00:00:00:00:00:03
Use SAP B-MAC     : Disabled
i-Vpls Count      : 1
Epipe Count       : 0
Use ESI B-MAC     : Disabled

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sdp:17405:4294967293 SB(192.0.2.5)       BgpAd        0       8974    Up   Up
sdp:17406:4294967294 SB(192.0.2.4)       BgpAd        0       8974    Up   Up
sdp:17407:4294967295 SB(192.0.2.2)       BgpAd        0       8974    Up   Up
===============================================================================
* indicates that the corresponding row element may have been truncated.
```

Multiple MAC Registration Protocol (MMRP) is not used in the B-VPLS instance. If it were enabled, MMRP would have to be disabled in the network before this migration. If there are ISIDs using B-VPLS SDP-bindings to reach some remote locations and B-VPLS EVPN destinations to reach others, the default multicast list must be used in the current release (the MFIB cannot be used if there is a mix of both types). Therefore, during the migration process, the ISIDs must be added to the default-multicast-list.

Step 1. Add service-level SHG (if not already there).

From the first node being migrated to PBB-EVPN to all nodes migrated, PBB-VPLS and PBB-EVPN have to coexist within the same meshed network. That is, EVPN-MPLS destinations and SDP-bindings need to be defined in the same split-horizon-group. Therefore, if there is no split-horizon-group defined in the B-VPLS, the first step is to add it. In this example, the **split-horizon-group** is defined at the **config>service>pw-template>level**; therefore, it has to be added at the B-VPLS level.

Note: When the **service>split-horizon-group** is removed, an eval-pw-template must be performed.

Note: After adding the **split-horizon-group** at the service level, an eval-pw-template must be performed again so that the SDP-bindings take the new SHG configuration.

Note: During the time between the **split-horizon-group** being removed and added back again, the SDP-bindings can forward BUM traffic to each other, so this operation must be done carefully to avoid loops.

Assuming that the first node to be migrated is PE-3, the following output shows the procedure for adding the **split-horizon-group** at the service level.

```
*A:PE-3# configure service pw-template 1 no split-horizon-group


*A:PE-3# tools perform service id 1000 eval-pw-template 1 allow-service-impact
eval-pw-template succeeded for Svc 1000 17405:4294967293 Policy 1
eval-pw-template succeeded for Svc 1000 17406:4294967294 Policy 1
eval-pw-template succeeded for Svc 1000 17407:4294967295 Policy 1


*A:PE-3# configure
    service
        vpls 1000
            split-horizon-group "CORE" create
            exit
            bgp
                pw-template-binding 1 split-horizon-group "CORE"
                exit
            exit


*A:PE-3# tools perform service id 1000 eval-pw-template 1 allow-service-impact
eval-pw-template succeeded for Svc 1000 17405:4294967293 Policy 1
eval-pw-template succeeded for Svc 1000 17406:4294967294 Policy 1
eval-pw-template succeeded for Svc 1000 17407:4294967295 Policy 1


*A:PE-3>config>service>vpls# info
----------------------------------------------
            service-mtu 2000
            pbb
                source-bmac 00:00:00:00:00:03
            exit
            split-horizon-group "CORE" create
            exit
            bgp
                pw-template-binding 1 split-horizon-group "CORE"
                exit
            exit
            bgp-ad
                vpls-id 64500:1000
                no shutdown
            exit
            stp
                shutdown
            exit
            no shutdown
```

Step 2. Add BGP-EVPN and ISID-policy configuration to the B-VPLS.

After the B-VPLS is configured with the split horizon group, the BGP-EVPN configuration and ISID-policy can be added (still in **shutdown**), as follows.

```
configure
    service
        vpls 1000
            bgp-evpn
                evi 1000
                mpls
                    split-horizon-group "CORE"
                    auto-bind-tunnel
                        resolution any
                    exit
                    shutdown
                exit
            exit
            isid-policy
                entry 10 create
                    use-def-mcast
                    range 1001 to 3000
                exit
            exit
```

Step 3. Enable BGP-EVPN MPLS on the PE.

When the configuration is ready, the **bgp-evpn mpls** context can be **no shutdown**.

```
*A:PE-3# configure service vpls 1000 bgp-evpn mpls no shutdown
```

The preceding **no shutdown** triggers a route-refresh message for the EVPN family from PE-3, but no changes happen because PE-3 does not create any EVPN destinations until it imports EVPN routes from the other PEs. The three spoke-SDPs to the remote PEs are still up.

Step 4. Repeat steps 1 to 3 for the second PE.

The same steps 1 to 3 are repeated for PE-5. When **bgp-evpn mpls no shutdown** is executed, PE-5 sends a route-refresh and gets the BGP-EVPN routes from PE-3. As a result of that, PE-3 brings down the spoke-SDP to PE-5 and creates an EVPN destination to PE-5. The same process happens in PE-5. The following CLI output shows the received routes in PE-3 and spoke-SDP going down.

```
3 2017/05/05 10:58:50.21 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 117
    Flag: 0x90 Type: 14 Len: 47 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.5
        Type: EVPN-Incl-mcast Len: 17 RD: 64500:1000, tag: 0, orig_addr len: 32,
                              orig_addr: 192.0.2.5
        Type: EVPN-Incl-mcast Len: 17 RD: 64500:1000, tag: 1001, orig_addr len: 32,
                              orig_addr: 192.0.2.5
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.5
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        1.1.1.1
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:64500:1000
        bgp-tunnel-encap:MPLS
    Flag: 0xc0 Type: 22 Len: 9 PMSI:
        Tunnel-type Ingress Replication (6)
        Flags [Type: None BM: 0 U: 0 Leaf not required]
        MPLS Label 4194112
        Tunnel-Endpoint 192.0.2.5
"
```

Log 99 shows that spoke SDP 17405:4294967293 is operationally down:

```
149 2017/05/05 10:58:50.20 UTC MINOR: SVCMGR #2326 Base
"Status of SDP Bind 17405:4294967293 in service 1000 (customer 1) local PW status bits
changed to pwNotForwarding "

150 2017/05/05 10:58:50.20 UTC MINOR: SVCMGR #2306 Base
"Status of SDP Bind 17405:4294967293 in service 1000 (customer 1) changed to admin=
up oper=down flags=evpnRouteConflict"

151 2017/05/05 10:58:52.22 UTC MINOR: SVCMGR #2313 Base
"Status of SDP Bind 17405:4294967293 in service 1000 (customer 1) peer PW status bits
changed to pwNotForwarding "
```

Spoke SDP 17405:4294967293 is the spoke SDP toward PE-5 and it is kept down:

```
*A:PE-3# show service id 1000 base
---snip---

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sdp:17405:4294967293 SB(192.0.2.5)       BgpAd        0       8974    Up   Down
sdp:17406:4294967294 SB(192.0.2.4)       BgpAd        0       8974    Up   Up
sdp:17407:4294967295 SB(192.0.2.2)       BgpAd        0       8974    Up   Up
```

```
===============================================================================
* indicates that the corresponding row element may have been truncated.
```

The reason why the spoke SDP toward PE-5 is down is an EVPN route conflict:

```
*A:PE-3# show service id 1000 sdp 17405:4294967293 detail | match Flag context all
Flags              : PWPeerFaultStatusBits
                     EvpnRouteConflict
```

An EVPN destination to PE-5 is created:

```
*A:PE-3# show service id 1000 evpn-mpls

===============================================================================
BGP EVPN-MPLS Dest
===============================================================================
TEP Address     Egr Label     Num. MACs   Mcast           Last Change
                 Transport
-------------------------------------------------------------------------------
192.0.2.5       262132        1           Yes             05/05/2017 10:58:50
                ldp
-------------------------------------------------------------------------------
Number of entries : 1
-------------------------------------------------------------------------------
===============================================================================
---snip---
```

Step 5. Repeat Steps 1 to 3 for the rest of the PEs.

The same process is repeated in all the PEs, node by node. The service impact for the I-VPLS 1001 is minimal.

Step 6. (Optional) Remove the ISID policy.

When all the PEs in the B-VPLS 1000 are migrated, the ISID-policy can optionally be removed, node by node. This forces the B-VPLS instance to start using the MFIB to send I-VPLS BUM traffic to the remote nodes. This has no effect on Epipes (traffic is always unicast for Epipes).

Before removing the ISID-policy and starting to use the MFIB, it is recommended to check that the Inclusive Multicast routes for an ISID to the remote PEs are all active. Otherwise, connectivity for BUM traffic could be interrupted if any of the expected routes are not active. This is illustrated for PE-3.

```
*A:PE-3# show service id 1000 evpn-mpls

===============================================================================
BGP EVPN-MPLS Dest
===============================================================================
TEP Address     Egr Label     Num. MACs   Mcast           Last Change
                 Transport
-------------------------------------------------------------------------------
192.0.2.2       262134        1           Yes             05/05/2017 12:06:43
                ldp
192.0.2.4       262130        1           Yes             05/05/2017 12:06:49
```

```
                ldp
192.0.2.5       262132        1           Yes             05/05/2017 10:58:50
                ldp
-------------------------------------------------------------------------------
Number of entries : 3
-------------------------------------------------------------------------------
===============================================================================
```

The routes for ISID 1001 are valid and used by BGP (flags **u*>i**):

```
*A:PE-3# show router bgp routes evpn inclusive-mcast tag 1001
===============================================================================
 BGP Router ID:192.0.2.3        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN Inclusive-Mcast Routes
===============================================================================
Flag  Route Dist.         OrigAddr
      Tag                 NextHop
-------------------------------------------------------------------------------
u*>i  64500:1000          192.0.2.5
      1001                192.0.2.5

u*>i  64500:1000          192.0.2.4
      1001                192.0.2.4

u*>i  64500:1000          192.0.2.2
      1001                192.0.2.2

-------------------------------------------------------------------------------
Routes : 3
===============================================================================
```

There are no entries in the MFIB:

```
*A:PE-3# show service id 1000 mfib

===============================================================================
Multicast FIB, Service 1000
===============================================================================
Source Address  Group Address         Sap/Sdp Id                   Svc Id   Fwd
                                                                            Blk
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Number of entries: 0
===============================================================================
```

The ISID policy is removed as follows:

```
configure service vpls 1000 isid-policy no entry 10
```

After removing the ISID-policy, the MFIB is populated with entries for the ISID 1001 group BMAC to the three remote PEs where ISID 1001 is defined:

```
*A:PE-3# show service id 1000 mfib

===============================================================================
Multicast FIB, Service 1000
===============================================================================
Source Address  Group Address         Sap/Sdp Id                  Svc Id   Fwd
                                                                           Blk
-------------------------------------------------------------------------------
*               01:1e:83:00:03:e9     b-eMpls:192.0.2.2:262134     Local    Fwd
                                      b-eMpls:192.0.2.4:262130     Local    Fwd
                                      b-eMpls:192.0.2.5:262132     Local    Fwd
-------------------------------------------------------------------------------
Number of entries: 1
===============================================================================
```

Step 7. (Optional) Remove the BGP-AD configuration.

The BGP-AD configuration can stay in the B-VPLS services. However, when the entire network is migrated to PBB-EVPN, all the spoke-SDPs will be operationally down and, even if they are not forwarding traffic, they consume resources in the system. Consider removing the BGP-AD configuration and, therefore, the spoke-SDPs.

The following example shows the removal of BGP-AD in PE-4. Be aware that when BGP-AD is removed from the configuration, if the RD/RT was derived from the VPLS-id (as in this example), a new RD/RT must be auto-derived for the service. Therefore, new updates will be sent for all the EVPN NLRIs, as shown in the following output.

```
*A:PE-4# show service id 1000 bgp

===============================================================================
BGP Information
===============================================================================
Vsi-Import           : None
Vsi-Export           : None
Route Dist           : None
Oper Route Dist      : 64500:1000
Oper RD Type         : derivedVpls
Rte-Target Import    : None                 Rte-Target Export: None
Oper RT Imp Origin   : derivedVpls          Oper RT Import   : 64500:1000
Oper RT Exp Origin   : derivedVpls          Oper RT Export   : 64500:1000
PW-Template Id       : 1                    PW-Template SHG  : CORE
Oper Group           : None
Mon Oper Group       : None
BFD Template         : None
BFD-Enabled          : no                   BFD-Encap        : ipv4
Import Rte-Tgt       : None
-------------------------------------------------------------------------------
===============================================================================
```

BGP-AD is disabled as follows:

```
*A:PE-4# configure service vpls 1000 bgp-ad shutdown
```

After BGP-AD is shutdown, the spoke SDP bindings are deleted.

```
175 2017/05/05 12:41:55.27 UTC MINOR: SVCMGR #2306 Base
"Status of SDP Bind 17407:4294967295 in service 1000 (customer 1) changed to admin=down
 oper=down flags=sdpBindAdminDown noIngressVcLabel noEgressVcLabel "

176 2017/05/05 12:41:55.27 UTC MAJOR: SVCMGR #2320 Base
"Service Id 1000, Dynamic bgp-l2vpn SDP Bind Id 17407:4294967295 was deleted."

177 2017/05/05 12:41:55.27 UTC MINOR: SVCMGR #2303 Base
"Status of SDP 17407 changed to admin=down oper=down"

178 2017/05/05 12:41:55.27 UTC MAJOR: SVCMGR #2319 Base
"Dynamic bgp-l2vpn SDP 17407 (192.0.2.2) was deleted."
```

The PW template binding is removed as follows:

```
*A:PE-4# configure service vpls 1000 bgp no pw-template-binding 1
```

The BGP-AD configuration is removed as follows:

```
*A:PE-4# configure service vpls 1000 no bgp-ad
```

Initially, the RD/RT was derived from the VPLS-id (64500:1000). After the BGP-AD configuration is removed, a new RD/RT must be auto-derived from the EVI:

```
*A:PE-4# show service id 1000 bgp

===============================================================================
BGP Information
===============================================================================
Vsi-Import           : None
Vsi-Export           : None
Route Dist           : None
Oper Route Dist      : 192.0.2.4:1000
Oper RD Type         : derivedEvi
Rte-Target Import    : None                 Rte-Target Export: None
Oper RT Imp Origin   : derivedEvi           Oper RT Import   : 64500:1000
Oper RT Exp Origin   : derivedEvi           Oper RT Export   : 64500:1000
PW-Template Id       : None
-------------------------------------------------------------------------------
===============================================================================
```

In this case, the system picks up the RD in the following order:

1. Manual RD or auto-RD always take precedence when configured.
2. If no manual/auto-RD, the RD is derived from the **bgp-ad vpls-id**.
3. If no manual/auto-rd/vpls-id configuration, the RD is derived from the **bgp evpn evi**.

4. If no manual/auto-rd/vpls-id/evi configuration, there will not be RD and the service will fail.

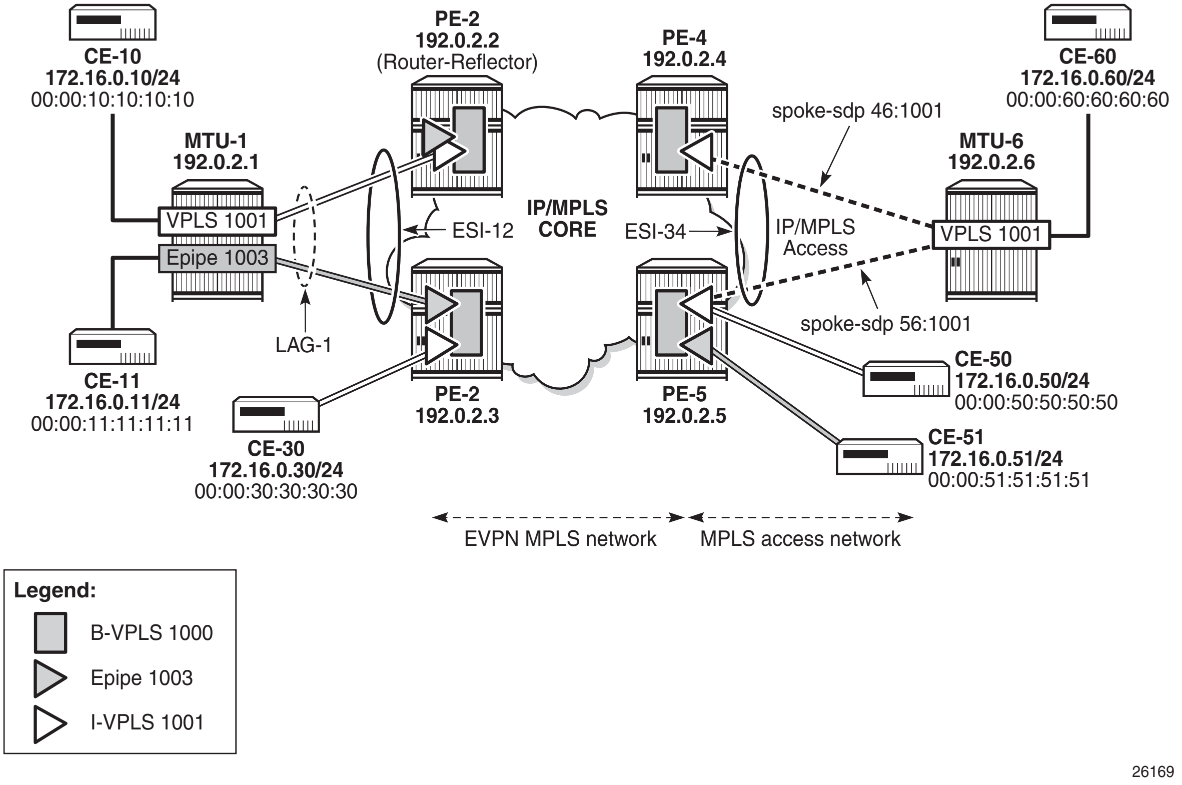
If in the migration from BGP-AD to BGP-EVPN, the advertisement of new updates is not needed, the initial configuration must include manual/auto-RDs. If manual/auto-RDs were not included, a **bgp-ad shutdown** would not cause the change of RD and the consequent BGP updates.

PBB-EVPN Multi-Homing

This section provides configuration guidelines for PBB-EVPN multi-homing. In the same way that EVPN-MPLS supports single-active and all-active multi-homing, PBB-EVPN can also be configured to support both modes. The same Ethernet-segment that is used for regular EVPN-MPLS service SAPs and spoke-SDPs can be shared with I-VPLS/Epipe SAPs and spoke-SDPs.

Figure 55 shows the example topology used in this section.

Figure 55 **PBB-EVPN Multi-homing**



MTU-1 and MTU-6 have been added to the network (compared to Figure 53). I-VPLS 1001 has two new sites that are multi-homed to the PBB-EVPN network. MTU-1 uses all-active multi-homing, whereas MTU-6 is connected to a single-active ES. As with EVPN-MPLS, all-active multi-homing is only supported when a LAG is used at the access. Single-active multi-homing can be supported with regular Ethernet ports (that can form an independent LAG per PE) or SDPs.

Draft-ietf-l2vpn-pbb-evpn describes two types of system BMAC assignments that a PE can implement in a B-VPLS when ESs are present:

- Shared BMAC addresses that can be used for all the single-homed CEs and a number of multi-homed CEs connected to Ethernet-segments.
- Dedicated BMAC addresses per Ethernet-segment.

In this chapter and in SR OS terminology:

- A shared-BMAC (in IETF) is a **source-bmac** as configured in **service>(b)vpls>pbb>source-bmac**. All the I-VPLS/Epipe traffic coming from single-homed CEs is sent encapsulated in a PBB packet with that **source-bmac**.
- A dedicated-BMAC per ES (in IETF) is an **es-bmac** as activated in **service>(b)vpls>pbb>use-es-bmac** and generated from the combination of **vpls>pbb>source-bmac** plus **ethernet-segment>source-bmac-lsb**. If configured, any I-VPLS/Epipe traffic coming from an ES is encapsulated in a PBB packet with the ES-BMAC as the source BMAC.

The system allows the following user choices per B-VPLS and ES:

- A dedicated **es-bmac** per ES can be used. In that case, the **pbb>use-es-bmac** command is configured in the B-VPLS. In all-active multi-homing, all the PEs that are part of the ES source the PBB packets with the same source **es-bmac**; single-active multi-homing requires the use of a different **es-bmac** per PE.
- A non-dedicated **source-bmac** can be used (this is only possible in single-active multi-homing). In this case, the user does not configure **pbb>use-es-bmac** and the regular **source-bmac** is used for the traffic. A different **source-bmac** has to be advertised per PE.

As discussed, single-active multi-homing can use **source-bmacs** or **es-bmacs**. Using one type or another has a different impact on CMAC flushing, as illustrated in Figure 56.

- If **es-bmacs** are used, as shown on the right-hand side of Figure 56, a less-impacting CMAC flush is achieved, therefore minimizing the flooding after ES failures. In the case of ES failure, PE-1 withdraws the **es-bmac** 00:12 and the remote PE-3 only flushes the CMACs associated with that **es-bmac** (only the CMACs behind the CE are flushed).

- If **source-bmacs** are used, as shown on the left-hand side of Figure 56, in the case of ES failure, a BGP update with higher sequence number is issued by PE-1 and the remote PE-3 flushes all the CMACs associated with the **source-bmac**. Therefore, all the CMACs behind the B-VPLS of the PEs will be flushed, as opposed to only the CMACs behind the CE of the Ethernet Service Instances (ESIs).

Figure 56 The Use of es-bmac to Minimize CMAC Flush

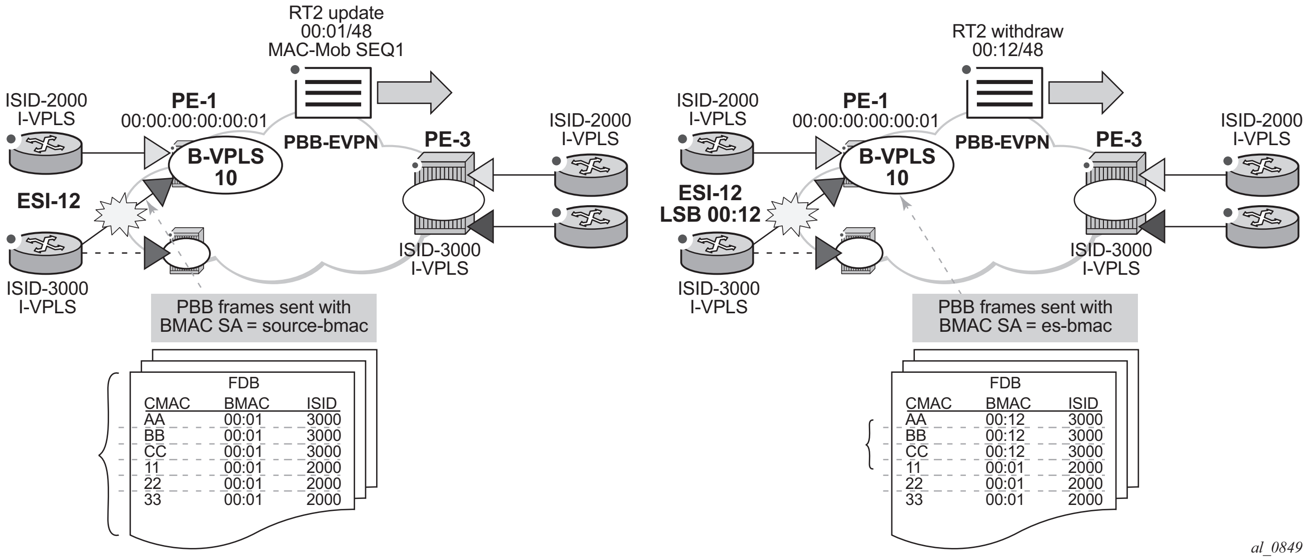


Table 6 shows the PBB-EVPN multi-homing combinations supported in the current release in the topology of Figure 55.

Table 6 PBB-EVPN Multi-Homing Supported Combinations in SR OS

CE Connectivity	PE Connectivity	PE Redundancy	BMAC Assignment	I-VPLS Support	Epipe Support
LAG (LACP optional)	LAG SAP	EVPN MH all-active	use-es-bmac (shared BMAC)	Yes	Yes
Ethernet ports (no LAG)	LAG SAP or port SAP	EVPN MH single-active	use-es-bmac (dedicated per PE)	Yes	No
Ethernet ports (no LAG)	LAG SAP or port SAP	EVPN MH single-active	source-bmac (dedicated per PE)	Yes	No
MPLS	spoke-SDP	EVPN MH single-active	source-bmac (dedicated per PE)	Yes	No
MPLS	spoke-SDP	EVPN MH single-active	use-es-bmac (dedicated per PE)	Yes	No

As an example, the configurations of the first, and last two, rows (LAG SAP all-active, MPLS source-BMAC, and MPLS ES-BMAC, respectively) will be discussed in the following three sections.

PBB-EVPN All-Active Multi-Homing for I-VPLS and Epipes

Figure 55 shows a PBB-EVPN network where ESI-12 is configured as an all-active multi-homing ES on PE-2 and PE-3. Two services are using ESI-12: I-VPLS 1001 and Epipe 1003. The following output shows the relevant configuration in PE-2:

```
configure
    service
        pbb
            mac-name "PE-5" 00:00:00:00:00:05
        exit
        system
            bgp-evpn
                ethernet-segment "ESI-12" create
                    esi 01:00:00:00:00:12:00:00:00:01
                    source-bmac-lsb 12-12 es-bmac-table-size 8
                    es-activation-timer 3
                    service-carving
                        mode auto
                    exit
                    multi-homing all-active
                    lag 1
                    no shutdown
                exit
            exit
        exit

        vpls 1000 customer 1 b-vpls create
            service-mtu 2000
            pbb
                source-bmac 00:00:00:00:00:02
                use-es-bmac
            exit
            split-horizon-group "CORE" create
            exit
            bgp
            exit
            bgp-evpn
                evi 1000
                mpls
                    split-horizon-group "CORE"
                    ecmp 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            no shutdown
        exit
        vpls 1001 customer 1 i-vpls create
            pbb
                backbone-vpls 1000
                exit
            exit
            sap lag-1:1001 create
            exit
            no shutdown
        exit
        epipe 1003 customer 1 create
            pbb
                tunnel 1000 backbone-dest-mac "PE-5" isid 1003
            exit
            sap lag-1:1003 create
            exit
            no shutdown
        exit
```

The following output shows the relevant configuration in PE-3:

```
configure
    service
        pbb
            mac-name "PE-5" 00:00:00:00:00:05
        exit
        system
            bgp-evpn
                ethernet-segment "ESI-12" create
                    esi 01:00:00:00:00:12:00:00:00:01
                    source-bmac-lsb 12-12 es-bmac-table-size 8
                    es-activation-timer 3
                    service-carving
                        mode auto
                    exit
                    multi-homing all-active
                    lag 1
                    no shutdown
                exit
            exit
        exit

        vpls 1000 customer 1 b-vpls create
            service-mtu 2000
            pbb
                source-bmac 00:00:00:00:00:03
                use-es-bmac
            exit
            split-horizon-group "CORE" create
            exit
            bgp
            exit
            bgp-evpn
                evi 1000
                mpls
                    split-horizon-group "CORE"
                    ecmp 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            no shutdown
        exit
        vpls 1001 customer 1 i-vpls create
            pbb
                backbone-vpls 1000
                exit
            exit
            sap 1/2/1:1001 create
            exit
            sap lag-1:1001 create
            exit
            no shutdown
        exit
        epipe 1003 customer 1 create
            pbb
                tunnel 1000 backbone-dest-mac "PE-5" isid 1003
            exit
            sap lag-1:1003 create
            exit
            no shutdown
        exit
```

The preceding configuration shows that Epipe 1003 has a PBB tunnel pointing at the PE-5 source-BMAC. Epipe 1003 has the following configuration in PE-5 (the PBB tunnel points at the ESI-12 ES-BMAC):

```
configure
    service
        pbb
            mac-name "ES-MAC-12" 00:00:00:00:12:12
        exit
        epipe 1003 customer 1 create
            pbb
                tunnel 1000 backbone-dest-mac "ES-MAC-12" isid 1003
            exit
            sap 1/2/1:1003 create
            exit
            no shutdown
        exit
```

Source-BMACs and ES-BMACs are distributed in BGP-EVPN. PE-2 and PE-3 will each advertise their own source-BMAC in a MAC route with ESI-0 and the shared ES-BMAC with ESI-MAX (as per the RFC 7623). The ES-BMAC that each PE uses in a B-VPLS is derived from the configured **service>(b)vpls>pbb>source-bmac** (four high-order bytes) and the ESI-12 configured **source-bmac-lsb**. In this example, PE-2 and PE-3 will both derive ES-BMAC 00:00:00:00:12:12. For both PEs to derive the required same ES-BMAC, the four high-order bytes of the source-BMAC must match on both PEs.

The **es-bmac-table-size** parameter modifies the default value (8) for the maximum number of ES-BMACs that can be associated with the Ethernet-segment across different B-VPLS services. When **source-bmac-lsb** is configured, the associated **es-bmac-table-size** is reserved out of the total FDB space.

The following outputs show the source-BMACs and ES-BMAC and how they are advertised and installed in the B-VPLS FDB.

```
*A:PE-2# show service system bgp-evpn ethernet-segment name "ESI-12" | match BMAC
Source BMAC LSB         : 12-12
```

The following output shows that ES-BMAC is used and that the operational source-BMAC is 00:00:00:00:00:02.

```
*A:PE-2# show service id 1000 base

================================================================================
Service Basic Information
================================================================================
Service Id        : 1000                Vpn Id            : 0
Service Type      : b-VPLS
---snip---
Oper Backbone Src : 00:00:00:00:00:02
Use SAP B-MAC     : Disabled
i-Vpls Count      : 1
Epipe Count       : 1
Use ESI B-MAC     : Enabled
---snip---
```

The source BMAC LSB is configured with the same value on PE-2 and PE-3. The two low-order bytes of the ES-BMAC will be 12:12.

```
*A:PE-3# show service system bgp-evpn ethernet-segment name "ESI-12" | match BMAC
Source BMAC LSB         : 12-12
```

On PE-3, ES-BMAC is used and the operational source BMAC is 00:00:00:00:00:03, as follows:

```
*A:PE-3# show service id 1000 base

================================================================================
Service Basic Information
================================================================================
Service Id        : 1000                Vpn Id            : 0
Service Type      : b-VPLS

--snipped--

Oper Backbone Src : 00:00:00:00:00:03
Use SAP B-MAC     : Disabled
i-Vpls Count      : 1
Epipe Count       : 2
Use ESI B-MAC     : Enabled
```

On PE-2, the FDB for B-VPLS 1000 has an entry for each of the other PEs. PEs do not show their own system BMACs in the FDB:

```
*A:PE-2# show service id 1000 fdb detail

===============================================================================
Forwarding Database, Service 1000
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1000      00:00:00:00:00:03 eMpls:                   EvpnS    05/05/17 12:49:42
                                                     P
                            192.0.2.3:262136
1000      00:00:00:00:00:04 eMpls:                   EvpnS    05/05/17 12:49:41
                                                     P
                            192.0.2.4:262133
1000      00:00:00:00:00:05 eMpls:                   EvpnS    05/05/17 12:49:37
                                                     P
                            192.0.2.5:262142
-------------------------------------------------------------------------------
No. of MAC Entries: 3
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
```

On PE-4, the FDB for B-VPLS 1000 has an entry for each of the other PEs and an entry for the ES-BMAC of ES “ESI-12”:

```
*A:PE-4# show service id 1000 fdb detail

===============================================================================
Forwarding Database, Service 1000
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1000      00:00:00:00:00:02 eMpls:                   EvpnS    05/05/17 12:49:56
                                                     P
                            192.0.2.2:262134
1000      00:00:00:00:00:03 eMpls:                   EvpnS    05/05/17 12:49:55
                                                     P
                            192.0.2.3:262136
1000      00:00:00:00:00:05 eMpls:                   EvpnS    05/05/17 12:49:50
                                                     P
                            192.0.2.5:262142
1000      00:00:00:00:12:12 eES:                     EvpnS    05/05/17 16:18:01
                                                     P
                            MAX-ESI
-------------------------------------------------------------------------------
No. of MAC Entries: 4
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
```

On PE-4, there are two BGP routes for ES-BMAC 00:00;00;00:12:12: one with next hop PE-2 and the other with next hop PE-3, as follows:

```
*A:PE-4# show router bgp routes evpn mac mac-address 00:00:00:00:12:12
===============================================================================
 BGP Router ID:192.0.2.4        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN MAC Routes
===============================================================================
Flag  Route Dist.         MacAddr           ESI
      Tag                 Mac Mobility      Label1
                          Ip Address
                          NextHop
-------------------------------------------------------------------------------
u*>i  192.0.2.2:1000      00:00:00:00:12:12 ESI-MAX
      0                   Static            LABEL 262134
                          N/A
                          192.0.2.2

u*>i  192.0.2.3:1000      00:00:00:00:12:12 ESI-MAX
      0                   Static            LABEL 262136
                          N/A
                          192.0.2.3

-------------------------------------------------------------------------------
Routes : 2
===============================================================================
```

PBB-EVPN all-active multi-homing is based on the same concepts as EVPN-MPLS all-active multi-homing: DF election, split-horizon, and aliasing.

Designated Forwarder (DF) Election

Only the DF PE for an ISID will send multicast traffic to the ES. The DF PE for an ISID can be shown with the following command:

```
*A:PE-3# show service system bgp-evpn ethernet-segment name "ESI-12" isid 1003

===============================================================================
ISID DF and Candidate List
===============================================================================
Isid          SvcId         Actv Timer Rem      DF  DF Last Change
-------------------------------------------------------------------------------
1003          1003          0                   yes 05/05/2017 16:35:04
===============================================================================

===============================================================================
DF Candidates                           Time Added
-------------------------------------------------------------------------------
192.0.2.2                               05/05/2017 16:35:01
192.0.2.3                               05/05/2017 16:34:59
-------------------------------------------------------------------------------
Number of entries: 2
===============================================================================
```

The DF PE identifies multicast traffic by looking at either the destination BMAC or the EVPN label (which can be unicast or multicast).

In the case of Epipes, there are also DF and non-DF PEs. However, traffic is usually unicast (sent to the PBB tunnel backbone-destination-bmac). The non-DF PE will usually not discard Epipe traffic to the ES, unless the packet comes with an EVPN multicast label. To avoid packet duplication at the CE for Epipes, it is recommended to either:

- configure **discard-unknown** on all the B-VPLS instances where there are PBB-Epipes. This will prevent the ingress PE from flooding Epipe traffic if the PBB tunnel BMAC is unknown in the FDB.
- configure **ingress-replication-bum-label** so that, when the PBB tunnel BMAC is unknown in the FDB, the ingress PE sends traffic with a multicast label. The non-DF will discard traffic identified as multicast at Epipes.

Ethernet-Segment Split-horizon

In PBB-EVPN all-active multi-homing, the split-horizon function is not based in the ESI label but in a source BMAC check. When BUM traffic is received on an I-VPLS, the PE will encapsulate it in PBB using the ES-BMAC as source BMAC and the group BMAC for the ISID. When the DF PE for the ISID receives that packet, it will not send it back to the ES if the packet is identified as being originated from the ES itself (based on the ES-BMAC shared between the PEs).

Aliasing

Aliasing is based on the advertisement of the same ES-BMAC with MAX-ESI from the PEs part of the same ES. PE-2 and PE-3 advertise the ES-BMAC 00:00:00:00:12:12 with MAX-ESI (ESI = all FFs, as per the RFC 7623) and as Static (protected). When the remote PEs, PE-4, and PE-5, receive the two routes for the same BMAC and MAX-ESI, they will create a single EVPN-MPLS destination that will give more than one next-hop (in this case 2), as long as ECMP > 1:

```
*A:PE-4# show service id 1000 evpn-mpls
```

```
===============================================================================
BGP EVPN-MPLS Dest
===============================================================================
TEP Address     Egr Label     Num. MACs   Mcast           Last Change
                 Transport
-------------------------------------------------------------------------------
192.0.2.2       262134        1           Yes             05/05/2017 12:06:35
                ldp
192.0.2.3       262136        1           Yes             05/05/2017 12:06:35
                ldp
192.0.2.5       262142        1           Yes             05/05/2017 12:06:35
                ldp
-------------------------------------------------------------------------------
Number of entries : 3
-------------------------------------------------------------------------------
===============================================================================
===============================================================================
BGP EVPN-MPLS Ethernet Segment Dest
===============================================================================
Eth SegId                       Num. Macs               Last Change
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================
===============================================================================
BGP EVPN-MPLS ES BMAC Dest
===============================================================================
ES BMAC Addr                            Last Change
-------------------------------------------------------------------------------
00:00:00:00:12:12                       05/05/2017 16:18:01
-------------------------------------------------------------------------------
Number of entries: 1
-------------------------------------------------------------------------------
===============================================================================
```

The EVPN-MPLS ES BMAC destination has two next hops: PE-2 and PE-3.

```
*A:PE-4# show service id 1000 evpn-mpls es-bmac 00:00:00:00:12:12

===============================================================================
BGP EVPN-MPLS ES BMAC Dest
===============================================================================
ES BMAC Addr                            Last Change
-------------------------------------------------------------------------------
00:00:00:00:12:12                       05/05/2017 12:06:35
===============================================================================

===============================================================================
BGP EVPN-MPLS ES BMAC Dest TEP Info
===============================================================================
TEP Address               Egr Label               Last Change
                          Transport
-------------------------------------------------------------------------------
192.0.2.2                 262134                  05/05/2017 12:06:35
                          ldp
192.0.2.3                 262136                  05/05/2017 12:06:35
                          ldp
-------------------------------------------------------------------------------
```

```
Number of entries : 2
-------------------------------------------------------------------------------
===============================================================================
```

A similar output will be obtained in PE-5. Unicast traffic entering I-VPLS 1001 in either PE-4 or PE-5 will be hashed and load-balanced to PE-2 and PE-3 if the destination CMAC lookup yields an **es-bmac-dest**:

```
*A:PE-5# show service id 1001 fdb detail pbb

==============================================================================
Forwarding Database, i-Vpls Service 1001
==============================================================================
MAC               Source-Identifier     B-Svc     b-Vpls MAC        Type/Age
------------------------------------------------------------------------------
00:00:10:10:10:10 eES-BMAC:             1000      00:00:00:00:12:12 L/21
                  00:00:00:00:12:12
00:00:30:30:30:30 b-eMpls:              1000      00:00:00:00:00:03 L/0
                  192.0.2.3:262136
00:00:50:50:50:50 sap:1/2/1:1001        1000      N/A               L/0
00:00:60:60:60:60 sdp:56:1001           1000      N/A               L/0
==============================================================================
```

Verify the FDB of I-VPLS 1001 for ES BMAC destination 00:00:00:00:12:12 as follows:

```
*A:PE-5# show service id 1001 fdb evpn-mpls es-bmac-dest 00:00:00:00:12:12

===============================================================================
Forwarding Database, Service 1001
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1001      00:00:10:10:10:10 eES-BMAC:                L/51     05/05/17 20:57:12
                            00:00:00:00:12:12
-------------------------------------------------------------------------------
No. of Entries: 1
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
```

If a failure occurs in the ES, the PE will withdraw the ES-BMAC and the remote PEs will remove one next-hop of the ES-BMAC EVPN-MPLS destination.

For PBB-Epipes, aliasing will also work, as long as shared-queuing or policing are enabled on the ingress PE Epipe. In Figure 55, Epipe 1003 on PE-5 requires shared-queuing or policing at the ingress SAP. Otherwise, the traffic will be sent to only one PE of the ES (usually to the lower-IP PE).

For more information about the configuration of the Ethernet-segment and its parameters, see the EVPN for MPLS Tunnels chapter.

PBB-EVPN Single-Active Multi-Homing for I-VPLS with source-bmacs

ESI-34 is a single-active Ethernet-segment (see Figure 55) with SDPs linked to it. As indicated in Table 6, only I-VPLS services can be used in this configuration. As described in section PBB-EVPN Multi-Homing, single-active ES and B-VPLS services can be configured to either use source-BMACs or ES-BMACs. The following configuration shows the former option on PE-4:

```
configure
    service
        sdp 46 mpls create
            far-end 192.0.2.6
            ldp
            no shutdown
        exit
        system
            bgp-evpn
                ethernet-segment "ESI-34" create
                    esi 01:00:00:00:00:34:00:00:00:01
                    source-bmac-lsb 34-04 es-bmac-table-size 8
                    es-activation-timer 3
                    service-carving
                        mode auto
                    exit
                    multi-homing single-active
                    sdp 46
                    no shutdown
                exit
            exit
        exit
        vpls 1000 customer 1 b-vpls create
            service-mtu 2000
            pbb
                source-bmac 00:00:00:00:00:04
            exit
            split-horizon-group "CORE" create
            exit
            bgp
            exit
            bgp-evpn
                evi 1000
                mpls
                    split-horizon-group "CORE"
                    ecmp 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            no shutdown
        exit
        vpls 1001 customer 1 i-vpls create
            pbb
```

```
                backbone-vpls 1000
                exit
            exit
            spoke-sdp 46:1001 create
            exit
            no shutdown
        exit
```

The configuration on PE-5 is similar:

```
configure
    service
        sdp 56 mpls create
            far-end 192.0.2.6
            ldp
            no shutdown
        exit
        system
            bgp-evpn
                ethernet-segment "ESI-34" create
                    esi 01:00:00:00:00:34:00:00:00:01
                    source-bmac-lsb 34-05 es-bmac-table-size 8
                    es-activation-timer 3
                    service-carving
                        mode auto
                    exit
                    multi-homing single-active
                    sdp 56
                    no shutdown
                exit
            exit
        exit
        vpls 1000 customer 1 b-vpls create
            service-mtu 2000
            pbb
                source-bmac 00:00:00:00:00:05
            exit
            split-horizon-group "CORE" create
            exit
            bgp
            exit
            bgp-evpn
                evi 1000
                mpls
                    split-horizon-group "CORE"
                    ecmp 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            no shutdown
        exit
        vpls 1001 customer 1 i-vpls create
            pbb
                backbone-vpls 1000
                exit
            exit
            sap 1/2/1:1001 create
            exit
            spoke-sdp 56:1001 create
            exit
            no shutdown
        exit
```

With the preceding configuration, PE-4 and PE-5 will not advertise ES-BMACs with MAX-ESI. Therefore, all the remote BMACs on PE-2/PE-3 are associated with regular backbone EVPN-MPLS destinations. The CMACs will be learned in the data plane associated with local SAP/SDP-bindings or remote BMACs. An example for the I-VPLS and B-VPLS FDB in PE-2 follows:

```
*A:PE-2# show service id 1001 fdb detail pbb

==============================================================================
Forwarding Database, i-Vpls Service 1001
==============================================================================
MAC               Source-Identifier     B-Svc     b-Vpls MAC        Type/Age
------------------------------------------------------------------------------
00:00:10:10:10:10 sap:lag-1:1001        1000      N/A               L/206
00:00:50:50:50:50 b-eMpls:              1000      00:00:00:00:00:05 L/309
                  192.0.2.5:262142
00:00:60:60:60:60 b-eMpls:              1000      00:00:00:00:00:05 L/180
                  192.0.2.5:262142
==============================================================================
```

The B-VPLS FDB on PE-2 looks as follows:

```
*A:PE-2# show service id 1000 fdb detail

===============================================================================
Forwarding Database, Service 1000
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1000      00:00:00:00:00:03 eMpls:                   EvpnS    05/05/17 12:49:42
                                                     P
                            192.0.2.3:262136
1000      00:00:00:00:00:04 eMpls:                   EvpnS    05/05/17 12:49:41
                                                     P
                            192.0.2.4:262133
1000      00:00:00:00:00:05 eMpls:                   EvpnS    05/05/17 12:49:37
                                                     P
                            192.0.2.5:262142
-------------------------------------------------------------------------------
No. of MAC Entries: 3
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
```

In the preceding example, the DF for ISID 1001 is PE-5. With a failure event on the SDP to MTU-6, PE-5 will not withdraw the advertised source-BMAC (because it is still being used as source-BMAC for other services and even CEs within the same service). PE-5 will send an update of the same source-BMAC instead, increasing the sequence number in the MAC mobility extended community. That will be a **flush-all-from-me** indication for the remote PEs (they will flush all the CMACs associated with the updated source-BMAC, irrespective of the service).

When the former DF (PE-5) comes back up, PE-4 will become non-DF and will send a CMAC flush indication using the same mechanism as described above.

The following example shows a failure of SDP 56 in PE-5 and the corresponding DF switchover and CMAC flush.

```
*A:PE-5#
18 2017/05/05 18:23:57.24 UTC MINOR: SVCMGR #2303 Base
"Status of SDP 56 changed to admin=up oper=down"

20 2017/05/05 18:23:57.24 UTC MINOR: SVCMGR #2095 Base
"Ethernet Segment:ESI-34, ISID:1001, Designated Forwarding state changed to:false"
```

PE-5 sends a BGP update with the same source-BMAC, increasing the sequence number in the MAC mobility extended community—CMAC flush:

```
4 2017/05/05 18:23:57.24 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 96
    Flag: 0x90 Type: 14 Len: 44 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.5
        Type: EVPN-MAC Len: 33 RD: 192.0.2.5:1000 ESI: ESI-0, tag: 0, mac len: 48
                         mac: 00:00:00:00:00:05, IP len: 0, IP: NULL, label1: 4194272
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:1000
        bgp-tunnel-encap:MPLS
        mac-mobility:Seq:1/Static
```

Individual SAP or spoke-SDP failures do not trigger any MAC flush or DF re-election. This is as per RFC 7623. In EVPN-MPLS, individual SAP/spoke-SDP failures are captured by the AD per-EVI withdrawal, which triggers a DF switchover.

PBB-EVPN Single-Active Multi-Homing for I-VPLS with ES-BMACs

As discussed throughout this chapter, the use of ES-BMACs for single-active multi-homing can minimize the number of CMACs flushed in a network. A simple change is necessary: activate the **use-es-bmac** command and ensure that the generated ES-BMACs in PE-4 and PE-5 are different (the **source-bmac-lsb** in the previous configuration had different values for PE-4 and PE-5 already):

```
*A:PE-4# configure service vpls 1000 pbb use-es-bmac
*A:PE-5# configure service vpls 1000 pbb use-es-bmac
```

On PE-4, the source BMAC LSB in ESI-34 is configured with a value of 34-04:

```
*A:PE-4# show service system bgp-evpn ethernet-segment name "ESI-34" | match BMAC
Source BMAC LSB         : 34-04
```

On PE-5, the source BMAC LSB in ESI-34 is configured with a value of 34-05:

```
*A:PE-5# show service system bgp-evpn ethernet-segment name "ESI-34" | match BMAC
Source BMAC LSB         : 34-05
```

The remote PEs (such as PE-2 in the following output) will receive two more BMACs in BGP.

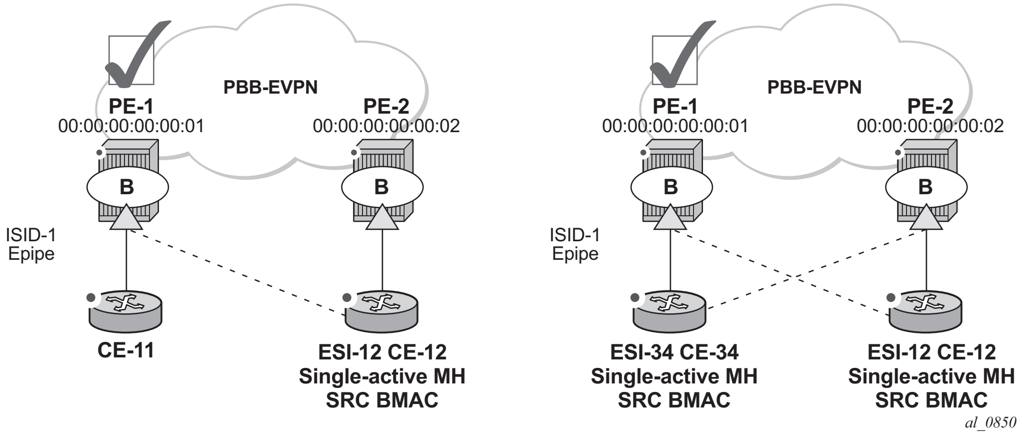
```
*A:PE-2# show service id 1000 fdb detail

===============================================================================
Forwarding Database, Service 1000
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1000      00:00:00:00:00:03 eMpls:                   EvpnS    05/05/17 12:49:42
                                                     P
                            192.0.2.3:262136
1000      00:00:00:00:00:04 eMpls:                   EvpnS    05/05/17 12:49:41
                                                     P
                            192.0.2.4:262133
1000      00:00:00:00:00:05 eMpls:                   EvpnS    05/05/17 12:49:37
                                                     P
                            192.0.2.5:262142
1000      00:00:00:00:34:04 eMpls:                   EvpnS    05/05/17 19:41:03
                                                     P
                            192.0.2.4:262133
1000      00:00:00:00:34:05 eMpls:                   EvpnS    05/05/17 19:41:00
                                                     P
                            192.0.2.5:262142
-------------------------------------------------------------------------------
No. of MAC Entries: 5
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
```

The benefit is that in case of a failure in ESI-34 (as before) the ES-BMAC is withdrawn and the remote PEs will only flush the CMACs associated with the remote ES-34, as opposed to all the CMACs associated with PE-5.

PBB-EVPN Single-Active Multi-Homing for Epipes

In the network in Figure 55, Epipes can only support single-homing or all-active multi-homing but not single-active. For non-local-switching PBB-Epipes (there is a single SAP per Epipe), only all-active multi-homing is supported. Single-active multi-homing for local-switching enabled PBB-Epipes (two SAPs are defined within the PBB-Epipe instance) is only supported in the following scenarios.

Figure 57 **PBB-EVPN Single-Active Support for Epipes**



Single-active multi-homing is supported for redundancy in a two-node, three or four SAP, scenario, as displayed in Figure 57. In these two cases, the Epipe PBB tunnel will be configured with the source BMAC of the remote PE node. When two SAPs are active in the same Epipe, local-switching is used to exchange frames between the CEs.

All-active multi-homing is not supported for redundancy in this scenario because the PE-1 PBB tunnel cannot point at a locally defined ES-BMAC.

PBB-EVPN Multi-Homing Operation

See the EVPN for MPLS Tunnels chapter for the commands to operate Ethernet-segments. Consider that there are no AD routes in PBB-EVPN. Also, the DF election algorithm will be based on the ISID values as opposed to EVIs.

Troubleshooting and Debug Commands

When troubleshooting PBB-EVPN networks, most of the troubleshooting commands discussed in EVPN for MPLS Tunnels can be used in the B-VPLS service and the base **service>system>bgp-evpn** instance. Some examples of useful commands are:

- show redundancy bgp-evpn-multi-homing
- show router bgp routes evpn (and filters)
- show service evpn-mpls [<TEP ip-address>]
- show service id bgp-evpn
- show service id evpn-mpls (and modifiers)
- show service id fdb pbb (and modifiers)
- show service system bgp-evpn
- show service system bgp-evpn ethernet-segment (and modifiers)
- debug router bgp update
- log-id 99

In addition, the following **tools dump** commands also discussed in EVPN for MPLS Tunnels can help too:

- tools dump service evpn usage
- tools dump service system bgp-evpn ethernet-segment <name> isid df (Note: **isid** is used instead of **evi**.)

There are two aspects that are specific to PBB-EVPN and not EVPN:

1. Consumption of virtual BMACs in the system— source-bmacs, sap-bmacs, sdp-bmacs, and es-bmacs are system BMACs that use FDB space but are not shown in the FDB together with the rest of the learned MACs. The following command provides information about the virtual system MACs consumed in the system.

```
*A:PE-3# tools dump redundancy src-bmac-lsb
Src-bmac-lsb:      3 (00-03) User: B-Vpls - 1 service(s)
Src-bmac-lsb:   4626 (12-12) User: ES

Total Src-bmac-lsbs = 2
```

2. Consumption of MFIBs — when ISIDs are not using the default-multicast list in the B-VPLS context for sending BUM traffic, an MFIB is consumed per ISID. The following command provides information about the consumption of MFIBs per system and per B-VPLS.

```
*A:PE-2# tools dump service vpls-pbb-mfib-stats detail
```

```
Service Manager VPLS PBB MFIB statistics at 05/05/2017 13:19:21:

Usage per Service
   ServiceId    MFIB User      Count
  ------------+-------------+-------
   1000         Evpn               1
  ------------+-------------+-------
                       Total       1

MMRP
  Current Usage     :     0
  System Limit      :  8191 Full, 40959 ESOnly
  Per Service Limit :  2048 Full,  8192 ESOnly

SPB
  Current Usage     :     0
  System Limit      :  8191
  Per Service Limit :  8191

Evpn
  Current Usage     :     1
  System Limit      : 40959
  Per Service Limit :  8191
```

Conclusion

In addition to a full RFC 7432 EVPN-MPLS implementation, SR OS supports PBB-EVPN as per RFC 7623 for large Layer 2 deployments, including single-active and all-active multi-homing. This example has shown how to configure and operate a PBB-EVPN network focusing on the specific aspects of PBB-EVPN compared to EVPN-MPLS.

EVPN for VXLAN Tunnels (Layer 2)

This chapter provides information about Ethernet Virtual Private Network (EVPN) for Virtual eXtensible Local Area Network (VXLAN) tunnels in VPLS services.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter is applicable to SR OS and was initially written for release 12.0.R4. The CLI in the current edition is based on SR OS Release 15.0.R2. Ethernet Virtual Private Network (EVPN) is a control plane technology and does not have line card hardware dependencies.

Overview

SR OS supports the EVPN control plane with Virtual eXtensible Local Area Network (VXLAN) data plane in VPLS services.

EVPN is an IETF technology (draft-ietf-l2vpn-evpn) that uses a dedicated BGP address family which allows VPLS services to be operated in a similar way to IP-VPNs, where the MAC addresses, IP addresses, and the information to set up the flooding tree are distributed by BGP. EVPN can be used as the control plane for different data plane encapsulations, such as VXLAN and MPLS.

VXLAN (draft-mahalingam-dutt-dcops-vxlan) is an overlay IP tunneling technology used to carry Ethernet traffic over any IP network, and it is becoming the de facto standard for overlay data centers and networks. Compared to other IP overlay tunneling technologies, such as GRE, VXLAN supports multi-tenancy and multi-pathing:

- A tenant identifier, the VXLAN Network Identifier (VNI), is encoded in the VXLAN header and allows each tenant to have an isolated Layer 2 domain.

- VXLAN supports multi-pathing scalability through ECMP. VXLAN uses the outer source UDP port as an entropy field that can be used by the core IP routers to balance the load across different paths.

In SR OS, EVPN and VXLAN can be enabled in VPLS or R-VPLS services. In this chapter, EVPN-VXLAN services will refer to VPLS or R-VPLS services with EVPN and VXLAN enabled. These services can terminate/originate VXLAN tunnels and may have SAPs and/or SDP bindings at the same time. Some other SR OS implementation-specific considerations are the following:

- VXLAN is only supported on network or hybrid ports with null or dot1q encapsulation on Ethernet/LAG/POS/APS interfaces.
- VXLAN packets are originated/terminated with the system IPv4 address, in other words, a system originating VXLAN packets will use the system IP address as source outer IPv4 address and systems will only process VXLAN packets if their destination outer IPv4 address matches its own system IP address.
- Data plane MAC learning is not supported over VXLAN bindings. Only the control plane (EVPN) will be used for populating the FDB with MAC addresses associated to VXLAN bindings.
- EVPN provides support for the following features that are described in this chapter:
  - The BGP advertisement of the MAC addresses learned on SAPs, SDP-bindings and conditional static MACs to the remote BGP peers. The advertisement of MAC addresses in BGP can optionally be disabled.
  - The optional advertisement of an unknown MAC route, that allows the remote EVPN PEs or Network Virtualization Edge devices (NVEs) to suppress the unknown unicast flooding and send any unknown unicast frame to the owner of the unknown MAC route.
  - Ingress replication of Broadcast, Unknown unicast and Multicast (BUM) packets over VXLAN.
  - A proxy-ARP table per service populated by the MAC-IP pairs received in BGP MAC advertisements. When an ARP request is received on a SAP or SDP-binding, the system will perform a lookup on this table and will reply to the ARP request if the lookup yields a valid result.
  - MAC mobility and static-MAC protection as described in draft-ietf-l2vpn-evpn, as well as MAC duplication detection.
- Multi-homing redundancy for SAPs and SDP-bindings in EVPN-VXLAN services is supported through BGP Multi-homing (L2VPN BGP address family). Only one BGP-MH site is supported in an EVPN-VXLAN service.

One of the main applications for EVPN-VXLAN services in SR OS is the Data Center Gateway (DC GW) function. In such an application, EVPN and VXLAN are expected to be used within the data center and VPLS SDP-bindings or SAPs are expected to be used for the connectivity to the WAN. When the system is used as a DC GW, a VPLS service is configured per Layer 2 domain that has to be extended to the WAN. In those VPLS services, BGP EVPN automatically sets up the VXLAN auto-bindings that connect the DC GW to the data center Network Virtual edge Devices (NVEs). The WAN connectivity is based on regular VPLS constructs where SAPs (null, dot1q and QinQ), spoke-SDPs (FEC type 128 and 129, not BGP-VPLS), and mesh-SDPs are supported. B-VPLS or I-VPLS services are not supported.

Although the DC GW application is one of the most common uses for this feature, this chapter focuses on the configuration and operation of EVPN-VXLAN for Layer 2 services in general, and its integration with regular VPLS services in MPLS networks.

Configuration

This section describes the configuration of EVPN-VXLAN on SR OS as well as the available troubleshooting and show commands. This example focuses on the following configuration aspects:

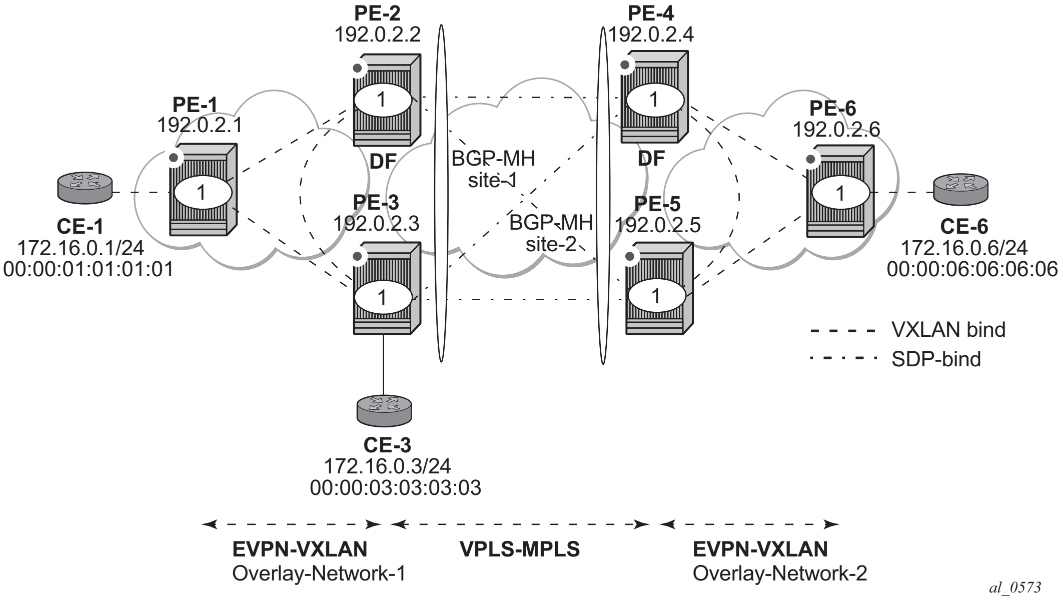
- Enabling EVPN and VXLAN in a VPLS service, including the use of BGP-EVPN, BGP Auto-discovery (BGP-AD), and BGP-Multi-homing (BGP-MH) in the same VPLS instance.
- Scaling BGP-MH resiliency with the use of operational groups (oper-groups).
- Use of proxy-ARP in EVPN-VXLAN services
- MAC mobility, MAC duplication, and MAC protection in EVPN-VXLAN services.

The configuration will be shown for PE-1, PE-2, and PE-3 only; the PEs in Overlay-Network-2 (Figure 58) have an equivalent configuration.

Enabling EVPN-VXLAN in a VPLS Service

Figure 58 shows the topology used in this example.

Figure 58 **EVPN-VXLAN Example Topology**



The example topology shows two overlay (VXLAN) networks interconnected by an MPLS network:

- PE-1, PE-2, and PE-3 are part of Overlay-Network-1
- PE-4, PE-5, and PE-6 are part of Overlay-Network-2

CE-1, CE-3, and CE-6 belong to the same IP subnet, therefore, Layer 2 connectivity must be provided to them.

The example topology can illustrate a Data Center Interconnect (DCI) example, where Overlay-Network-1 and Overlay-Network-2 are two data centers interconnected through an MPLS WAN. In this application, CE-1, CE-3, and CE-6 would simulate virtual machines or appliances, PE-2/3/4/5 would act as DC GWs and PE-1/6 as NVEs (or virtual PEs running on compute infrastructure).

The following protocols and objects are configured beforehand:

- The ports interconnecting the six PEs in Figure 58 are configured as network ports (or hybrid) and have router network interfaces defined on them. Only the ports connected to the CEs are configured as access ports.
- The six PEs shown in the Figure 58 are running IS-IS for the global routing table with the four core PEs interconnected using IS-IS Level-2 point-to-point interfaces and each overlay network is using IS-IS Level-1 point-to-point interfaces.

- LDP is used as the MPLS protocol to signal transport tunnel labels among PE-2, PE-3, PE-4 and PE-5. There is no LDP running in the two overlay networks.
- The network port MTU (in all the ports sending/receiving VXLAN packets) must be at least 50-bytes (54 if dot1q encapsulation is used) greater than the service MTU in order to accommodate the size of the VXLAN header.

Once the IGP infrastructure and LDP are enabled in the core, BGP has to be configured. In this example, two BGP families have to be enabled: EVPN within each overlay-network for the exchange of MAC/IP addresses and setting up the flooding domains, and L2-VPN for the use of BGP-MH and BGP-AD in the VPLS-MPLS network.

As an example, the following CLI output shows the relevant BGP configuration of PE-1, which only needs the EVPN family. PE-6 would have a similar BGP configuration. The use of route reflectors (RRs) in these type of scenarios is common. Although this example does not use RRs, an EVPN RR could have been used in Overlay-Network-1 and Overlay-Network-2 and an L2-VPN RR could have been used in the core VPLS-MPLS network.

```
configure
    router
        autonomous-system 64500
        bgp
            vpn-apply-import
            vpn-apply-export
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            rapid-update evpn
            group "DC"
                family evpn
                peer-as 64500
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
            exit
        exit
```

The BGP configuration on PE-2 is as follows:

```
configure
    router
        autonomous-system 64500
        bgp
            vpn-apply-import
            vpn-apply-export
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            rapid-update l2-vpn evpn
            group "DC"
                family l2-vpn evpn
                peer-as 64500
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.3
                exit
            exit
            group "WAN"
                family l2-vpn
                peer-as 64500
                neighbor 192.0.2.4
                exit
                neighbor 192.0.2.5
                exit
            exit
        exit
```

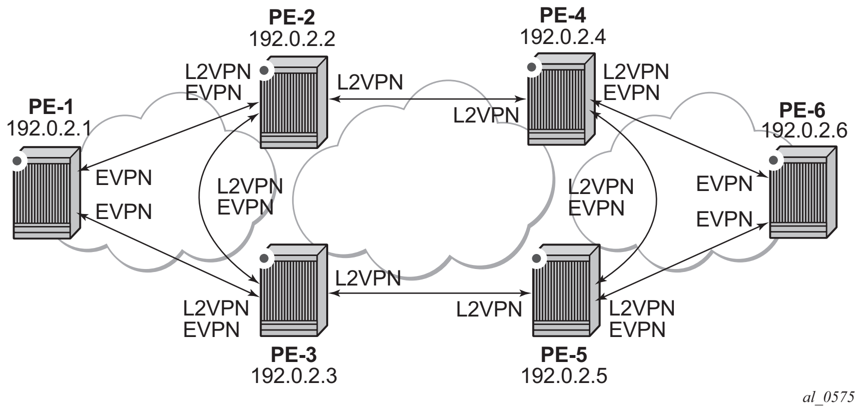
The BGP configuration on PE-3 is as follows:

```
configure
    router
        autonomous-system 64500
        bgp
            vpn-apply-import
            vpn-apply-export
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            rapid-update l2-vpn evpn
            group "DC"
                family l2-vpn evpn
                peer-as 64500
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
            exit
            group "WAN"
                family l2-vpn
                peer-as 64500
                neighbor 192.0.2.4
                exit
                neighbor 192.0.2.5
                exit
            exit
        exit
```

The BGP configuration on PE-4 and PE-5 is equivalent.

Figure 59 shows the BGP peering sessions among the PEs and the enabled BGP families. PE-1 will only establish an EVPN peering session with its peers (only the EVPN family is enabled on PE-1), even though PE-2 and PE-3 have EVPN and L2-VPN families configured.

Figure 59 **BGP Adjacencies and Enabled Families**



Once the network infrastructure is running properly, the actual service configuration can be carried out. The following CLI outputs show the configuration of VPLS 1 in PE-1, PE-2, and PE-3 as per the topology illustrated in Figure 58.

VPLS 1 in those three PEs are interconnected using VXLAN bindings, whereas PE-2 and PE-3 are connected to the remote PEs by means of BGP-AD SDP-bindings. Although BGP-AD SDP-bindings are used in this example for the connectivity of the EVPN-VXLAN PEs to a regular VPLS network, SAPs, manual spoke-SDPs or mesh-SDPs could have been used instead. BGP-VPLS cannot be enabled in an EVPN-VXLAN VPLS service.

VPLS 1 is configured on PE-1, as follows:

```
configure
    service
        vpls 1 customer 1 create
            vxlan vni 1 create
            exit
            bgp
                route-distinguisher 64500:1
                route-target export target:64500:12 import target:64500:12
            exit
            bgp-evpn
                vxlan
                    no shutdown
                exit
            exit
            sap 1/2/1:1 create
            exit
            no shutdown
```

EVPN-VXLAN is enabled by the configuration of a valid VXLAN Network Identifier (VNI) and the **bgp-evpn>vxlan>no shutdown** command. These two commands, along with the required BGP Route Distinguisher (RD) and Route Target (RT) information, are the minimum mandatory attributes:

- The VNI is a 24-bit identifier with valid values in the [1..16777215] range. This defines the VNI that SR OS will use in the EVPN routes generated for the VPLS service, and therefore the VNI that the system expects to see in the VXLAN packets destined to that particular VPLS service. The configured VNI determines the VNI that has to be received in the packets for the VPLS service, but not the VNI that will be sent in VXLAN packets to remote PEs for the service. In other words, in this example, VPLS 1 is configured with VNI=1 in all the PEs; however, each PE could have used a different VNI. The VNI is a system-wide significant value and two VPLS services cannot be configured with the same VNI.
- The **bgp-evpn>vxlan>no shutdown** command enables the use of EVPN for VXLAN. It requires the previous configuration of the VNI, RD, and RT. As soon as this command is executed, EVPN will advertise an inclusive multicast route to all of the BGP EVPN peers (regardless of the existing SAP/SDP-binding operational status). The exchange of inclusive multicast routes allows the establishment of the VXLAN bindings among the PEs.

Upon the reception of the EVPN inclusive multicast routes from PE-2 and PE-3, PE-1 will automatically set up its VXLAN bindings for VPLS 1. A VXLAN binding is represented by an (egress VTEP, egress VNI) pair, where VTEP is a VXLAN Termination End Point. This can be shown with the following show commands:

```
*A:PE-1# show service id 1 vxlan
===============================================================================
Vxlan Src Vtep IP: N/A
===============================================================================
VPLS VXLAN, Ingress VXLAN Network Id: 1
Creation Origin: manual
Assisted-Replication: none
RestProtSrcMacAct: none

===============================================================================
VPLS VXLAN service Network Specifics
===============================================================================
Ing Net QoS Policy : none                            Vxlan VNI Id     : 1
Ingress FP QGrp    : (none)                          Ing FP QGrp Inst : (none)

===============================================================================
Egress VTEP, VNI
===============================================================================
VTEP Address                            Egress VNI  Num. MACs   Mcast Oper  L2
                                                                      State PBR
-------------------------------------------------------------------------------
192.0.2.2                               1           0           BUM   Up    No
192.0.2.3                               1           0           BUM   Up    No
-------------------------------------------------------------------------------
Number of Egress VTEP, VNI : 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#


*A:PE-1# show service vxlan
```

```
===============================================================================
VXLAN Tunnel Endpoints (VTEPs)
===============================================================================
VTEP Address                                  Number of Egress VNIs    Oper
                                                                       State
-------------------------------------------------------------------------------
192.0.2.2                                     1                        Up
192.0.2.3                                     1                        Up
-------------------------------------------------------------------------------
Number of VTEPs: 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

To actually see this output, the VPLS service needs to be configured on all PEs, with import and export policy “vsi-policy-1” defined on the core PEs; see further. As can be seen in the CLI output, PE-1 has two VXLAN bindings: one to PE-2 and one to PE-3. Both use egress VNI=1 (the actual VNI used in its egress VXLAN packets) and both are part of the flooding multicast list (BUM) for VPLS 1 and are up. There is no layer 2 Policy-Based Routing (L2 PBR).

- The **Mcast= BUM** entry is set when the proper inclusive multicast route is received from the remote VTEP. The VXLAN binding will be used to flood BUM (Broadcast, Unknown unicast, Multicast) packets.
- The **Oper State** is based on the existence of the VTEP in the global routing table.

The VPLS 1 configuration of PE-2 and PE-3 is as follows:

```
configure
    service
        pw-template 1 create
        exit
        vpls 1 customer 1 create
            vxlan vni 1 create
            exit
            bgp
                route-distinguisher 192.0.2.2:1
                vsi-export "vsi-policy-1"
                vsi-import "vsi-policy-1"
                pw-template-binding 1 split-horizon-group "CORE"
                exit
            exit
            bgp-ad
                vpls-id 64500:1
                no shutdown
            exit
            bgp-evpn
                vxlan
                    no shutdown
                exit
            exit
            site "site-1" create
                site-id 1
                split-horizon-group "CORE"
                no shutdown
            exit
            no shutdown
```

On PE-3:

```
configure
    service
        pw-template 1 create
        exit
        vpls 1 customer 1 create
            vxlan vni 1 create
            exit
            bgp
                route-distinguisher 192.0.2.3:1
                vsi-export "vsi-policy-1"
                vsi-import "vsi-policy-1"
                pw-template-binding 1 split-horizon-group "CORE"
                exit
            exit
            bgp-ad
                vpls-id 64500:1
                no shutdown
            exit
            bgp-evpn
                vxlan
                    no shutdown
                exit
            exit
            site "site-1" create
                site-id 1
                split-horizon-group "CORE"
                no shutdown
            exit
            sap 1/2/1:1 create
            exit
            no shutdown
```

In addition to the VNI and **bgp-evpn>vxlan>no shutdown** commands for enabling EVPN-VXLAN in VPLS 1, PE-2 and PE-3 require the configuration of BGP-AD for the discovery and establishment of FEC129 spoke SDPs to the remote PEs in the core, as well as BGP-MH for redundancy. As outlined in Figure 58, there are two BGP-MH sites defined in the network: site-1 is used on PE-2/PE-3 and site-2 is used on PE-4/PE-5. Only one of the two gateway PEs in each overlay network will be the designated forwarder (DF) for VPLS 1, and only the DF will send/receive traffic for VPLS 1 in the overlay network. The following considerations must be taken into account when configuring the connectivity of EVPN-VXLAN services to regular VPLS objects:

- As discussed, in this example, BGP-AD spoke-SDPs are used, but SAPs, manual spoke-SDPs, or mesh-SDPs are also supported.

- In this example, BGP-AD spoke-SDPs are auto-instantiated using **pw-template-binding 1 split-horizon-group “CORE”.**
    - This requires the creation of the pw-template 1 (**config>service>pw-template 1 create**).
- The split-horizon-group CORE is added to the BGP-MH site “site-1”. This statement will ensure that all the spoke SDPs automatically established to the remote PEs are part of the BGP-MH site.
- Although the route-targets for the Overlay-Network and the VPLS-MPLS network can have the same value for the same VPLS service, they are usually different. This example assumes the use of RT-DC-1 in Overlay-Network-1 and RT-WAN-1 in the VPLS-MPLS core for VPLS 1. The “vsi-policy-1” allows the system to export and import the right RTs for VPLS 1 on the core PEs:

```
configure
    router
        policy-options
            begin
            community "RT-DC-1" members "target:64500:12"
            community "RT-WAN-1" members "target:64500:11"
            policy-statement "vsi-policy-1"
                entry 10     # to import all the EVPN routes with RT-DC-1
                    from
                        community "RT-DC-1"
                        family evpn
                    exit
                    action accept
                    exit
                exit
                entry 20     # to import all the BGP-AD/MH routes from the WAN
                    from
                        community "RT-WAN-1"
                        family l2-vpn
                    exit
                    action accept
                    exit
                exit
                entry 30      # to export all the EVPN routes with "RT-DC-1"
                    from
                        family evpn
                    exit
                    action accept
                        community add "RT-DC-1"
                    exit
                exit
                entry 40      # to export all the BGP-AD/MH routes with "RT-WAN-1"
                    from
                        family l2-vpn
                    exit
                    action accept
                        community add "RT-WAN-1"
                    exit
                exit
                default-action drop
                exit
            exit
            commit
```

Once PE-2 and PE-3 are configured as shown, they will set up the spoke SDPs and will run the DF election algorithm to determine the operational status of those spoke SDPs. See chapters LDP VPLS Using BGP Auto-Discovery and BGP Multi-Homing for VPLS Networks for more information about the use of BGP-AD and BGP-MH.

In the configuration for VPLS 1, both gateway PEs, PE-2 and PE-3, will attempt to establish two parallel Layer 2 paths between each other (a BGP-AD spoke SDP and a EVPN VXLAN binding). Because that would create a Layer 2 loop, the SR OS implementation gives priority to the EVPN path and only the VXLAN binding will be active. In other words, when a VXLAN (egress VTEP, VNI) and a spoke SDP are attempted to be set up to the same far-end IP address at the same time, the VXLAN path will prevail and the spoke SDP will be kept down. The spoke SDP will only be brought up if the VXLAN (egress VTEP, VNI) goes down.

This behavior can be easily observed in this setup by using the following show commands. In PE-2, the spoke SDP to far-end PE-3 will be down with a **EvpnRouteConflict** Flag. The (egress VTEP, VNI) = (192.0.2.3, 1) VXLAN bind will be up.

```
*A:PE-2# show service id 1 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 1                   Vpn Id            : 0
Service Type      : VPLS
---snip---
Admin State       : Up                  Oper State        : Up
MTU               : 1514                Def. Mesh VC Id   : 1
SAP Count         : 0                   SDP Bind Count    : 3
---snip---
-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sdp:17405:4294967293 SB(192.0.2.3)       BgpAd        0       8978    Up   Down
sdp:17406:4294967294 SB(192.0.2.5)       BgpAd        0       8978    Up   Up
sdp:17407:4294967295 SB(192.0.2.4)       BgpAd        0       8978    Up   Up
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-2#


*A:PE-2# show service id 1 all | match Flag context all
Flags              : PWPeerFaultStatusBits
                     EvpnRouteConflict
Flags              : None
Flags              : None
*A:PE-2#
```

```
*A:PE-2# show service id 1 vxlan
===============================================================================
Vxlan Src Vtep IP: N/A
===============================================================================
VPLS VXLAN, Ingress VXLAN Network Id: 1
Creation Origin: manual
Assisted-Replication: none
RestProtSrcMacAct: none

===============================================================================
VPLS VXLAN service Network Specifics
===============================================================================
Ing Net QoS Policy : none                                Vxlan VNI Id     : 1
Ingress FP QGrp    : (none)                              Ing FP QGrp Inst : (none)

===============================================================================
Egress VTEP, VNI
===============================================================================
VTEP Address                            Egress VNI  Num. MACs   Mcast Oper  L2
                                                                      State PBR
-------------------------------------------------------------------------------
192.0.2.1                               1           0           BUM   Up    No
192.0.2.3                               1           0           BUM   Up    No
-------------------------------------------------------------------------------
Number of Egress VTEP, VNI : 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-2#
```

At the non-DF, PE-3, all the spoke SDPs will be down due to BGP-MH:

```
*A:PE-3# show service id 1 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 1                   Vpn Id            : 0
Service Type      : VPLS
---snip---
Admin State       : Up                  Oper State        : Up
MTU               : 1514                Def. Mesh VC Id   : 1
SAP Count         : 1                   SDP Bind Count    : 3
---snip---
-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/2/1:1                              q-tag        1518    1518    Up   Up
sdp:17405:4294967293 SB(192.0.2.4)       BgpAd        0       8978    Up   Down
sdp:17406:4294967294 SB(192.0.2.5)       BgpAd        0       8978    Up   Down
sdp:17407:4294967295 SB(192.0.2.2)       BgpAd        0       8978    Up   Down
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-3#


*A:PE-3# show service id 1 all | match Flag context all
Flags              : StandbyForMHProtocol
Flags              : StandbyForMHProtocol
Flags              : StandbyForMHProtocol
                     PWPeerFaultStatusBits
                     EvpnRouteConflict
Flags              : None
```

MAC Learning and unknown-mac-route

Once the VPLS service (VPLS 1) is configured, the network allows the CEs to exchange unicast and BUM traffic over the overlay and VPLS-MPLS service infrastructure. BUM traffic sent by CE-1 will be ingress-replicated by PE-1 to PE-2 and PE-3, and propagated by PE-2 (the DF) to the remote network. From this point on, MAC addresses will be learned on active SAPs and spoke SDPs and advertised in EVPN MAC routes. No data plane MAC learning is carried out on VXLAN bindings. MACs associated with (egress VTEP, VNI) bindings will always be learned through EVPN.

The following CLI output shows the reception of an EVPN MAC route on PE-1 and how the (CE-3) MAC address appears in the FDB for VPLS 1.

```
4 2017/05/03 11:14:06.95 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 88
    Flag: 0x90 Type: 14 Len: 44 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.3
        Type: EVPN-MAC Len: 33 RD: 192.0.2.3:1 ESI: ESI-0, tag: 0, mac len: 48
                      mac: 00:00:03:03:03:03, IP len: 0, IP: NULL, label1: 1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:64500:12
        bgp-tunnel-encap:VXLAN
"


*A:PE-1# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         00:00:01:01:01:01 sap:1/2/1:1              L/0      05/03/17 11:29:07
1         00:00:03:03:03:03 vxlan:                   Evpn     05/03/17 11:13:02
                            192.0.2.3:1
1         00:00:06:06:06:06 vxlan:                   Evpn     05/03/17 11:24:05
                            192.0.2.2:1
```

```
-------------------------------------------------------------------------------
No. of MAC Entries: 3
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-1#
```

When a frame destined to 00:00:03:03:03:03 enters SAP 1/2/1:1, it is encapsulated into a VXLAN packet with outer destination IP 192.0.2.3 and VNI 1, and sent on the wire.

In virtualized data center networks where all the MACs are known beforehand (all the virtual machine and appliance MACs are distributed by EVPN before any traffic flows), unknown MAC addresses are always outside the data center. If that is the case, the DC GWs can make use of the **unknown-mac-route** so that the DC NVEs supporting the concept of this route send the unknown unicast traffic only to the DC GW. This minimizes the flooding within the data center, as explained in draft-rabadan-l2vpn-dci-evpn-overlay.

In this example, the unknown MAC route is configured in the gateway PEs (PE-2, PE-3 and PE-4, PE-5) in the following way:

```
*A:PE-2# configure service vpls 1 bgp-evpn unknown-mac-route


22 2017/05/03 12:04:34.03 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 88
    Flag: 0x90 Type: 14 Len: 44 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.2
        Type: EVPN-MAC Len: 33 RD: 192.0.2.2:1 ESI: ESI-0, tag: 0, mac len: 48
                       mac: 00:00:00:00:00:00, IP len: 0, IP: NULL, label1: 1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:64500:12
        bgp-tunnel-encap:VXLAN
"
```

Note that:

- Although SR OS can generate the unknown MAC route, it will never honor it and normal flooding applies when an unknown unicast packet arrives at an ingress SAP/SDP-binding.

- When unknown-mac-route is configured, it will only be generated when: a) no BGP-MH site is configured within the same VPLS service or b) a site is configured and the site is DF (Designated Forwarder) in the PE. If the site becomes a non-DF site, the unknown-mac-route will be withdrawn.
- If the unknown-mac-route is used in the DC GW and all the NVEs in the DC understand it, the advertisement of MAC addresses can be disabled with the [**no**] **mac-advertisement** command. If so, SR OS will only advertise the unknown-mac-route.

```
*A:PE-2# configure service vpls 1 bgp-evpn unknown-mac-route
*A:PE-2# configure service vpls 1 bgp-evpn no mac-advertisement
```

Scaling BGP-MH Resiliency with the Use of Operational Groups

In Figure 58, VPLS 1 in PE-2/PE-3 is configured with a BGP-MH site that controls which of the two PEs forwards the traffic to the remote PEs (in this case, PE-2 is the DF and the gateway responsible for forwarding packets to the remote PEs).

When new VPLS services are required in PE-2/PE-3, the same BGP-MH configuration can be used. However, if the number of VPLS services grows significantly, the use of individual BGP-MH sites per service will not scale. Because all the services in these two PEs share the same physical topology, the use of operational groups can provide a simple and scalable way of providing resiliency to as many services as the user needs (up to the maximum number of VPLS services per system).

The way operational groups can be used to scale this type of deployments is the following (using the network topology in Figure 58 and focusing on Overlay-Network-1):

- A control-VPLS service is defined in PE-2 and PE-3. For instance, VPLS 1.
  - This service is configured with a BGP-MH site in both PEs.
  - An oper-group **control-vpls-1** is created and associated to the pw-template-binding 1 in VPLS 1.
- Data VPLS services are defined in both PEs. For instance: VPLS 2, VPLS 3,... VPLS 999.
  - In all these services, the pw-template-binding is configured with **monitor-oper-group "control-vpls-1".**

   - The status of the spoke SDPs in the data VPLS services depends on the status of the operational group. If there is a DF switchover in VPLS 1 and VPLS 1 spoke SDPs go down on PE-2, all the spoke SDPs in all the data VPLS services controlled by “control-vpls-1” in PE-2 will go down too. In the same way, the spoke SDPs in PE-3 will come up.
- To allow per-service load balancing a second control-VPLS service with a different BGP-MH site should be configured.
   - For instance, VPLS 1 might have PE-2 as the DF and VPLS 1000 might be a second control-VPLS service with PE-3 as the DF.
   - Each control-VPLS would control a group of data VPLS services based on the definition and association of a second operational group.

The following example shows the modification of VPLS 1 as the control-VPLS and the configuration of VPLS 2 as a data-VPLS on PE-2. VPLS 1 controls the VPLS 2 spoke SDP status.

```
configure
    service
        oper-group "control-vpls-1" create
        exit
        vpls 1 customer 1 create
            description "control-VPLS"
            bgp
                pw-template-binding 1 split-horizon-group "CORE"
                    oper-group "control-vpls-1"
                exit
            exit
        exit
        vpls 2 customer 1 create
            description "data-VPLS"
            vxlan vni 2 create
            exit
            bgp
                route-distinguisher 192.0.2.2:2
                vsi-export "vsi-policy-2"
                vsi-import "vsi-policy-2"
                pw-template-binding 1
                    monitor-oper-group "control-vpls-1"
                exit
            exit
            bgp-ad
                vpls-id 64500:2
                no shutdown
            exit
            bgp-evpn
                unknown-mac-route
                vxlan
                    no shutdown
                exit
            exit
            no shutdown
        exit
```

Use of Proxy-ARP in EVPN-VXLAN Services

EVPN-VXLAN services support proxy-ARP functionality that is enabled by the **proxy-arp [no] shutdown** command. The default value is shutdown. When proxy-arp is enabled:

- MAC and IP addresses contained in the received valid EVPN MAC routes are populated in the proxy-ARP table.
- ARP-request messages received on SAPs and SDP-bindings are intercepted and the target IP address is looked up. If the IP address is found, an ARP reply will be issued based on the information found in the proxy-ARP table, otherwise the ARP request would be flooded in the VPLS service (except for the source SAP/SDP binding).
- ARP-reply messages received on SAPs and SDP-bindings are also intercepted and sent to the CPM. These ARP-reply messages are re-injected in the data plane and forwarded based on the FDB information to the destination MAC address. If the destination MAC address is not in the FDB, the ARP-reply message will be flooded in the VPLS service (except for the source SAP/SDP binding).

The following CLI output shows the proxy-ARP configuration in PE-3 and a received valid MAC route that includes the MAC and IP of CE-1. This MAC-IP pair is installed in the proxy-ARP table for VPLS 1.

```
configure service vpls 1 proxy-arp no shutdown
```

```
27  2017/05/03 13:50:30.49 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 92
    Flag: 0x90 Type: 14 Len: 48 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.1
        Type: EVPN-MAC Len: 37 RD: 192.0.2.1:1 ESI: ESI-0, tag: 0, mac len: 48
              mac: 00:00:01:01:01:01, IP len: 4, IP: 172.16.0.1, label1: 1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:64500:12
        bgp-tunnel-encap:VXLAN
"
```

```
*A:PE-3# show service id 1 proxy-arp detail
-------------------------------------------------------------------------------
Proxy Arp
-------------------------------------------------------------------------------
Admin State       : enabled
```

```
Dyn Populate      : disabled
Age Time          : disabled            Send Refresh      : disabled
Table Size        : 250                 Total             : 1
Static Count      : 0                   EVPN Count        : 1
Dynamic Count     : 0                   Duplicate Count   : 0

Dup Detect
-------------------------------------------------------------------------------
Detect Window     : 3 mins              Num Moves         : 5
Hold down         : 9 mins
Anti Spoof MAC    : None

EVPN
-------------------------------------------------------------------------------
Garp Flood        : enabled             Req Flood         : enabled
Static Black Hole : disabled
-------------------------------------------------------------------------------

===============================================================================
VPLS Proxy Arp Entries
===============================================================================
IP Address          Mac Address         Type      Status    Last Update
-------------------------------------------------------------------------------
172.16.0.1          00:00:01:01:01:01   evpn      active    05/03/2017 12:28:30
-------------------------------------------------------------------------------
Number of entries : 1
-------------------------------------------------------------------------------
===============================================================================
*A:PE-3#
```

SR OS does not include a host IP address in any EVPN MAC advertisement for a MAC learned on a SAP or SDP-binding. Host IP addresses are only included in the EVPN MAC advertisements corresponding to R-VPLS IP interfaces. When deployed as DC GW in a Nuage architecture, the Nuage Networks Virtual Services Controller (VSC) or Virtual Services Gateway (VSG) will send virtual machine and host MAC/IP pairs in EVPN MAC routes. See the Nokia Nuage documentation for more information about the Nuage DC architecture. The 7x50 DC GW will populate the proxy-ARP tables with those MAC/IP pairs.

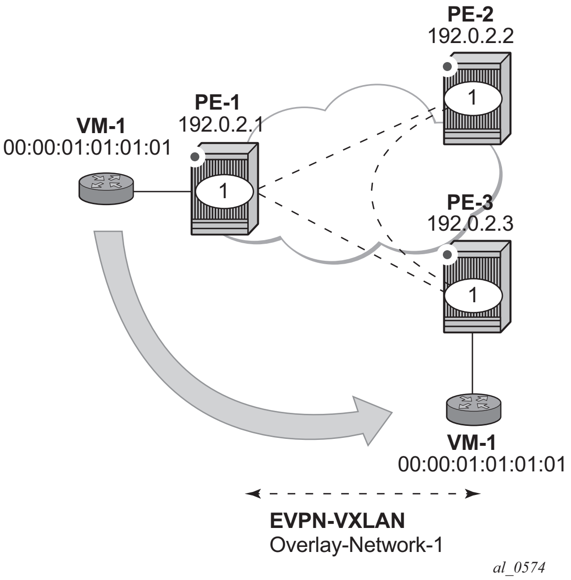
In the preceding CLI excerpt, assume that PE-1 is replaced by a Nuage VSC that sends the pair <172.16.0.1, 00:00:01:01:01:01> in an EVPN MAC route. PE-3 receives the advertisement and adds the entry to its proxy-ARP table for VPLS 1.

The proxy-ARP feature was significantly improved in SR OS release 13.0; see chapter EVPN for MPLS Tunnels.

MAC Mobility, MAC Duplication, and MAC Protection in EVPN

MAC mobility, duplication and protection are fully supported as specified in draft-ietf-l2vpn-evpn. Figure 60 illustrates the concept of mobility (Virtual Machine VM-1 moves from PE-1 to PE-3).

Figure 60 **EVPN MAC Mobility**



MAC mobility is handled in EVPN by the use of sequence numbers in the MAC routes. When 00:00:01:01:01:01 moves from PE-1 to PE-3, SR OS will gracefully handle it in this way:

- 00:00:01:01:01:01 moves to PE-3 SAP 1/2/1:1
- PE-3 advertises 00:00:01:01:01:01 using a higher sequence number (the first time a MAC is advertised, EVPN uses sequence number 0).
- PE-2 at this point has two valid MAC routes for 00:00:01:01:01:01. It picks up the one coming from PE-3 since the sequence number is higher.
- PE-1 receives the MAC route, and since the sequence number is higher than the one for its own route, it updates the FDB and withdraws its own MAC route.

However, if MAC 00:00:01:01:01:01 is constantly learned on the PE-1 and PE-3 SAPs, the preceding process causes an endless exchange of MAC route advertisements and withdraws that has a negative impact on all the PEs in the EVPN network. This issue is known as “MAC duplication” and is originated by a loop at the access or a duplicated MAC address in two hosts of the same service. SR OS solves this issue through the use of the mac-duplication detection feature. MAC-duplication is always enabled with the following default settings:

```
*A:PE-1>config>service>vpls>bgp-evpn# info detail | match mac-duplication context all
----------------------------------------------
                mac-duplication
                    detect num-moves 5 window 3
                    retry 9
                    no black-hole-dup-mac
```

Where:

- **num-moves** — Identifies the number of MAC moves in a VPLS service. The counter is incremented when a MAC is locally relearned in the FDB or flushed from the FDB due to the reception of a better remote EVPN route for that MAC. When the threshold is reached for a MAC address, this MAC address is put in hold-down state (this ‘hold-down’ state is described below). Range: <3..10>. Default value: 5.
- **window** — Identifies the timer within which a MAC is considered duplicate if it reaches the configured num-moves. Range: <1..15> minutes. Default value: 3 minutes.
- **Retry —** The timer after which the MAC in hold-down state is automatically flushed and the mac-duplication process starts again. This value is expected to be equal to two times or more than the window. If no retry is configured, this implies that, once mac-duplication is detected, MAC updates for that MAC will be held down until the user intervenes or a network event (that flushes the MAC) occurs. Range: <2..60> minutes. Default value: 9 minutes.
- **black-hole-dup-mac** — If enabled and a duplicate MAC address is detected, the router adds the MAC address to the duplicate MAC list and it programs the MAC in the FDB as a protected MAC associated with a black-hole (with type EvpnD:P and source ID “black-hole”)

When a MAC is considered a duplicate or in the ‘hold-down’ state, no further BGP advertisements are issued for this MAC and an alarm is triggered (by the first MAC in hold-down state). The following CLI output shows how PE-3 detects that MAC 00:00:01:01:01:01 is a duplicate (after reaching the **num-moves** in **window**) and the corresponding alarm.

```
142 2017/05/03 09:34:28.59 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 96
    Flag: 0x90 Type: 14 Len: 44 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.3
        Type: EVPN-MAC Len: 33 RD: 192.0.2.3:1 ESI: ESI-0, tag: 0, mac len: 48
                       mac: 00:00:01:01:01:01, IP len: 0, IP: NULL, label1: 1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:12
        bgp-tunnel-encap:VXLAN
        mac-mobility:Seq:5
"
```

Log 99 on PE-3 shows the following message when EVPN has detected a duplicate MAC address in VPLS 1:

```
2 2017/05/03 09:34:28.59 UTC MINOR: SVCMGR #2331 Base
"VPLS Service 1 has MAC(s) detected as duplicates by EVPN mac-duplication detect
ion."
```

The **show service id bgp-evpn** command shows the mac-duplication settings and the list of duplicate MACs on hold-down.

```
*A:PE-3# show service id 1 bgp-evpn

===============================================================================
BGP EVPN Table
===============================================================================
MAC Advertisement  : Enabled            Unknown MAC Route  : Enabled
CFM MAC Advertise  : Disabled
VXLAN Admin Status : Enabled            Creation Origin    : manual
MAC Dup Detn Moves : 3                  MAC Dup Detn Window: 1
MAC Dup Detn Retry : 2                  Number of Dup MACs : 2
MAC Dup Detn BH    : Disabled
IP Route Advert    : Disabled

EVI                : n/a
Ing Rep Inc McastAd: Enabled
Accept IVPLS Flush : Disabled
Send EVPN Encap    : Enabled

-------------------------------------------------------------------------------
Detected Duplicate MAC Addresses             Time Detected
-------------------------------------------------------------------------------
00:00:01:01:01:01                            05/03/2017 09:40:29
-------------------------------------------------------------------------------
===============================================================================
---snip---
```

SR OS stops sending and processing any BGP MAC advertisement routes for that MAC address until:

- The MAC is flushed due to a local event (SAP/SDP-binding associated to the MAC fails) or the reception of a remote withdraw for the MAC (due to a MAC flush at the remote 7x50) or
- The **retry <in_minutes>** timer expires, which flushes the MAC and restart the process.

When the last duplicate MAC address is removed from the duplicate list, log 99 on PE-3 will show the following message:

```
3 2017/05/03 09:44:28.60 UTC MINOR: SVCMGR #2332 Base
"VPLS Service 1 no longer has MAC(s) detected as duplicates by EVPN mac-duplication
detection."
```

EVPN also provides a mechanism to protect certain MACs that do not move for which connectivity must be guaranteed. These addresses must be protected in case there is an attempt to dynamically learn them in a different place in the EVPN-VXLAN VPLS service (on the same or different PE).

The protected MACs are configured in SR OS as conditional static MACs. A conditional static MAC defined in an EVPN-VXLAN VPLS service is advertised by BGP-EVPN as a static address. An example of the configuration of a conditional static MAC is as follows:

```
configure
    service
        vpls 1
            static-mac
                mac 00:00:05:05:05:05 create sap 1/2/1:1 monitor fwd-status
            exit
        exit
```

The protected MACs advertised in EVPN are shown in the receiving BGP RIB as Static (MAC mobility extended community with Sequence 0 and sticky bit set) and **EvpnS:P** (Evpn Static: Protected) in the FDB. The advertising PE shows the protected MAC as **CStatic:P** (Conditional Static: Protected) in the FDB:

On the advertising PE:

```
*A:PE-1# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         00:00:05:05:05:05 sap:1/2/1:1              CStatic: 05/03/17 10:06:08
                                                     P
---snip---
-------------------------------------------------------------------------------
No. of MAC Entries: 2
```

```
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
```

On the receiving PE:

```
*A:PE-3# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         00:00:05:05:05:05 vxlan:                   EvpnS    05/03/17 10:06:34
                                                     P
                            192.0.2.1:1
-------------------------------------------------------------------------------
No. of MAC Entries: 1
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================


*A:PE-3# show router bgp routes evpn mac mac-address 00:00:05:05:05:05 hunt
===============================================================================
 BGP Router ID:192.0.2.3       AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN MAC Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : N/A
Nexthop        : 192.0.2.1
From           : 192.0.2.1
Res. Nexthop   : 192.168.13.1
Local Pref.    : 100                    Interface Name : int-PE-3-PE-1
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : target:64500:12 bgp-tunnel-encap:VXLAN
                 mac-mobility:Seq:0/Static
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.1
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
EVPN type      : MAC
ESI            : ESI-0
Tag            : 0
IP Address     : N/A
```

```
Route Dist.     : 192.0.2.1:1
Mac Address     : 00:00:05:05:05:05
MPLS Label1     : VNI 1                 MPLS Label2     : N/A
Route Tag       : 0
Neighbor-AS     : N/A
Orig Validation: N/A
Source Class    : 0                     Dest Class      : 0
Add Paths Send : Default
Last Modified  : 02h34m57s

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-3#
```

The following procedures are supported in order to protect the configured static MAC addresses:

- All the SAP/SDP-bindings are internally configured as MAC protect restrict-protected-src as soon as BGP-EVPN is enabled in the VPLS service.
- Local static MACs or remote EVPN static MACs are considered as protected.
- If a frame with a source MAC address matching one of the protected MACs is received on a different SAP/SDP-binding than the owner of the protected MAC, the frame is discarded and an alarm triggered. This MAC protection is not performed for frames received on VXLAN bindings.
- The same throttled alarm mechanism used in MAC protect for restrict-protected-src with discard-frame is used here: the offending frames are captured to a list to be polled by the CPM every ~10min.

In this example, PE-3 has 00:00:05:05:05:05 in its FDB as EvpnS. If SAP 1/2/1:1 receives a frame with source MAC address 00:00:05:05:05:05, the frame is discarded and an alarm triggered:

```
4 2017/05/03 14:05:51.96 UTC MINOR: SVCMGR #2208 Base Slot 1
"Protected MAC 00:00:05:05:05:05 received on SAP 1/2/1:1 in service 1. "
```

Debug and Show Commands

In addition to the previously mentioned **show service id vxlan**, **show service id bgp-evpn** and **show service id fdb detail** commands, the following commands provide valuable information when troubleshooting an EVPN-VXLAN VPLS service.

The **show router bgp routes evpn** command supports filtering by route type as well as many other route fields.

```
*A:PE-3# show router bgp routes evpn
  - evpn <evpn-type>

      auto-disc       - Display BGP EVPN Auto-Disc Routes
      eth-seg         - Display BGP EVPN Eth-Seg Routes
      inclusive-mcast - Display BGP EVPN Inclusive-Mcast Routes
      ip-prefix       - Display BGP EVPN IPv4-Prefix Routes
      ipv6-prefix     - Display BGP EVPN IPv6-Prefix Routes
      mac             - Display BGP EVPN Mac Routes


*A:PE-3# show router bgp routes evpn mac
 - mac [hunt|detail] [rd <rd>] [next-hop <ip-address>] [mac-address <mac-address>]
       [community <comm-id>] [tag <tag>] [aspath-regex <reg-exp>]

 <hunt|detail>        : keywords
 <rd>                 : {<ip-addr:comm-val>|
                        <2byte-asnumber:ext-comm-val>|
                        <4byte-asnumber:comm-val>}
 <ip-address>         : a.b.c.d
 <mac-address>        : xx:xx:xx:xx:xx:xx or xx-xx-xx-xx-xx-xx
 <comm-id>            : <as-number1:comm-val1>|<ext-comm>|
                        <well-known-comm>
                        ext-comm        - <type>:{<ip-address:comm-val1>|
                                                  <as-number1:comm-val2>|
                                                  <as-number2:comm-val1>}
                        as-number1      - [0..65535]
                        comm-val1       - [0..65535]
                        type            - target|origin
                        ip-address      - a.b.c.d
                        comm-val2       - [0..4294967295]
                        as-number2      - [0..4294967295]
                        well-known-comm - null|no-export|no-export-subconfed|
                                          no-advertise
 <tag>                : [0..16777215] | MAX-ET
 <reg-exp>            : [80 chars max]


*A:PE-3# show router bgp routes evpn mac tag 0
===============================================================================
 BGP Router ID:192.0.2.3        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN MAC Routes
===============================================================================
Flag  Route Dist.         MacAddr           ESI
      Tag                 Mac Mobility      Ip Address
                                            NextHop
                                            Label1
-------------------------------------------------------------------------------
u*>i  192.0.2.1:1         00:00:05:05:05:05 ESI-0
      0                   Static            N/A
                                            192.0.2.1
                                            VNI 1
```

```
u*>i  192.0.2.2:1         00:00:00:00:00:00 ESI-0
      0                   Seq:0             N/A
                                            192.0.2.2
                                            VNI 1

u*>i  192.0.2.2:2         00:00:00:00:00:00 ESI-0
      0                   Seq:0             N/A
                                            192.0.2.2
                                            VNI 2

-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:PE-3#
```

The **tools dump service id vxlan** displays the number of times a service could not add a VXLAN binding or <VTEP, Egress VNI> due to the following limits:

- The per System VTEP limit has been reached
- The per System (egress VTEP, egress VNI) limit has been reached
- The per Service (egress VTEP, egress VNI) limit has been reached
- The per System Bind limit: Total bind limit or VXLAN bind limit has been reached.

Tools dump service evpn usage displays the consumed resources in the system, whereas **tools dump service vxlan dup-vtep-egrvni** displays the (egress VTEP, egress VNI) bindings that have been detected as duplicate attempts, in other words, an attempt to add the same binding to more than one service:

```
*A:PE-3# tools dump service id 1 vxlan

VTEP, Egress VNI Failure statistics at 05/03/2017 10:38:32:

statistics last cleared at 05/03/2017 06:09:14:

Failures: None


*A:PE-3# tools dump service id 1 evpn usage

Evpn Tunnel Interface IP Next Hop: N/A


*A:PE-3# tools dump service evpn usage

EVPN usage statistics at 05/03/2017 10:38:32:

MPLS-TEP                                        :             0
VXLAN-TEP                                       :             2
Total-TEP                                       :       2/ 16383

Mpls Dests (TEP, Egress Label + ES + ES-BMAC)   :             0
Mpls Etree Leaf Dests                           :             0
Vxlan Dests (TEP, Egress VNI)                   :             3
Total-Dest                                      :       3/196607
```

```
Sdp Bind +  Evpn Dests                            :      9/245759
ES L2/L3 PBR                                      :      0/ 32767
Evpn Etree Remote BUM Leaf Labels                 :             0

*A:PE-3# tools dump service vxlan dup-vtep-egrvni

Duplicate VTEP, Egress VNI usage attempts at 05/03/2017 10:38:32:

1. 192.0.2.1:100
```

Conclusion

SR OS supports the EVPN control plane for VXLAN tunnels terminated in VPLS services. VXLAN is an overlay IP tunneling mechanism that is being used in data center, data center interconnect and other applications. EVPN is a scalable and flexible control plane that provides control over the MAC addresses being learned and advertised, as well as other mechanisms to optimize Layer 2 services such as proxy-ARP, MAC mobility, MAC duplication detection and MAC protection. SR OS provides a resilient and scalable EVPN-VXLAN solution for Layer 2 services, including interoperability to existing VPLS networks. This chapter showed all of those functions and how they are configured and operated.

EVPN for VXLAN Tunnels (Layer 3)

This chapter provides information about EVPN for VXLAN tunnels (Layer 3).

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter is applicable to SR OS and was initially written for release 12.0.R4. The CLI in the current edition is based on release 15.0.R2. Ethernet Virtual Private Network (EVPN) is a control plane technology and does not have line card hardware dependencies.

Chapter EVPN for VXLAN Tunnels (Layer 2) example is prerequisite reading.

Overview

As discussed in the EVPN for VXLAN Tunnels (Layer 2) chapter, EVPN and VXLAN can be enabled on VPLS or R-VPLS services in SR OS. While that chapter focuses on the use of EVPN-VXLAN layer 2 services, in other words, how EVPN-VXLAN is configured in VPLS services, this chapter describes how EVPN-VXLAN can be used to provide inter-subnet forwarding in R-VPLS and VPRN services. Inter-subnet forwarding can be provided by regular R-VPLS and VPRN services. However, EVPN provides an efficient and unified way to populate Forwarding Databases (FDBs), Address Resolution Protocol (ARP) tables and routing tables using a single BGP address family. Inter-subnet forwarding in overlay networks would otherwise require data plane learning and the use of routing protocols on a per VPRN basis.

The SR OS solution for inter-subnet forwarding using EVPN is based on building blocks described in draft-sajassi-l2vpn-evpn-inter-subnet-forwarding and the use of the EVPN IP-prefix routes (route type-5) as explained in draft-rabadan-l2vpn-evpn-prefix-advertisement. This example describes three supported common scenarios and provides the CLI configuration and required tools to troubleshoot EVPN-VXLAN in each case. The scenarios configured and explained are:

- EVPN-VXLAN in R-VPLS services
- EVPN-VXLAN in Integrated Routing Bridging (IRB) backhaul R-VPLS services
- EVPN-VXLAN in EVPN tunnel R-VPLS services

In all these scenarios, redundant PEs are usually deployed. If that is the case, the interaction of EVPN, IP-VPN, and the Routing Table Manager (RTM) may lead to some routing loop situations that must be avoided by the use of routing policies (this also may happen in traditional IP-VPN deployments when eBGP and MP-BGP interact to populate VPRN routing tables in multi-homed networks). This chapter explains when those routing loops can happen and how to avoid them.

The term IRB interface refers to an R-VPLS service bound to a VPRN IP interface. The terms IRB interface and R-VPLS interface are used interchangeably throughout this chapter.

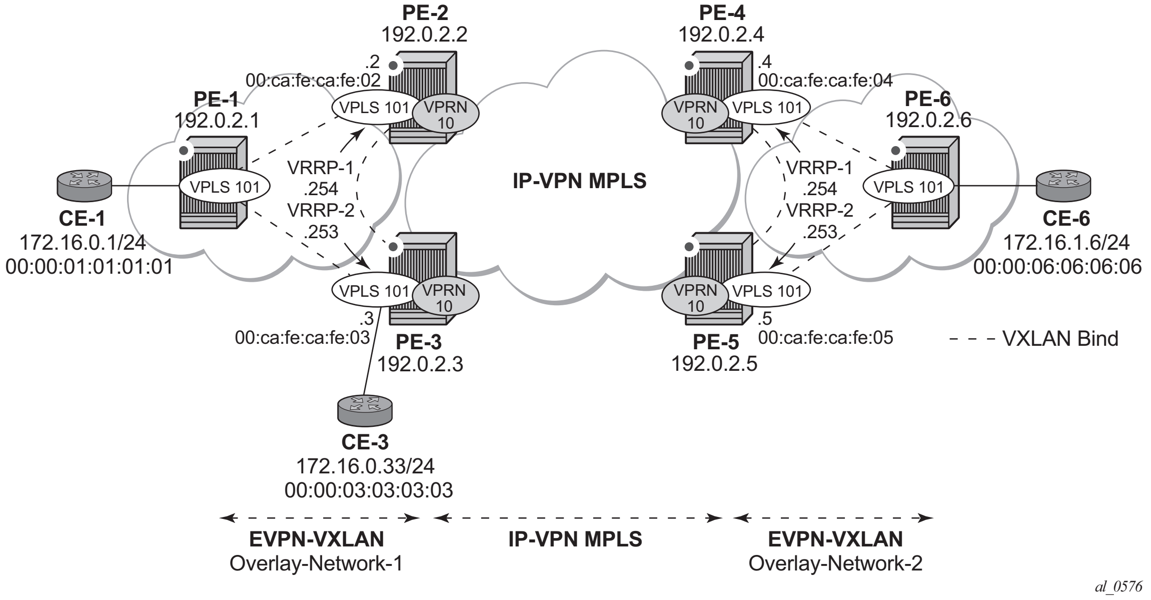
Configuration

This section describes the configuration of EVPN-VXLAN for Layer 3 services on SR OS, as well as the available troubleshooting and show commands. The three scenarios described in the overview are analyzed independently.

EVPN-VXLAN in an R-VPLS Service

Figure 61 shows the topology used in the first scenario.

Figure 61 **EVPN-VXLAN for R-VPLS Services**



The network topology shows two overlay (VXLAN) networks interconnected by an MPLS network:

- PE-1, PE-2, and PE-3 are part of Overlay-Network-1
- PE-4, PE-5, and PE-6 are part of Overlay-Network-2

A Layer 2/Layer 3 service is provided to a customer to connect CE-1, CE-3, and CE-6. In this scenario, Layer 2 connectivity is provided within each overlay network and inter-subnet connectivity (Layer 3) is provided between the overlay networks. VPLS 101 is defined within each overlay network and VPRN 10 connects both Layer 2 services through an IP-VPN MPLS network.

This topology can illustrate a Data Center Interconnect (DCI) example, where Overlay-Network-1 and Overlay-Network-2 are two data centers interconnected through an MPLS WAN. In this application, CE-1, CE-3, and CE-6 would simulate virtual machines or appliances, PE-2/3/4/5 would act as Data Center Gateways (DC GWs) and PE-1/6 as Network Virtualization Edge devices (or virtual PEs running on a compute infrastructure).

The following protocols and objects are configured beforehand:

- The ports interconnecting the six PEs in Figure 61 are configured as network ports (or hybrid) and will have router network interfaces defined in them. Only the ports connected to the CEs are configured as access ports.

- The six PEs are running IS-IS for the global routing table with the four core PEs interconnected using IS-IS Level-2 point-to-point interfaces and each overlay network using IS-IS Level-1 point-to-point interfaces.
- LDP is used as the MPLS protocol to signal transport tunnel labels among PE-2, PE-3, PE-4 and PE-5. There is no LDP running within each overlay network.
- The network port MTU (in all the ports sending/receiving VXLAN packets) must be at least 50-bytes (54 if dot1q encapsulation is used) greater than the service MTU in order to accommodate the size of the VXLAN header.

Once the IGP infrastructure and LDP in the core are enabled, BGP has to be configured. In this scenario, two BGP families have to be enabled: EVPN within each overlay-network for the exchange of MAC/IP addresses and setting up the flooding domains, and VPN-IPv4 among the four core PEs so that IP-prefixes can be exchanged and resolved to MPLS tunnels in the core.

As an example, the following CLI output shows the relevant BGP configuration of PE-1, which only needs the EVPN family. PE-6 has a similar BGP configuration, that is, only EVPN family is configured for its peers. The use of Route-Reflectors (RRs) in these type of scenarios is common. Although this scenario does not use RRs, an EVPN RR could have been used in Overlay-Network-1 and Overlay-Network-2 and a separate VPN-IPv4 RR could have been used in the core IP-VPN MPLS network.

```
configure
    router
        autonomous-system 64500
        bgp
            vpn-apply-import
            vpn-apply-export
            enable-peer-tracking
            rapid-withdrawal
            rapid-update evpn
            group "DC"
                family evpn
                peer-as 64500
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
            exit
```

The BGP configuration of PE-2 is as follows:

```
configure
    router
        autonomous-system 64500
        bgp
            vpn-apply-import
            vpn-apply-export
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            rapid-update evpn
            group "DC"
                family vpn-ipv4 evpn
                peer-as 64500
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.3
                exit
            exit
            group "WAN"
                family vpn-ipv4
                peer-as 64500
                neighbor 192.0.2.4
                exit
                neighbor 192.0.2.5
                exit
            exit
```

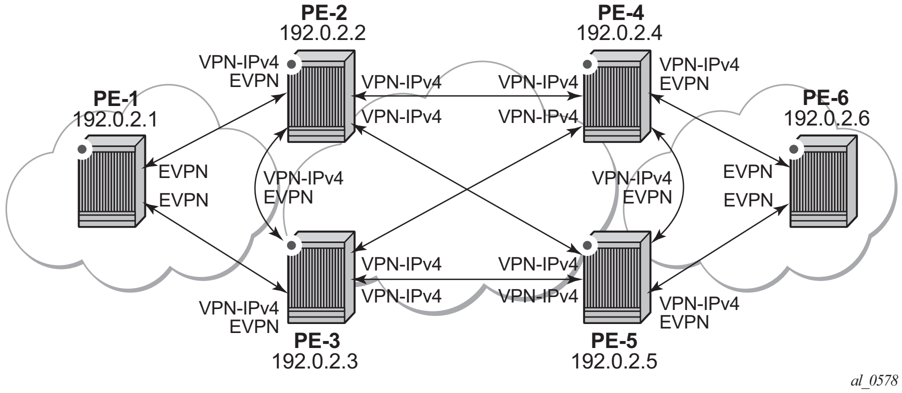
The BGP configuration on PE-3 is as follows:

```
configure
    router
        autonomous-system 64500
        bgp
            vpn-apply-import
            vpn-apply-export
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            rapid-update evpn
            group "DC"
                family vpn-ipv4 evpn
                peer-as 64500
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
            exit
            group "WAN"
                family vpn-ipv4
                peer-as 64500
                neighbor 192.0.2.4
                exit
                neighbor 192.0.2.5
                exit
            exit
```

PE-4 and PE-5 have an equivalent BGP configuration.

Figure 62 shows the BGP peering sessions among the PEs and the enabled BGP families. PE-1 and PE-6 only establish an EVPN peering session with their peers (only the EVPN family is enabled on both PEs, even if the peer PEs are VPN-IPv4 capable as well).

Figure 62 BGP adjacencies and enabled families



Once the network infrastructure is running properly, the actual service configuration, as illustrated in Figure 61, can be carried out. The following CLI shows the configuration for VPLS 101 and VPRN 10 in PE-1, PE-2, and PE-3. The other overlay network has a similar configuration.

On PE-1:

```
configure
    service
        vpls 101 customer 1 create
            vxlan vni 101 create
            exit
            bgp
                route-distinguisher 192.0.2.1:101
                route-target export target:64500:101 import target:64500:101
            exit
            bgp-evpn
                vxlan
                    no shutdown
                exit
            exit
            service-name "evi-101"
            sap 1/2/1:101 create
            exit
            proxy-arp
                no shutdown
            exit
            no shutdown
```

Proxy-ARP is disabled (default) on PE-2, as well as on the other core PEs:

```
configure
    service
        vpls 101 customer 1 create
            allow-ip-int-bind
            exit
            vxlan vni 101 create
            exit
            bgp
                route-distinguisher 192.0.2.2:101
                route-target export target:64500:101 import target:64500:101
            exit
            bgp-evpn
                vxlan
                    no shutdown
                exit
            exit
            service-name "evi-101"
            no shutdown
        exit
        vprn 10 customer 1 create
            ecmp 2
            route-distinguisher 192.0.2.2:10
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64500:10
            interface "int-1" create
                address 172.16.0.2/24
                mac 00:ca:fe:ca:fe:02
                vrrp 1
                    backup 172.16.0.254
                    priority 254
                    ping-reply
                    traceroute-reply
                    mac 00:ca:fe:ca:fe:54
                exit
                vrrp 2
                    backup 172.16.0.253
                    ping-reply
                    traceroute-reply
                    mac 00:ca:fe:ca:fe:53
                exit
                vpls "evi-101"
                exit
            exit
            no shutdown
```

On PE-3:

```
configure
    service
        vpls 101 customer 1 create
            allow-ip-int-bind
            exit
            vxlan vni 101 create
            exit
            bgp
                route-distinguisher 192.0.2.3:101
                route-target export target:64500:101 import target:64500:101
            exit
            bgp-evpn
                vxlan
            no shutdown
        exit
    exit
    service-name "evi-101"
    sap 1/2/1:101 create
    exit
    no shutdown
exit
vprn 10 customer 1 create
    ecmp 2
    route-distinguisher 192.0.2.3:10
    auto-bind-tunnel
        resolution-filter
            ldp
        exit
        resolution filter
    exit
    vrf-target target:64500:10
    interface "int-1" create
        address 172.16.0.3/24
        mac 00:ca:fe:ca:fe:03
        vrrp 1
            backup 172.16.0.254
            ping-reply
            traceroute-reply
            mac 00:ca:fe:ca:fe:54
        exit
        vrrp 2
            backup 172.16.0.253
            priority 254
            ping-reply
            traceroute-reply
            mac 00:ca:fe:ca:fe:53
        exit
        vpls "evi-101"
        exit
    exit
    no shutdown
```

For details about the EVPN and VXLAN configuration on PE-1 VPLS 101, see chapter EVPN for VXLAN Tunnels (Layer 2). The configuration of VPLS 101 on PE-2 and PE-3 has the following important aspects:

- The **allow-ip-int-bind** command is required so that the R-VPLS can be bound to VPRN 10.
- The **service-name** command is required and the configured name must match the name configured in the VPRN 10 VPLS interface.
- Even though EVPN and VXLAN are properly configured, proxy-ARP cannot be enabled in VPLS 101. In an R-VPLS with EVPN-VXLAN, proxy-ARP is not supported and the VPRN ARP table is used instead. When an EVPN MAC route that includes an IP address is received in an R-VPLS, the MAC-IP pair encoded in the route is added to the VPRN's ARP table, as opposed to the proxy-arp table.

```
*A:PE-2# configure service vpls 101 proxy-arp no shutdown
MINOR: SVCMGR #8007 Cannot modify proxy arp - service is routed
```

When configuring VPRN 10 on PE-2 and PE-3, the following considerations must be taken into account:

- When trying to enable existing VPRN features on interfaces linked to EVPN-VXLAN R-VPLS interfaces, the following commands are not supported:
  - arp-populate.
  - authentication-policy.

```
*A:PE-2# configure service vprn 10 interface "int-1" authentication-policy "authPol1"
INFO: PIP #1875 Cannot configure auth-policy on routed-vpls interface
```

- Dynamic routing protocols such as IS-IS, RIP, or OSPF are not supported.
- In general, no SR OS control plane generated packets are sent to the egress VXLAN bindings except for ARP, VRRP, ICMP, BFD, and Eth-CFM.
- As shown in Figure 61 and in the CLI excerpts, VRRP can be configured on the VPRN 10 VPLS interfaces to provide default gateway redundancy to the hosts connected to VPLS 101. Two VRRP instances are configured so that VPLS 101 upstream traffic can be load-balanced to PE-2 and PE-3. With VRRP on EVPN-VXLAN R-VPLS interfaces:
  - **Ping** and **traceroute** reply can be configured and are supported. BFD is also supported to speed up the fault detection.
  - **standby-forwarding**, even if it were configured for VRRP, would not have any effect in this configuration: the standby PE will never see any flooded traffic sent to it, therefore this command is not applicable to this scenario.
- When a VPRN 10 VPLS interface is bound to VPLS 101, EVPN advertises all the IP addresses configured for that VPLS interface as MAC routes with a static MAC indication. For the remote EVPN peers, that means that those MAC addresses linked to remote IP interfaces are protected. VRRP virtual IP/MACs are also advertised by EVPN as “static” and so protected. In the example of Figure 61, the VPLS 101 FDB in PE-1 shows the IP interface MACs and VRRP MACs as **EvpnS:P** (Static and protected MAC) as shown in the following output:

```
*A:PE-1# show service id 101 fdb detail

===============================================================================
Forwarding Database, Service 101
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
101       00:00:01:01:01:01 sap:1/2/1:101            L/0      05/03/17 12:19:04
101       00:ca:fe:ca:fe:02 vxlan:                   EvpnS    05/03/17 12:07:12
                                                     P
                            192.0.2.2:101
101       00:ca:fe:ca:fe:03 vxlan:                   EvpnS    05/03/17 12:07:24
```

```
                                                     P
                            192.0.2.3:101
101       00:ca:fe:ca:fe:53 vxlan:                   EvpnS    05/03/17 12:07:27
                                                     P
                            192.0.2.3:101
101       00:ca:fe:ca:fe:54 vxlan:                   EvpnS    05/03/17 12:07:15
                                                     P
                            192.0.2.2:101
-------------------------------------------------------------------------------
No. of MAC Entries: 5
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
```

The VPRN 10 VRRP instances on PE-2 are the following:

```
*A:PE-2# show router 10 vrrp instance

===============================================================================
VRRP Instances
===============================================================================
Interface Name                   VR Id Own Adm  State       Base Pri  Msg Int
                                 IP        Opr  Pol Id      InUse Pri Inh Int
-------------------------------------------------------------------------------
int-1                            1     No  Up   Master       254      1
                                 IPv4      Up   n/a         254       No
  Backup Addr: 172.16.0.254
int-1                            2     No  Up   Backup       100      1
                                 IPv4      Up   n/a         100       No
  Backup Addr: 172.16.0.253
-------------------------------------------------------------------------------
Instances : 2
===============================================================================
```

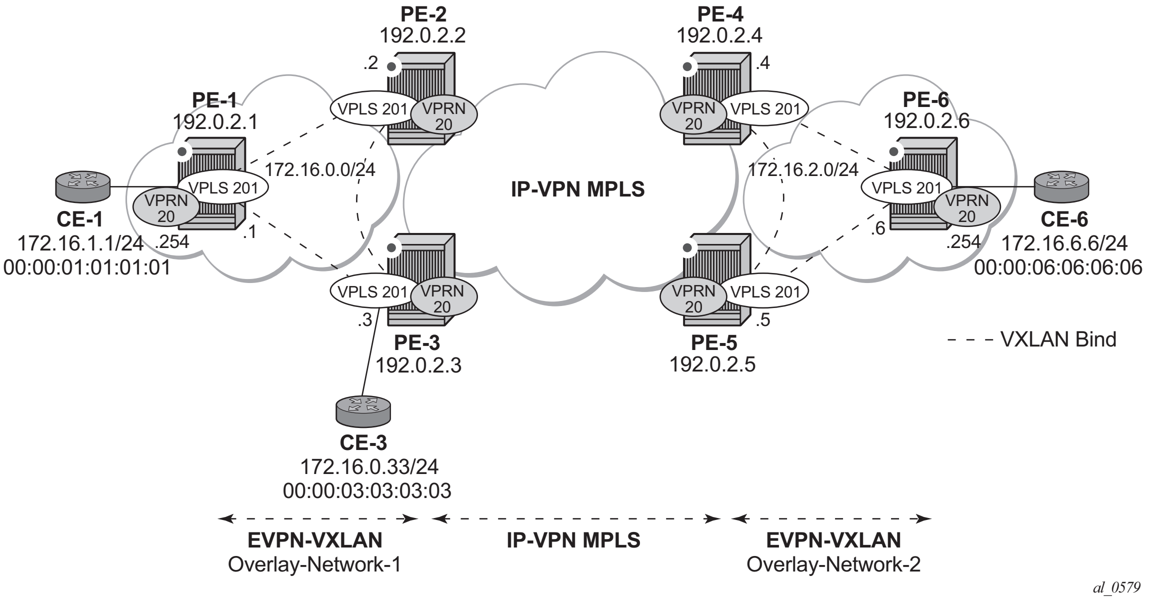
The ARP entries for PE-2 are the following:

```
*A:PE-2# show router 10 arp

===============================================================================
ARP Table (Service: 10)
===============================================================================
IP Address      MAC Address       Expiry    Type   Interface
-------------------------------------------------------------------------------
172.16.0.2      00:ca:fe:ca:fe:02 00h00m00s Oth[I] int-1
172.16.0.3      00:ca:fe:ca:fe:03 00h00m00s Evp[I] int-1
172.16.0.253    00:ca:fe:ca:fe:53 00h00m00s Oth    int-1
172.16.0.254    00:ca:fe:ca:fe:54 00h00m00s Oth[I] int-1
-------------------------------------------------------------------------------
No. of ARP Entries: 4
===============================================================================
```

EVPN-VXLAN in IRB Backhaul R-VPLS Services

Figure 63 illustrates the second inter-subnet forwarding scenario, where Layer 3 connectivity must be provided not only between the overlay networks but also within each overlay network. In the example shown in Figure 63, a customer (tenant) has different subnets and connectivity must be provided across all of them (CE-1, CE-3, and CE-6 must be able to communicate), bearing in mind that EVPN-VXLAN is enabled in each overlay network and IP-VPN MPLS is used to interconnect both overlay networks. VPLS 201 is an IRB Backhaul R-VPLS service because it provides connectivity to the VPRN instances.

Figure 63 **EVPN-VXLAN for IRB Backhaul R-VPLS Services**



From a BGP peering perspective, there is no change in this scenario compared to the previous one: PE-1 and PE-6 only support the EVPN address family. However, in this scenario, CE-1 is now connected to an R-VPLS directly linked to the VPRN instances in PE-2/PE-3. As a result of that, IP prefixes must be exchanged between PE-1 and PE-2/PE-3. EVPN is able to advertise not only MAC routes and Inclusive Multicast routes, but also IP prefix routes that contain IP prefixes that can be installed in the attached VPRN routing table.

As an example, the VPRN 20 and VPLS 201 configurations on PE-1, PE-2, and PE-3 are shown. Similar configurations are needed in PE-3, PE-4 and PE-6.

On PE-1, VPRN 20 is configured as follows:

```
configure
```

```
    service
        vprn 20 customer 1 create
            route-distinguisher 192.0.2.1:20
            vrf-target target:64500:20
            interface "int-evi-201" create
                address 172.16.0.1/24
                vpls "evi-201"
                exit
            exit
            interface "int-PE-1-CE-1" create
                address 172.16.1.254/24
                sap 1/2/1:20 create
                exit
            exit
            no shutdown
```

On PE-1, VPLS 201 is configured as follows:

```
configure
    service
        vpls 201 customer 1 create
            allow-ip-int-bind
            exit
            vxlan vni 201 create
            exit
            bgp
                route-distinguisher 192.0.2.1:201
                route-target export target:64500:201 import target:64500:201
            exit
            bgp-evpn
                ip-route-advertisement
                vxlan
                    no shutdown
                exit
            exit
            service-name "evi-201"
            no shutdown
```

On PE-2, VPRN 20 is configured with auto-bind-tunnel, as follows:

```
configure
    service
        vprn 20 customer 1 create
            route-distinguisher 192.0.2.2:20
            auto-bind-tunnel
                resolution any
            exit
            vrf-target target:64500:20
            interface "int-evi-201" create
                address 172.16.0.2/24
                vpls "evi-201"
                exit
            exit
            no shutdown
```

On PE-2, VPLS 201 is configured as follows:

```
configure
    service
        vpls 201 customer 1 create
            allow-ip-int-bind
            exit
            vxlan vni 201 create
            exit
            bgp
                route-distinguisher 192.0.2.2:201
                route-target export target:64500:201 import target:64500:201
            exit
            bgp-evpn
                ip-route-advertisement
                vxlan
                    no shutdown
                exit
            exit
            service-name "evi-201"
            no shutdown
```

On PE-3, VPRN 20 is configured with auto-bind-tunnel, as follows:

```
configure
    service
        vprn 20 customer 1 create
            route-distinguisher 192.0.2.3:20
            auto-bind-tunnel
                resolution any
            exit
            vrf-target target:64500:20
            interface "int-evi-201" create
                address 172.16.0.3/24
                vpls "evi-201"
                exit
            exit
            no shutdown
```

On PE-3, VPLS 201 is configured as follows:

```
configure
    service
        vpls 201 customer 1 create
            allow-ip-int-bind
            exit
            vxlan vni 201 create
            exit
            bgp
                route-distinguisher 192.0.2.3:201
                route-target export target:64500:201 import target:64500:201
            exit
            bgp-evpn
                ip-route-advertisement
                vxlan
                    no shutdown
                exit
            exit
            service-name "evi-201"
            sap 1/2/1:20 create
            exit
            no shutdown
```

As shown in the CLI excerpt, the configuration in the three nodes (PE-1/2/3) for VPLS 201 and VPRN 20 is very similar. The main difference is the **auto-bind-tunnel** command existing in PE-2/3's VPRN 20. This command allows the VPRN 20 on PE-2/3 to receive IP-VPN routes from the core and resolve them to MPLS tunnels. VPRN 20 on PE-1 does not require such command because all its IP prefixes are resolved to local interfaces or to EVPN peers.

The **ip-route-advertisement** command enables:

- The advertisement of IP prefixes in EVPN, in routes type 5. All the existing IP prefixes in the attached VPRN 20 routing table are advertised in EVPN within the VPLS 201 context (except for the ones associated to VPLS 201 itself).
- The installation of IP prefixes in the attached VPRN 20 routing table with a preference of 169 (BGP-VPN routes for IP-VPN have a preference of 170) and a next-hop of the gateway IP (GW IP) address included in the EVPN IP prefix route.

For instance, the following output shows that PE-1 advertises the IP prefix 172.16.1.0/24 as a EVPN route to PE-3 (similar route is sent to PE-2), captured by a **debug router bgp update** session.

```
4 2017/05/03 12:17:31.60 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 89
    Flag: 0x90 Type: 14 Len: 45 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.1
        Type: EVPN-IP-Prefix Len: 34 RD: 192.0.2.1:201, tag: 0,
                             ip_prefix: 172.16.1.0/24 gw_ip 172.16.0.1 Label: 201
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:64500:201
        bgp-tunnel-encap:VXLAN
"
```

The VPRN 20 routing table in PE-1 is as follows:

```
*A:PE-1# show router 20 route-table

===============================================================================
Route Table (Service: 20)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age         Pref
```

```
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.0.0/24                                 Local   Local     23h57m48s  0
       int-evi-201                                                   0
172.16.1.0/24                                 Local   Local     23h57m35s  0
       int-PE-1-CE-1                                                0
172.16.2.0/24                                 Remote  BGP EVPN  00h00m17s  169
       172.16.0.2                                                   0
172.16.6.0/24                                 Remote  BGP EVPN  00h00m17s  169
       172.16.0.2                                                   0
-------------------------------------------------------------------------------
No. of Routes: 4
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
```

The subnet 172.16.0.0/24 is used on the interfaces “int-evi-201” in overlay network 1 and subnet 172.16.2.0/24 is used on similar interfaces in overlay network 2. CE-1 has an IP address in subnet 172.16.1.0/24 and CE-6 has an IP address in subnet 172.16.6.0/24. The next hop to reach 172.16.2.0/24 (overlay network 2) or CE-6, is 172.16.0.2 (PE-2), but it could have been PE-3,

There is redundancy in the example setup and therefore, loops can occur. This is why routing policies need to be configured on the core PEs (PE-2, PE-3, PE-4, PE-5). These policies are described in section Use of Routing Policies to Avoid Routing Loops in Redundant PEs for routing loop use-case 1.

The routing table on PE-2 shows a BGP EVPN route toward CE-1 (subnet 172.16.1.0/24) via PE-1. The route toward CE-6 uses a tunnel toward PE-4 in overlay network 2.

```
*A:PE-2# show router 20 route-table

===============================================================================
Route Table (Service: 20)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.0.0/24                                 Local   Local     01d00h08m  0
       int-evi-201                                                   0
172.16.1.0/24                                 Remote  BGP EVPN  00h45m09s  169
       172.16.0.1                                                   0
172.16.2.0/24                                 Remote  BGP VPN   01d00h07m  170
       192.0.2.4  (tunneled)                                         0
172.16.6.0/24                                 Remote  BGP VPN   01d00h07m  170
       192.0.2.4  (tunneled)                                         0
-------------------------------------------------------------------------------
No. of Routes: 4
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
```

```
===============================================================================
```

The routing table on PE-3 is as follows:

```
*A:PE-3# show router 20 route-table

===============================================================================
Route Table (Service: 20)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.0.0/24                                 Local   Local     00h09m25s  0
       int-evi-201                                                  0
172.16.1.0/24                                 Remote  BGP EVPN  00h09m25s  169
       172.16.0.1                                                   0
172.16.2.0/24                                 Remote  BGP VPN   00h09m24s  170
       192.0.2.4 (tunneled)                                         0
172.16.6.0/24                                 Remote  BGP VPN   00h09m24s  170
       192.0.2.4 (tunneled)                                         0
-------------------------------------------------------------------------------
No. of Routes: 4
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-3#
```

When checking the operation of EVPN in this scenario, it is important to observe that the right next hops and prefixes are successfully installed in the VPRN 20 routing table:

- EVPN IP prefixes are sent using a GW-IP matching the primary IP interface address of the R-VPLS for which the routes are sent. For instance, as shown above, IP prefix 172.16.1.0/24 is advertised from PE-1 with GW-IP 172.16.0.1 (the IP address configured for the VPRN 20 VPLS interface in PE-1). In the PE-2/3 VPRN 20 routing tables, IP prefix 172.16.1.0/24 is installed with next hop 172.16.0.1. Traffic arriving at PE-2/3 on VPRN 20 with IP Destination Address (DA) in the 172.16.1.0 subnet matches the mentioned routing table entry. As usual, the next hop is resolved by the ARP table to a MAC and the MAC resolved by the FDB table to an egress VTEP, VNI.
- IP prefixes in the VPRN 20 routing table are advertised in IP-VPN to the remote IP-VPN MPLS peers. Received IP-VPN prefixes are installed in the VPRN 20 routing table using the remote PE system IP address as the next hop, as usual. For instance, 172.16.6.0/24 is installed in PE-2 VPRN 20’s routing table with next hop (tunneled) 192.0.2.4 and preference 170.

The following considerations of how the routing table manager (RTM) handles EVPN and IP-VPN prefixes must be taken into account:

- Only VPRN interface primary addresses are advertised as GW-IP in EVPN IP prefix routes. Secondary addresses are never sent as GW-IP addresses.
- EVPN IP prefixes are advertised by default as soon as the **ip-route-advertisement** command is enabled and there are active IP prefixes in the attached VPRN routing table.
- If the same IP prefix is received on a PE via EVPN and IP-VPN at the same time for the same VPRN, by default the EVPN prefix is selected because its preference (169) is better than the IP-VPN preference (170).
- Because EVPN has a better preference compared to IP-VPN, when the VPRNs on redundant PEs are attached to the same R-VPLS service, routing loops may occur. The use case described here is an example where routing loops can occur. Check Use of Routing Policies to Avoid Routing Loops in Redundant PEs to avoid routing loops in redundant PEs for more information.
- When the command **ip-route-advertisement** is enabled, the subnet IP prefixes are advertised in EVPN but not the “host” IP prefixes (/32 prefixes associated with the local interfaces). If the user wants to advertise the host IP prefixes as well, the **incl-host** keyword must be added to the **ip-route-advertisement** command. The following example illustrates this. The host routes can be shown with the **show router route-table all** command. When the **incl-host** keyword is added to PE-1’s VPLS 201, PE-1 advertises the host routes as well and these are installed in the remote PEs’ routing tables.

```
*A:PE-1# show router 20 route-table

===============================================================================
Route Table (Service: 20)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.0.0/24                                 Local   Local     01d04h18m  0
       int-evi-201                                                  0
172.16.1.0/24                                 Local   Local     01d04h17m  0
       int-PE-1-CE-1                                                0
172.16.2.0/24                                 Remote  BGP EVPN  04h20m31s  169
       172.16.0.2                                                   0
172.16.6.0/24                                 Remote  BGP EVPN  04h20m31s  169
       172.16.0.2                                                   0
-------------------------------------------------------------------------------
No. of Routes: 4
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================

*A:PE-1# show router 20 route-table all

===============================================================================
Route Table (Service: 20)
```

```
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                         Active     Metric
-------------------------------------------------------------------------------
172.16.0.0/24                                 Local   Local     01d04h18m  0
       int-evi-201                                     Y            0
172.16.0.1/32                                 Local   Host      01d04h18m  0
       int-evi-201                                     Y            0
172.16.1.0/24                                 Local   Local     01d04h17m  0
       int-PE-1-CE-1                                   Y            0
172.16.1.254/32                               Local   Host      01d04h17m  0
       int-PE-1-CE-1                                   Y            0
172.16.2.0/24                                 Remote  BGP EVPN  04h20m34s  169
       172.16.0.2                                      Y            0
172.16.6.0/24                                 Remote  BGP EVPN  04h20m34s  169
       172.16.0.2                                      Y            0
-------------------------------------------------------------------------------
No. of Routes: 6
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
       E = Inactive best-external BGP route
===============================================================================
```

```
*A:PE-1# configure service vpls 201 bgp-evpn ip-route-advertisement incl-host
```

```
*A:PE-2# show router 20 route-table

===============================================================================
Route Table (Service: 20)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.0.0/24                                 Local   Local     01d04h22m  0
       int-evi-201                                                  0
172.16.1.0/24                                 Remote  BGP EVPN  04h25m22s  169
       172.16.0.1                                                   0
172.16.1.254/32                               Remote  BGP EVPN  00h03m52s  169
       172.16.0.1                                                   0
172.16.2.0/24                                 Remote  BGP VPN   01d04h22m  170
       192.0.2.4  (tunneled)                                        0
172.16.6.0/24                                 Remote  BGP VPN   01d04h22m  170
       192.0.2.4  (tunneled)                                        0
-------------------------------------------------------------------------------
No. of Routes: 5
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-2#
```

- ECMP is fully supported for the VPRN for EVPN IP prefix routes coming from different GW-IP next-hops. However, ECMP is not supported for IP prefixes routes belonging to different owners (EVPN and IP-VPN). ECMP is enabled in VPRN 20 on PE-1, as follows:

```
*A:PE-1# configure service vprn 20 ecmp 2
```

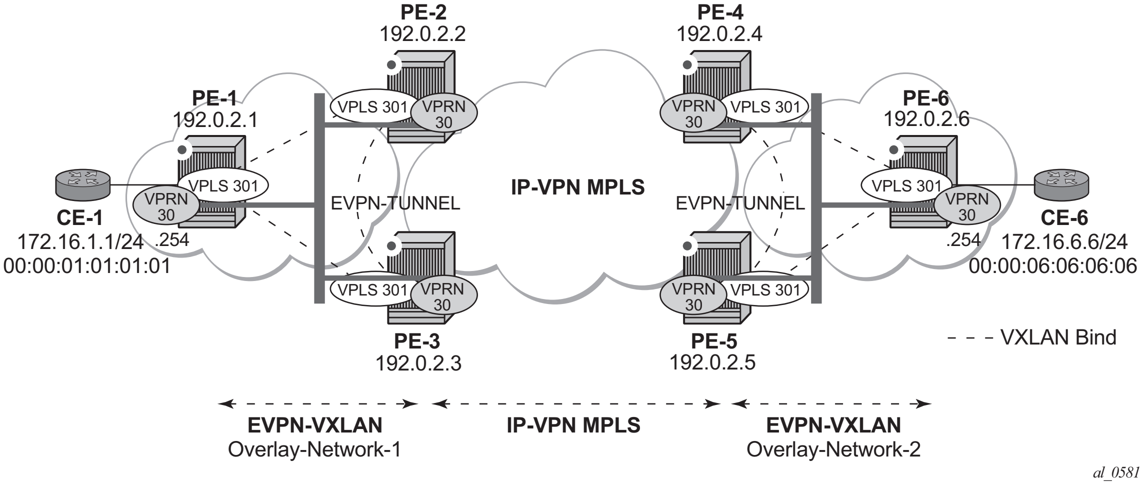
When policies are applied that prevent routing loops, as described in section Use of Routing Policies to Avoid Routing Loops in Redundant PEs, both PE-2 and PE-3 have IP-VPN tunnels for IP prefixes 172.16.2.0/24 and 172.16.6.0/24. In that case, an additional route with a different GW IP as next hop is installed in the routing table for these IP prefixes:

```
*A:PE-1# show router 20 route-table

===============================================================================
Route Table (Service: 20)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.0.0/24                                 Local   Local     00d00h10m  0
       int-evi-201                                                  0
172.16.1.0/24                                 Local   Local     00d00h10m  0
       int-PE-1-CE-1                                                0
172.16.2.0/24                                 Remote  BGP EVPN  00h00m01s  169
       172.16.0.2                                                   0
172.16.2.0/24                                 Remote  BGP EVPN  00h00m01s  169
       172.16.0.3                                                   0
172.16.6.0/24                                 Remote  BGP EVPN  00h00m01s  169
       172.16.0.2                                                   0
172.16.6.0/24                                 Remote  BGP EVPN  00h00m01s  169
       172.16.0.3                                                   0
-------------------------------------------------------------------------------
No. of Routes: 6
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
```

EVPN-VXLAN in EVPN Tunnel R-VPLS Services

The previous scenario shows how to use EVPN-VXLAN to provide inter-subnet forwarding for a tenant, where R-VPLS services can contain hosts and also offer transit services between VPRN instances. For example, in the use case shown in Figure 63, VPLS 201 in Overlay-Network-1 is an R-VPLS that can provide intra-subnet connectivity to all the hosts in subnet 172.16.0.0/24 (for example, CE-3 belongs to this subnet) but it can also provide “transit” or “backhaul” connectivity to hosts in subnet 172.16.1.0/24 (for example, CE-1) sending packets to subnets 172.16.2.0/24 or 172.16.6.0/24. In some cases, the R-VPLS where EVPN-VXLAN is enabled does not need to provide intra-subnet connectivity and it is purely a transit or backhaul service where VPRN IRB interfaces are connected. Figure 64 illustrates this use case.

Figure 64 **EVPN-VXLAN in EVPN-tunnel R-VPLS Services**



Compared to the use case in Figure 63, in this case the R-VPLS connecting the IRB interfaces in Overlay-Network-1 (VPLS 301) does not have any connected host. If that is the case, VPLS 301 can be configured as an EVPN tunnel.

EVPN tunnels are enabled using the **evpn-tunnel** command under the R-VPLS interface configured on the VPRN. EVPN tunnels bring the following benefits to EVPN-VXLAN IRB backhaul R-VPLS services:

- Easier and simpler provisioning of the tenant service: if an EVPN tunnel is configured in an IRB backhaul R-VPLS, there is no need to provision the IRB IP addresses in the VPRN. This makes the provisioning easier to automate and saves IP addresses from the tenant IP space.
- Higher scalability of the IRB backhaul R-VPLS: if EVPN tunnels are enabled, BUM traffic is suppressed in the EVPN-VXLAN IRB backhaul R-VPLS service (it is not required). As a result, the number of VXLAN bindings in IRB backhaul R-VPLS services with EVPN tunnels can be much higher.

As an example, the VPRN 30 and VPLS 301 configurations on PE-1, PE-2 and PE-3 are shown. Similar configurations are needed in PE-4, PE-5 and PE-6.

On PE-1:

```
configure
```

```
    service
        vprn 30 customer 1 create
            route-distinguisher 192.0.2.1:30
            vrf-target target:64500:30
            interface "int-PE-1-CE-1" create
                address 172.16.0.254/24
                sap 1/1/1:30 create
                exit
            exit
            interface "int-evi-301" create
                vpls "evi-301"
                    evpn-tunnel
                exit
            exit
            no shutdown

configure
    service
        vpls 301 customer 1 create
            allow-ip-int-bind
            exit
            vxlan vni 301 create
            exit
            bgp
                route-distinguisher 192.0.2.1:301
                route-target export target:64500:301 import target:64500:301
            exit
            bgp-evpn
                ip-route-advertisement
                vxlan
                    no shutdown
                exit
            exit
            service-name "evi-301"
            no shutdown
```

On PE-2:

```
configure
    service
        vprn 30 customer 1 create
            route-distinguisher 192.0.2.2:30
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64500:30
            interface "int-evi-301" create
                vpls "evi-301"
                    evpn-tunnel
                exit
            exit
            no shutdown

configure
```

```
    service
        vpls 301 customer 1 create
            allow-ip-int-bind
            exit
            vxlan vni 301 create
            exit
            bgp
                route-distinguisher 192.0.2.2:301
                route-target export target:64500:301 import target:64500:301
            exit
            bgp-evpn
                ip-route-advertisement
                vxlan
                    no shutdown
                exit
            exit
            service-name "evi-301"
            no shutdown
```

On PE-3:

```
configure
    service
        vprn 30 customer 1 create
            route-distinguisher 192.0.2.3:30
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64500:30
            interface "int-evi-301" create
                vpls "evi-301"
                    evpn-tunnel
                exit
            exit
            no shutdown
```

```
configure
    service
        vpls 301 customer 1 create
            allow-ip-int-bind
            exit
            vxlan vni 301 create
            exit
            bgp
                route-distinguisher 192.0.2.3:301
                route-target export target:64500:301 import target:64500:301
            exit
            bgp-evpn
                ip-route-advertisement
                vxlan
                    no shutdown
                exit
            exit
            service-name "evi-301"
            no shutdown
```

As shown in the preceding output, the configuration in the three nodes (PE-1/2/3) for VPLS 301 and VPRN 30 is similar to the configuration of VPLS 201 and VPRN 20 in the previous scenario, however, when the **evpn-tunnel** command is added to the VPRN interface, there is no need to configure an IP interface address. The option **evpn-tunnel** can be enabled independently of **ip-route-advertisement** (although no route-type 5 advertisements are sent in that case).

A VPRN supports regular IRB backhaul R-VPLS services as well as EVPN tunnel R-VPLS services. A maximum of eight R-VPLS services with **ip-route-advertisement** enabled per VPRN is supported (in any combination of regular IRB R-VPLS or EVPN tunnel R-VPLS services). EVPN tunnel R-VPLS services do not support SAPs or SDP-binds. No frames are flooded in an EVPN tunnel R-VPLS service, and, in fact no inclusive multicast routes are exchanged in R-VPLS services that are configured as EVPN tunnels.

The **show service id vxlan** command for an R-VPLS service configured as an EVPN tunnel shows <egress VTEP, VNI> bindings excluded from Mcast, in other words, the VXLAN bindings are not used to flood BUM traffic:

```
*A:PE-2# show service id 301 vxlan
===============================================================================
Vxlan Src Vtep IP: N/A
===============================================================================
VPLS VXLAN, Ingress VXLAN Network Id: 301
Creation Origin: manual
Assisted-Replication: none
RestProtSrcMacAct: none

===============================================================================
VPLS VXLAN service Network Specifics
===============================================================================
Ing Net QoS Policy : none                            Vxlan VNI Id     : 301
Ingress FP QGrp    : (none)                          Ing FP QGrp Inst : (none)

===============================================================================
Egress VTEP, VNI
===============================================================================
VTEP Address                            Egress VNI  Num. MACs   Mcast Oper  L2
                                                                      State PBR
-------------------------------------------------------------------------------
192.0.2.1                               301         1            -    Up    No
192.0.2.3                               301         1            -    Up    No
-------------------------------------------------------------------------------
Number of Egress VTEP, VNI : 2
-------------------------------------------------------------------------------
===============================================================================
```

The process followed upon receiving a route-type 5 on a regular IRB R-VPLS interface (previous scenario) differs from the one for an EVPN tunnel type (this scenario):

- IRB backhaul R-VPLS VPRN interface:

- When a route-type 2 that includes an IP address is received and it becomes active, the MAC/IP information is added to the FDB and ARP tables. This can be checked with the **show>router>arp** command and the **show>service>id>fdb detail** command.
- When a route-type 5 is received on (for instance) PE-2, and becomes active for the R-VPLS service, the IP prefix is added to the VPRN routing table regardless of the existence of a route-type 2 that can resolve the GW IP address. If a packet is received from the WAN side and the IP lookup hits an entry for which the GW IP (IP next-hop) does not have an active ARP entry, the system will ARP to get the MAC. If the ARP is resolved but the MAC is unknown in the FDB table, the system will flood the ARP message into the R-VPLS multicast list. Routes type 5 can be checked in the routing table with the **show>router>route-table** command and the **show>router>fib** command.

- EVPN tunnel R-VPLS VPRN interface:
  - When a route-type 2 is received and becomes active, the MAC address is added to the FDB (only). This MAC address is normally a GW-MAC.
  - When a route-type 5 is received on (for instance) PE-1, the system looks for the GW-MAC. The IP prefix is added to the VPRN routing table with next hop equal to EVPN-tunnel-GW-MAC; for example, ET-16:0a:ff:00:00:6a is an EVPN tunnel with GW-MAC 16:0a:ff:00:00:6a. The GW-MAC is added from the GW-MAC extended community sent along with the route-type 5 for prefix 172.16.6.0/24. If a packet is received from CE-1 and the IP lookup hits an entry for which the next hop is a EVPN tunnel:GW-MAC, the system looks up the GW-MAC in the FDB. Normally a route-type 2 with the GW-MAC has already been received so that the GW-MAC has been added to the FDB. If the GW-MAC is not present in the FDB, the packet will be dropped.
  - The IP prefixes with GW-MACs as next hops for the setup in Figure 64 are displayed in the show router route-table command, as follows:

```
*A:PE-1# show router 30 route-table

===============================================================================
Route Table (Service: 30)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age         Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.1.0/24                                 Local   Local     00h06m15s   0
       int-PE-1-CE-1                                               0
172.16.6.0/24                                 Remote  BGP EVPN  00h05m31s   169
       int-evi-301 (ET-16:0a:ff:00:00:6a)                           0
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
```

The same routing policies are applied on the core PEs to prevent loops; see Use of Routing Policies to Avoid Routing Loops in Redundant PEs.

The **show service id fdb detail** command can be used to look for the forwarding information for a GW MAC:

```
*A:PE-1# show service id 301 fdb detail

===============================================================================
Forwarding Database, Service 301
===============================================================================
ServId    MAC               Source-Identifier         Type     Last Change
                                                      Age
-------------------------------------------------------------------------------
301       16:09:ff:00:00:6a cpm                       Intf     05/03/17 13:00:25
301       16:0a:ff:00:00:6a vxlan:                    EvpnS    05/03/17 13:00:32
                                                      P
                            192.0.2.2:301
301       16:0b:ff:00:00:6a vxlan:                    EvpnS    05/03/17 13:00:38
                                                      P
                            192.0.2.3:301
-------------------------------------------------------------------------------
No. of MAC Entries: 3
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
```

IP prefix routes sent for EVPN tunnel R-VPLS services do not contain a GW-IP (the GW-IP will be zero) but convey a GW-MAC address that is used in the peer VPRN routing table. The following output shows PE-2's VPRN 30 interface MAC address sent to PE-1:

```
*A:PE-2# show router 30 interface "int-evi-301" detail | match MAC
MAC Address      : 16:0a:ff:00:00:6a    Mac Accounting    : Disabled
```

When ip-route-advertisement is enabled, PE-2 sends route type 5 messages to PE-1, as can be seen in the following BGP update for the route toward subnet 172.16.6.0/24 in overlay network 2, using the MAC as GW-MAC:

```
*A:PE-2# configure service vpls 301 bgp-evpn ip-route-advertisement


8 2017/05/03 12:04:54.36 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 97
    Flag: 0x90 Type: 14 Len: 45 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.2
        Type: EVPN-IP-Prefix Len: 34 RD: 192.0.2.2:301, tag: 0,
                             ip_prefix: 172.16.6.0/24 gw_ip 0.0.0.0 Label: 301
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:301
        mac-nh:16:0a:ff:00:00:6a
        bgp-tunnel-encap:VXLAN
"
```

In the VPRN 30 routing table, IP prefixes are shown with an EVPN tunnel next-hop (GW-MAC) as opposed to an IP next-hop, therefore, the user may think that no ARP entries are consumed by VPRN 30. However, internal ARP entries are still consumed in VPRN 30. Although not shown in the **show router 30 arp** command, the **summary** option shows the consumption of internal ARP entries for EVPN.

```
*A:PE-2# show router 30 route-table
===============================================================================
Route Table (Service: 30)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                  Metric
-------------------------------------------------------------------------------
172.16.1.0/24                                 Remote  BGP EVPN  00h31m34s  169
       int-evi-301 (ET-16:09:ff:00:00:6a)                         0
172.16.6.0/24                                 Remote  BGP VPN   00h59m09s  170
       192.0.2.4 (tunneled)                                      0
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
```

There are no entries in the ARP table:

```
*A:PE-2# show router 30 arp

===============================================================================
ARP Table (Service: 30)
===============================================================================
IP Address      MAC Address       Expiry    Type   Interface
-------------------------------------------------------------------------------
No Matching Entries Found
===============================================================================
```

One internal BGP-EVPN ARP entry is consumed, as can be seen as follows:

```
*A:PE-2# show router 30 arp summary

============================================================
ARP Table Summary (Service: 30)
============================================================
Local ARP Entries    : 1
```

```
Static ARP Entries   : 0
Dynamic ARP Entries  : 0
Managed ARP Entries  : 0
Internal ARP Entries : 0
BGP-EVPN ARP Entries : 1
------------------------------------------------------------
No. of ARP Entries   : 2
============================================================
```

The number of BGP-EVPN ARP entries in the **show router 30 arp summary** command matches the number of remote valid GW-MACs for VPRN 30.

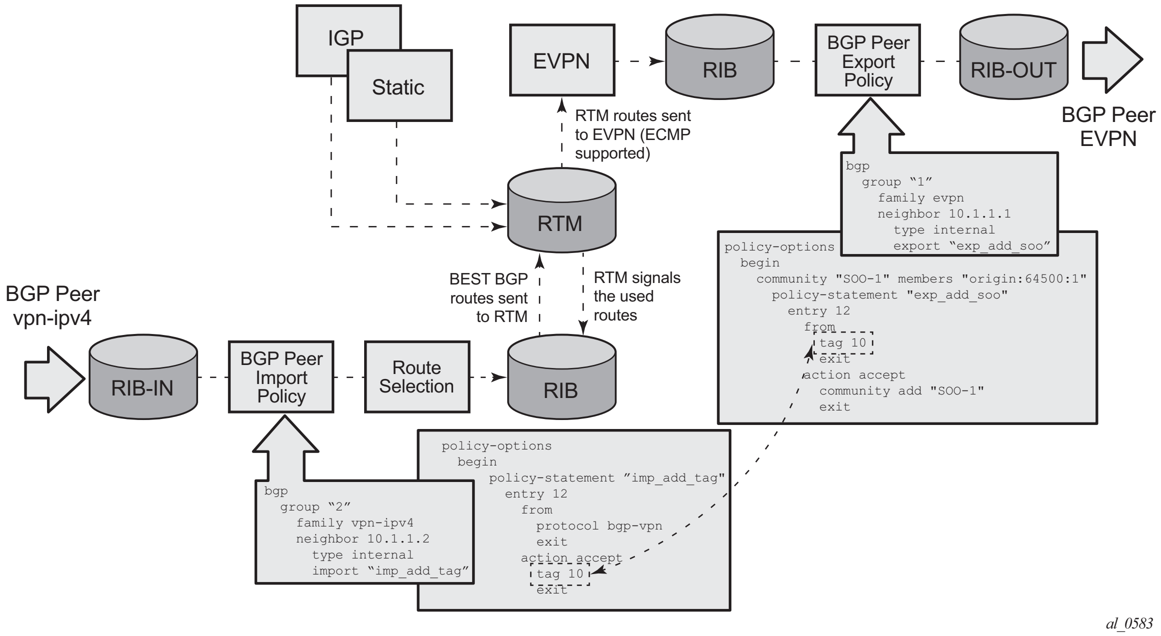
Routing Policies for IP Prefixes in EVPN

Routing policies are supported for IP prefixes imported/exported through BGP EVPN. The default import/export behavior for IP prefixes in EVPN can be modified by the use of routing policies applied either at peer level (**config>router>bgp>group/group>neighbor>import/export**) or VPLS level (**config>service>vpls>bgp>vsi-import/vsi-export**).

When applying routing policies to control the distribution of prefixes between EVPN and IP-VPN, the user must take into account that both families are completely separated as far as BGP is concerned and that when prefixes from a family are imported in the RTM, the BGP attributes are lost to the other family. The use of tags allows the controlled distribution of prefixes across the two families.

Figure 65 illustrates how VPN-IPv4 routes are imported into the RTM and then passed onto EVPN for its own processing. VPN-IPv4 routes can be tagged at ingress and this tag is preserved throughout the RTM and EVPN processing so that the tag can be “matched” by the egress BGP routing policy. In this particular example, egress EVPN routes matching tag 10, are modified to add a site-of-origin community origin:64500:1.

Figure 65 Routing Policies for Egress EVPN Routes



```
bgp
  group “1”
    family evpn
    neighbor 10.1.1.1
      type internal
      export “exp_add_soo”
```

```
policy-options
  begin
    community "SOO-1" members "origin:64500:1"
      policy-statement "exp_add_soo"
        entry 12
          from
            tag 10
            exit
          action accept
            community add "SOO-1"
            exit
```

```
bgp
  group “2”
    family vpn-ipv4
    neighbor 10.1.1.2
      type internal
      import “imp_add_tag”
```

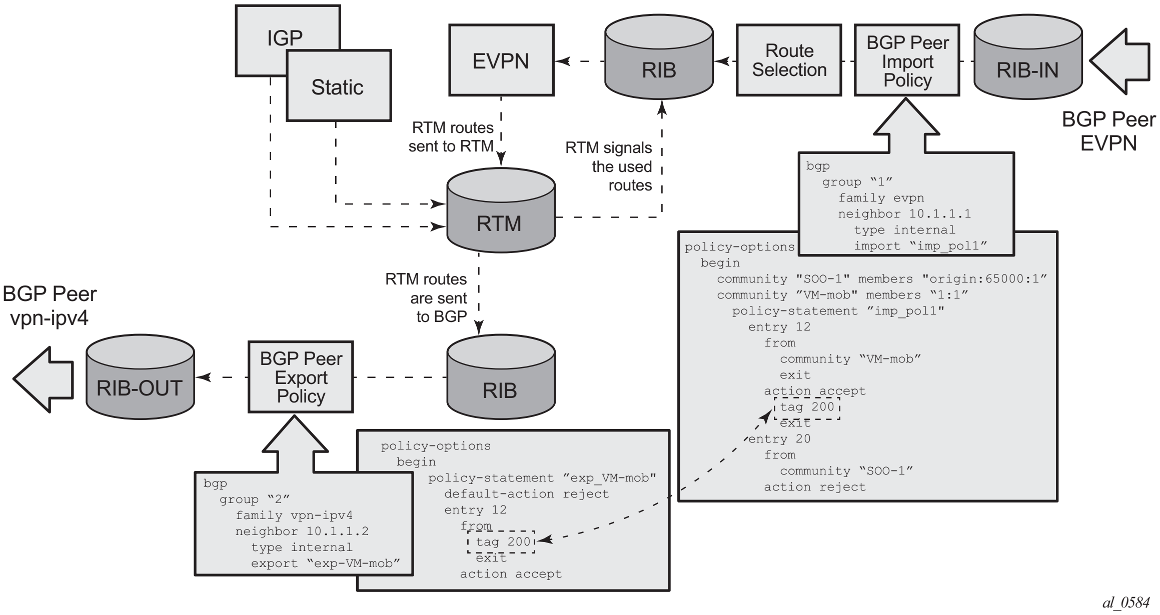
```
policy-options
  begin
      policy-statement ”imp_add_tag"
        entry 12
          from
            protocol bgp-vpn
            exit
          action accept
            tag 10
            exit
```

Policy tags can be used to match EVPN IP-prefixes that were learned not only from BGP VPN-IPv4 but also from other routing protocols. The tag range supported for each protocol is different:

```
<tag>   : accepts in decimal or hex
         [0x1..0xFFFFFFFF]H (for OSPF and ISIS)
         [0x1..0xFFFF]H (for RIP)
         [0x1..0xFF]H (for BGP)
```

Figure 66 illustrates the reverse workflow: routes imported from EVPN and exported from RTM to BGPVPN-IPv4. In this example, EVPN routes received with community VM-mob are tagged with TAG 200. At the egress VPN-IPv4 peers, only the routes with TAG 200 are advertised.

Figure 66 **Routing Policies for Ingress EVPN Routes**



```
bgp
  group "1"
    family evpn
    neighbor 10.1.1.1
      type internal
      import "imp_pol1"
```

```
policy-options
  begin
    community "SOO-1" members "origin:65000:1"
    community "VM-mob" members "1:1"
      policy-statement "imp_pol1"
        entry 12
          from
            community "VM-mob"
            exit
          action accept
            tag 200
            exit
        entry 20
          from
            community "SOO-1"
          action reject
```

```
bgp
  group "2"
    family vpn-ipv4
    neighbor 10.1.1.2
      type internal
      export "exp-VM-mob"
```

```
policy-options
  begin
      policy-statement "exp_VM-mob"
        default-action reject
        entry 12
          from
           tag 200
           exit
          action accept
```

The preceding behavior and the use of tags is also valid for **vsi-import** and **vsi-export** policies. The behavior can be summarized in the following statements:

- For EVPN prefix routes received and imported in RTM:
  - Routes can be matched on communities and tags can be added to them. This works at peer level or vsi-import level.
  - Well-known communities (no-export|no-export-subconfed|no-advertise) also require that the routing policies add a tag if the user wants to modify the behavior when exporting to BGP.
  - Routes can be matched based on family EVPN.
  - Routes cannot be matched on prefix-list.
- For exporting RTM to EVPN prefix routes:
  - Routes can be matched on tags and based on that, communities added, or routes accepted or rejected (dropped), etc. This works at peer level or vsi-export level.
  - Tags can be added for static-routes, RIP, OSPF, IS-IS, and BGP and then be matched in the vsi-export policy for EVPN IP-prefix route advertisement.
  - Tags cannot be added for direct routes.

Use of Routing Policies to Avoid Routing Loops in Redundant PEs

When redundant PE VPRN instances are connected to the same R-VPLS service (IRB backhaul or EVPN tunnel R-VPLS) with the ip-route-advertisement command enabled, routing loops can occur in two different use-cases:

1. Routing loop caused by EVPN and IP-VPN interaction in the RTM.
2. Routing loop caused by EVPN in “parallel” R-VPLS services.

Policy configuration examples for both cases are provided in the following sections.

Routing loop use-case 1: EVPN and IP-VPN interaction

This use case refers to scenarios with redundant PEs and VPRNs connected to the same R-VPLS with **ip-route-advertisement**. The scenarios in Figure 63 (EVPN-VXLAN for IRB Backhaul R-VPLS services) and Figure 64 (EVPN-VXLAN in EVPN tunnel R-VPLS services) are examples of this use case. In both scenarios, the following process causes a routing loop:

1. IP prefix 172.16.6.0/24 is advertised by PE-4 to PE-2 and PE-3.
2. PE-2 imports that prefix in the VPRN routing table and re-advertises the IP prefix in EVPN to PE-1 and PE-3 (the same thing happens in PE-3).
3. PE-3 already has the 172.16.6.0/24 prefix in the VPRN routing table with preference 170 (IP-VPN) but because it receives the IP prefix from EVPN with lower preference (169), the RTM will install the EVPN prefix in the VPRN routing table (the same thing happens in PE-2).
4. PE-3 advertises the EVPN learned IP prefix to all MP-BGP VPN-IPv4 peers (also PE-2).
5. PE-2 receives the IP prefix again from PE-3 and will advertise it in EVPN again, creating a routing loop (PE-3 will do the same thing as well).

This routing loop also happens in traditional multi-homed IP-VPN scenarios where the PE-CE eBGP and MP-BGP VPN-IPv4/v6 protocols interact in the same VPRN RTM, with different router preferences. In either case (EVPN or eBGP interaction with MP-BGP) the issue can be solved by the use of routing policies and site-of-origin communities.

Routing policies are applied to PE-2 and PE-3 (also to PE-4 and PE-5) and allow the redundant PEs to reject their own generated routes in order to avoid the loops. These routing policies can be applied at vsi-import/export level or BGP group/neighbor level. The following output shows an example of routing policies applied at BGP neighbor level for PE-2 (similar policies are applied on PE-3/4/5). Neighbor or group level policies are the preferred way in this kind of use case: a single set of policies is sufficient, as opposed to a set of policies per service (if the policies are applied at vsi-import/export level).

The following policies are applied in the BGP group or BGP group/neighbor context on PE-2:

```
configure
    router
        bgp
            group "DC"
                neighbor 192.0.2.1
                    import "add-tag_to_bgp-evpn_routes"
                exit
                neighbor 192.0.2.3
                    import "reject_based_on_SOO"
                    export "add-SOO_on_export"
                exit
            exit
            group "WAN"
                import "add-tag_to_bgp-vpn_routes"
            exit
```

The routing policies are configured as follows on PE-2:

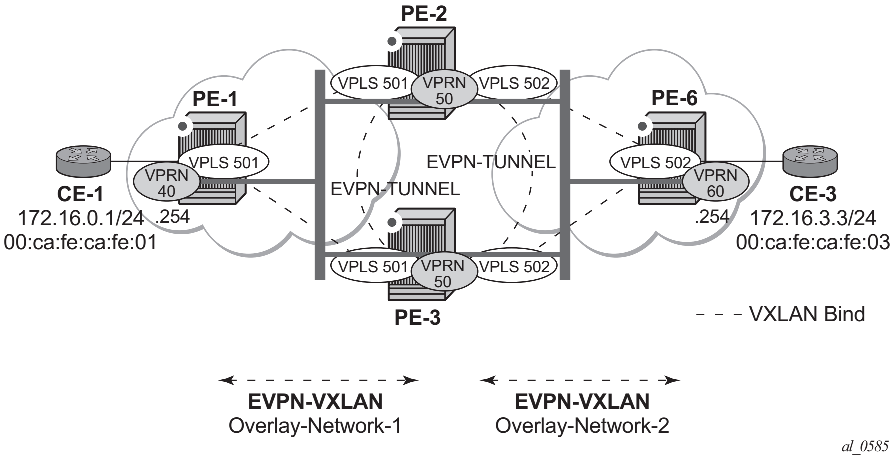
```
configure
    router
        policy-options
            begin
            community "SOO-PE-2" members "origin:2:1"
            community "SOO-PE-3" members "origin:3:1"
            policy-statement "add-SOO_on_export"
                entry 10
                    from
                        tag 2
                    exit
                    action accept
                        community add "SOO-PE-2"
                    exit
                exit
                entry 20
                    from
                        tag 3
                    exit
                    action accept
                        community add "SOO-PE-3"
                    exit
                exit
            exit
            policy-statement "reject_based_on_SOO"
                entry 10
                    from
                        community "SOO-PE-2"
                    exit
                    action drop
                    exit
                exit
                entry 20
                    from
                        community "SOO-PE-3"
                    exit
                    action drop
                    exit
                exit
            exit
            policy-statement "add-tag_to_bgp-vpn_routes"
                entry 10
                    from
                        protocol bgp-vpn
                    exit
                    action accept
                        tag 2
                    exit
                exit
            exit
            policy-statement "add-tag_to_bgp-evpn_routes"
                entry 10
                    from
                        family evpn
                    exit
                    action accept
                        tag 2
                    exit
                exit
            exit
            commit
```

EVPN and MP-BGP routes are tagged at import; on export, a site-of-origin community is added. Routes exchanged between the two redundant PEs are dropped if they are received by a PE with its own site-of-origin.

Routing loop use-case 2: EVPN in parallel R-VPLS services

If a VPRN is connected to more than one R-VPLS with **ip-route-advertisement** enabled, IP prefixes that belong to one R-VPLS are advertised into the other R-VPLS and vice versa. When redundant PEs are used, a routing loop will occur. Figure 67 illustrates this use case. The example shows R-VPLS with an EVPN tunnel configuration, but the same routing loop occurs for regular IRB backhaul R-VPLS services.

Figure 67 **EVPN in Parallel R-VPLS Services**



The configuration of VPRN 50 as well as VPLS 501/502 and the required policies are as follows. For this use case, policies must be applied at vsi-import/export level because more granularity is required when modifying the imported/exported routes.

On PE-2:

```
configure
    service
        vprn 50 customer 1 create
            route-distinguisher 192.0.2.2:50
            interface "int-evi-501" create
                vpls "evi-501"
                    evpn-tunnel
                exit
            exit
            interface "int-evi-502" create
                vpls "evi-502"
                    evpn-tunnel
                exit
            exit
            no shutdown
```

```
configure
    service
        vpls 501 customer 1 create
            allow-ip-int-bind
            exit
            vxlan vni 501 create
            exit
            bgp
                route-distinguisher 192.0.2.2:501
                vsi-export "vsi-export-policy-501"
                vsi-import "vsi-import-policy-501"
            exit
            bgp-evpn
                ip-route-advertisement
                vxlan
                    no shutdown
                exit
            exit
            service-name "evi-501"
            no shutdown

configure
    service
        vpls 502 customer 1 create
            allow-ip-int-bind
            exit
            vxlan vni 502 create
            exit
            bgp
                route-distinguisher 192.0.2.2:502
                vsi-export "vsi-export-policy-502"
                vsi-import "vsi-import-policy-502"
            exit
            bgp-evpn
                ip-route-advertisement
                vxlan
                    no shutdown
                exit
            exit
            service-name "evi-502"
            no shutdown

configure
    router
        policy-options
            begin
            community "exp_RVPLS501" members "origin:2:11" "target:64500:501"
            community "exp_RVPLS502" members "origin:2:11" "target:64500:502"
            community "SOO-PE-2-RVPLS" members "origin:2:11"
            community "SOO-PE-3-RVPLS" members "origin:3:11"
            community "SOO_PE-3_RVPLS501" members "origin:3:11" "target:64500:501"
            community "SOO_PE-3_RVPLS502" members "origin:3:11" "target:64500:502"
            policy-statement "vsi-export-policy-501"
                entry 10
                    from
                        tag 12
                    exit
                    action accept
                        community add "SOO_PE-3_RVPLS501"
                    exit
                exit
                entry 20
                    action accept
                        community add "exp_RVPLS501"
                    exit
                exit
            exit
            policy-statement "vsi-export-policy-502"
                entry 10
                    from
                        tag 12
                    exit
                    action accept
                        community add "SOO_PE-3_RVPLS502"
                    exit
                exit
                entry 20
                    action accept
                        community add "exp_RVPLS502"
                    exit
                exit
            exit
            policy-statement "vsi-import-policy-501"
                entry 10
                    from
                        community "SOO-PE-2-RVPLS"
                    exit
                    action drop
                    exit
                exit
                entry 20
                    from
                        community "SOO_PE-3_RVPLS501"
                    exit
                    action accept
                        tag 12
                    exit
                exit
                default-action accept
                exit
            exit
            policy-statement "vsi-import-policy-502"
                entry 10
                    from
                        community "SOO-PE-2-RVPLS"
                    exit
                    action drop
                    exit
                exit
                entry 20
                    from
                        community "SOO_PE-3_RVPLS502"
                    exit
                    action accept
                        tag 12
                    exit
                exit
                default-action accept
                exit
            exit
            commit
```

Troubleshooting and Debug Commands

For general information on EVPN and VXLAN troubleshooting and debug commands, see chapter EVPN for VXLAN Tunnels (Layer 2). The following information focuses on specific commands for Layer-3 applications.

When troubleshooting and operating a EVPN-VXLAN scenario with inter-subnet forwarding, it is important to check the IP prefixes and next-hops, as well as ARP tables and FDB tables (**show router x route-table**, **show router x arp**, **show service id y fdb detail**).

ICMP commands can also help checking the connectivity. When traceroute is used on EVPN-VXLAN in EVPN tunnel interfaces, EVPN tunnel interface hops in the traceroute commands are showing the VPRN loopback address or the other non evpn-tunnel interface address. In VPRN services where all of the interfaces are of type EVPN tunnel, ICMP packets fail until an IP address is configured. The following output shows a traceroute from VPRN 30 in PE-1 to CE-6 and from PE-2 to CE-1 (see Figure 64):

```
*A:PE-1# traceroute router 30 172.16.6.6
traceroute to 172.16.6.6, 30 hops max, 40 byte packets
  1  0.0.0.0  * * *
  2  0.0.0.0  * * *
  3  172.16.6.254 (172.16.6.254)    2.33 ms  2.21 ms  2.38 ms
  4  172.16.6.6 (172.16.6.6)    2.72 ms  2.97 ms  2.88 ms


*A:PE-2# traceroute router 30 172.16.1.1
traceroute to 172.16.0.1, 30 hops max, 0 byte packets

Send failed. Unable to find local ip address
```

When troubleshooting R-VPLS services, specifically R-VPLS services configured as EVPN tunnels, the limit of peer PEs per EVPN tunnel service is much higher than for a regular R-VPLS service because the egress <VTEP, VNI> bindings do not have to be added to the multicast flooding list. For this reason, the following **tools dump** command has been added to check the consumed/total EVPN tunnel next hops. The number of EVPN tunnel next hops matches the number of remote GW-MAC addresses per EVPN tunnel R-VPLS service.

```
*A:PE-1# tools dump service id 501 evpn usage

Evpn Tunnel Interface IP Next Hop: 2/8189
```

Finally, when troubleshooting EVPN routes and routing policies, the **show router bgp routes evpn** command and its filters can help:

- Check that the expected routes are received, properly imported and communities/tags added/replaced/removed.

- Check that the expected routes are sent, properly exported and communities added/replaced/removed.

Examples of EVPN IP prefix routes including communities and tags are the following.

```
*A:PE-2# show router bgp routes evpn ?
  - evpn <evpn-type>

      auto-disc       - Display BGP EVPN Auto-Disc Routes
      eth-seg         - Display BGP EVPN Eth-Seg Routes
      inclusive-mcast - Display BGP EVPN Inclusive-Mcast Routes
      ip-prefix       - Display BGP EVPN IPv4-Prefix Routes
      ipv6-prefix     - Display BGP EVPN IPv6-Prefix Routes
      mac             - Display BGP EVPN Mac Routes


*A:PE-2# show router bgp routes evpn ip-prefix
  - ip-prefix [hunt|detail] [rd <rd>] [prefix <ip-prefix/ip-prefix-length>]
            [community <comm-id>] [tag <tag>] [next-hop <ip-address>]
            [aspath-regex <reg-exp>]

 <hunt|detail>        : keywords
 <rd>                 : {<ip-addr:comm-val>|
                        <2byte-asnumber:ext-comm-val>|
                        <4byte-asnumber:comm-val>}
 <ip-prefix/ip-pref*> : ip-address     - a.b.c.d (host bits must be 0)
                        mask           - [0..32]
 <comm-id>            : <as-number1:comm-val1>|<ext-comm>|
                        <well-known-comm>
                        ext-comm        - <type>:{<ip-address:comm-val1>|
                                                  <as-number1:comm-val2>|
                                                  <as-number2:comm-val1>}
                        as-number1      - [0..65535]
                        comm-val1       - [0..65535]
                        type            - target|origin
                        ip-address      - a.b.c.d
                        comm-val2       - [0..4294967295]
                        as-number2      - [0..4294967295]
                        well-known-comm - null|no-export|no-export-subconfed|
                                          no-advertise
 <tag>                : [0..16777215] | MAX-ET
 <next-hop>           : ipv4-address   - a.b.c.d
                        ipv6-address   - x:x:x:x:x:x:x:x   (eight 16-bit pieces)
                                         x:x:x:x:x:x:d.d.d.d
                                         x - [0..FFFF]H
                                         d - [0..255]D
 <reg-exp>            : [80 chars max]
```

Community origin:2:11 target:64500:502 is added to the outgoing routes (by the routing policy “vsi-export-policy-502”), as can be verified as follows:

```
*A:PE-2# show router bgp routes evpn ip-prefix hunt prefix 172.16.1.0/24
================================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
================================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
```

```
                l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN IP-Prefix Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
---snip---
-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
Network        : N/A
Nexthop        : 192.0.2.2
To             : 192.0.2.1
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : origin:2:11 target:64500:502
                 mac-nh:16:0a:ff:00:01:33 bgp-tunnel-encap:VXLAN
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.1
Origin         : IGP
AS-Path        : No As-Path
EVPN type      : IP-PREFIX
ESI            : N/A
Tag            : 0
Gateway Address: 16:0a:ff:00:01:33
Prefix         : 172.16.1.0/24
Route Dist.    : 192.0.2.2:502
MPLS Label     : VNI 502
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0

---snip---
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-2#
```

Route tag 2 is added by PE-2 to all BGP EVPN routes (according to policy “add-tag_to_bgp-evpn_routes”), as can be verified in the following output:

```
*A:PE-2# show router bgp routes evpn ip-prefix prefix 172.16.1.0/24 detail
===============================================================================
 BGP Router ID:192.0.2.2       AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP EVPN IP-Prefix Routes
===============================================================================
-------------------------------------------------------------------------------
Original Attributes

Network        : N/A
Nexthop        : 192.0.2.1
From           : 192.0.2.1
Res. Nexthop   : 192.168.12.1
Local Pref.    : 100                    Interface Name : int-PE-2-PE-1
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : target:64500:201 bgp-tunnel-encap:VXLAN
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.1
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
EVPN type      : IP-PREFIX
ESI            : N/A
Tag            : 0
Gateway Address: 172.16.0.1
Prefix         : 172.16.1.0/24
Route Dist.    : 192.0.2.1:201
MPLS Label     : VNI 201
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h06m56s

Modified Attributes

Network        : N/A
Nexthop        : 192.0.2.1
From           : 192.0.2.1
Res. Nexthop   : 192.168.12.1
Local Pref.    : 100                    Interface Name : int-PE-2-PE-1
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : target:64500:201 bgp-tunnel-encap:VXLAN
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.1
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
EVPN type      : IP-PREFIX
ESI            : N/A
Tag            : 0
Gateway Address: 172.16.0.1
Prefix         : 172.16.1.0/24
Route Dist.    : 192.0.2.1:201
MPLS Label     : VNI 201
Route Tag      : 2
```

```
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h06m57s
------------------------------------------------------------------------------
---snip---
```

Conclusion

SR OS supports not only the EVPN control plane for VXLAN tunnels in Layer 2 applications but also the simultaneous use of EVPN and VXLAN for VPN customers (tenants) with intra and inter-subnet connectivity requirements. R-VPLS services can be configured to provide default gateway connectivity to hosts, IRB backhaul connectivity to VPRN services and EVPN tunnel connectivity to VPRN services. When configured to do so, EVPN can advertise IP prefixes and interact with the VPRN RTM to propagate IP prefix connectivity between EVPN and other routing protocols in the VPRN, including IP-VPN. This example has shown how to configure R-VPLS services for all these functions, as well as how to configure routing policies for EVPN-based IP prefixes.

EVPN Interconnect Ethernet Segments

This chapter provides information about EVPN Interconnect Ethernet Segments.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

The information and configuration in this chapter are based on SR OS Release 15.0.R4.

Chapters EVPN for MPLS Tunnels, EVPN for VXLAN Tunnels (Layer 2) and EVPN-MPLS Interconnect for EVPN-VXLAN VPLS Services are prerequisite reading.

Overview

SR OS supports Interconnect Ethernet Segments (I-ESs) for VXLAN as per the IETF Draft draft-ietf-bess-dci-evpn-overlay. An I-ES is a virtual Ethernet Segment (vES) that allows Data Center Gateways (DCGWs) with two BGP instances (one for EVPN-MPLS and one for EVPN-VXLAN) to handle redundancy in VXLAN access networks. I-ESs support the RFC 7432 multi-homing functions, including single-active and all-active, ESI-label based split-horizon filtering, Designated Forwarder (DF) election, aliasing, and backup functions on remote EVPN-MPLS PEs.

The chapter EVPN-MPLS Interconnect for EVPN-VXLAN VPLS Services describes how VPLS services with two BGP instances are configured and describes a redundant mechanism referred to as Multi-homed Anycast Configuration for Dual BGP-Instance VPLS Services. The use of I-ESs is recommended over this anycast configuration.

In addition to the EVPN multi-homing features, the main advantages of the I-ES solution compared to the redundant solution (described in Anycast Redundant Solution for Dual BGP-instance Services) are as follows:

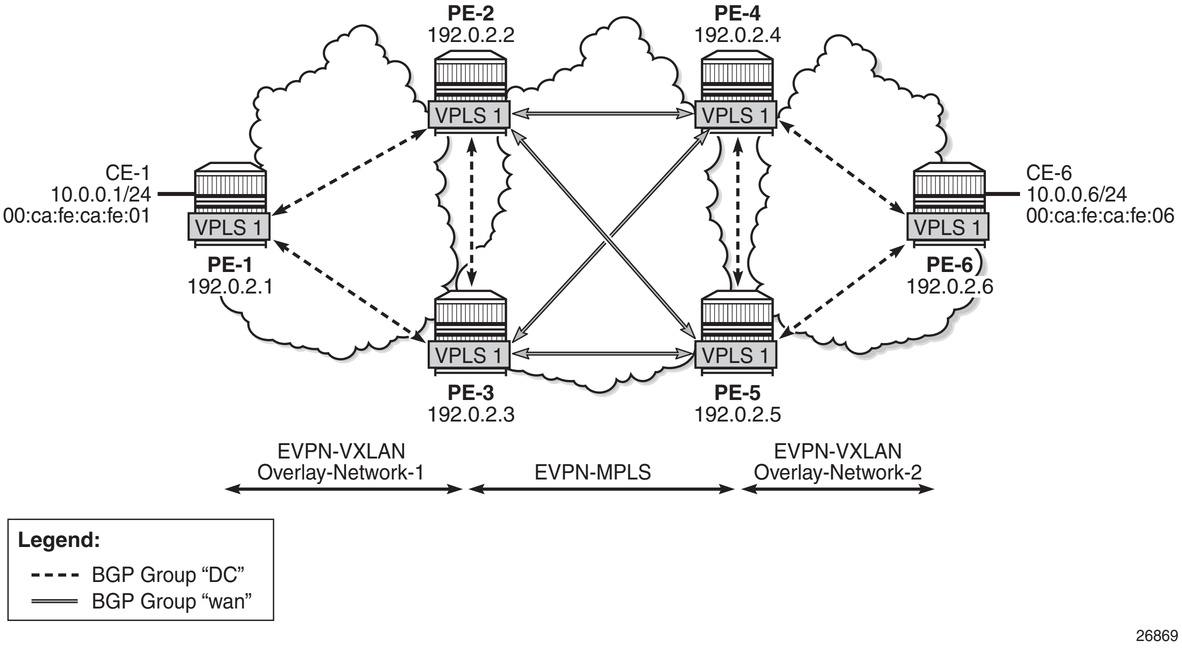
- The use of I-ES for redundancy in dual BGP-instance services allows local SAPs on the DCGWs. This is not supported in the anycast solution.
- P2MP mLDP can be provisioned to transport Broadcast, Unknown unicast, and Multicast (BUM) traffic between DCs that use I-ES, without any risk of packet duplication. As described in The Use of Provider-tunnels on Multi-Homed Anycast Solutions, packet duplication may occur in the anycast DCGW solution when mLDP is used in the WAN.

When EVPN-MPLS networks are interconnected to EVPN-VXLAN networks, the I-ES concept and procedures apply only to the access VXLAN network; the EVPN-MPLS network does not modify its existing behavior compared to any other ES.

Configuration

Figure 68 shows the topology and infrastructure configuration, which are the same as in chapter EVPN-MPLS Interconnect for EVPN-VXLAN VPLS Services. Read that chapter to see how the PEs are configured at port, IS-IS, and base BGP level.

Figure 68 **EVPN-MPLS Interconnect for EVPN-VXLAN - BGP Topology**



PE-1, PE-2, and PE-3 simulate a data center, shown as Overlay-Network-1, where PE-2 and PE-3 are DCGWs. In the same way, PE-4, PE-5, and PE-6 simulate a remote data center, Overlay-Network-2. Inside each DC, EVPN-VXLAN is used and the two DCGW pairs are connected by EVPN-MPLS. CE-1 and CE-6 are end-to-end connected by EVPN without any VLAN or Pseudowire (PW) hand-off, maintaining all the EVPN advantages across the DC Interconnect (DCI) network.

Interconnect Ethernet Segment (I-ES) Configuration

After the base infrastructure is configured (interfaces, IGP, LDP in the core, and BGP EVPN peering sessions, as per Figure 1), two I-ESs configured on the DCGWs show the use of the Interconnect Ethernet Segments.

The I-ES "I-ES231" is configured on PE-2 and PE-3 as follows:

```
A:PE-2>config>service>system>bgp-evpn# info
----------------------------------------------
                ethernet-segment "I-ES231" virtual create
                    esi 00:23:23:23:23:23:23:00:00:01
                    service-carving
                        mode manual
                        manual
                            preference non-revertive create
                                value 150
                            exit
                            evi 101 to 200
                        exit
                    exit
                    multi-homing all-active
                    network-interconnect-vxlan 1
                    service-id
                        service-range 1 to 100
                        service-range 101 to 200
                    exit
                    no shutdown
                exit
```

```
A:PE-3>config>service>system>bgp-evpn# info
----------------------------------------------
                ethernet-segment "I-ES231" virtual create
                    esi 00:23:23:23:23:23:23:00:00:01
                    service-carving
                        mode manual
                        manual
                            preference non-revertive create
                                value 50
                            exit
                            evi 101 to 200
                        exit
                    exit
                    multi-homing all-active
                    network-interconnect-vxlan 1
```

```
    service-id
        service-range 1 to 100
        service-range 101 to 200
    exit
    no shutdown
exit
```

On PE-1 and PE-2, the preceding configuration associates I-ES "I-ES231" with the VXLAN instance 1 in services contained in the range VPLS 1 to 100 and 101 to 200. The I-ES is modeled as a virtual ES, where:

- Two commands are needed within the ethernet-segment context: **network-interconnect-vxla**n and **service-id service-range <svc-id> [to <svc-id>]**.
    - The **[no] network-interconnect-vxlan** command identifies the VXLAN instance associated with the virtual ES. Only value 1 is supported in SR OS release 15.0.R4. This command is rejected in non-virtual ESs.
    - The **[no] service-range** command associates the specific service range with the ES. The ES must be configured as **network-interconnect-vxlan** before any service range can be added.
    - The other ES association options (port, lag, sdp, vc-id-range, dot1q, and qinq) are blocked in the ES when a **network-interconnect-vxlan** instance is configured.
    - The rest of the ES configuration options are supported. The **source-bmac-lsb** is blocked because the I-ES cannot be associated with I-VPLS or PBB-Epipe services.
    - All the services with two BGP instances associate the VXLAN destinations and ingress VXLAN instance with the ES.
- Multiple services (for example, 1 to 200 in the CLI above) can be associated with the same ES.
    - Up to eight service ranges per VXLAN instance can be configured. Ranges may overlap within the same ES (and not between different ESs). In this example, two non-overlapping ranges are configured to show the service range configuration, although a single range containing all the services could have been configured.
    - The service range may be configured before the service is, and it can be changed on the fly without having to shut down the ES first.
- When the **network-interconnect-vxlan** I-ES is configured, the ES operational state depends exclusively on the ES admin state.
    - Because the I-ES is not associated with a physical port or SDP, when testing the non-revertive service-carving manual mode, an ethernet-segment shutdown/no shutdown will result in the node sending its own administrative preference and "Do not preempt" (pref/DP) values, and taking over if pref/DP is higher than the current DF. This is because when the ES is no shutdown, the peer ES routes are not present at the EVPN

application layer, so the PE will send its own admin pref/DP values. Therefore, for I-ESs, the non-revertive mode will only work for node failures. See the *Preference-based and Non-revertive EVPN DF Election* chapter for more information about the preference-based and non-revertive DF election modes.

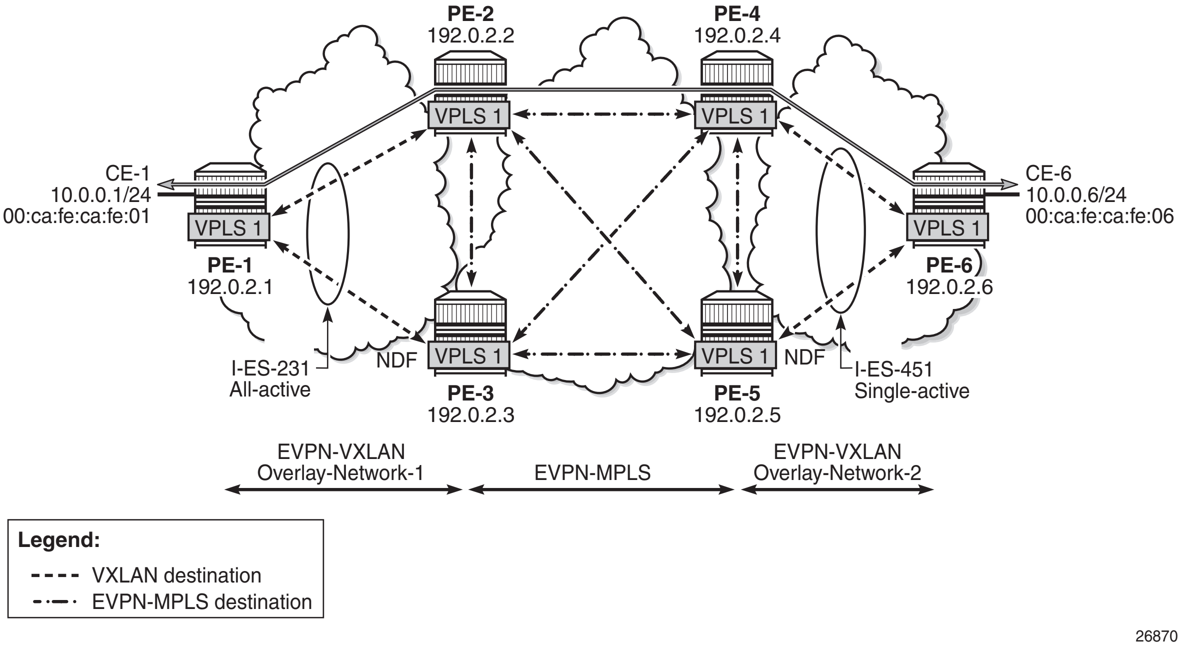
- There are no restrictions in the service-carving mode supported by I-ESs. In this example, preference-based service-carving is configured, but modes auto and (non-preference-based) manual are also supported.
- As described in the Preference-based and Non-revertive EVPN DF Election chapter, the service-carving context is configured with an EVI range that will pick up the lowest preference value when electing a DF for the service, whereas the non-configured EVI services will pick up the highest value when electing a DF. In this example, this means that, of the services allowed in the I-ES, that is, 1 to 200, services 1 to 100 will elect the highest Preference PE as DF, whereas services 101 to 200 will elect the lowest Preference PE.

PE-4 and PE-5 are configured with I-ES "I-ES451". The configuration of I-ES451 is similar to that of I-ES231; only single-active mode is configured, instead of all-active mode.

```
A:PE-4>config>service>system>bgp-evpn# info
----------------------------------------------
                ethernet-segment "I-ES451" virtual create
                    esi 00:45:45:45:45:45:45:00:00:01
                    service-carving
                        mode manual
                        manual
                            preference non-revertive create
                                value 150
                            exit
                            evi 101 to 200
                        exit
                    exit
                    multi-homing single-active
                    network-interconnect-vxlan 1
                    service-id
                        service-range 1 to 100
                        service-range 101 to 200
                    exit
                    no shutdown
                exit

A:PE-5>config>service>system>bgp-evpn# info
----------------------------------------------
                ethernet-segment "I-ES451" virtual create
                    esi 00:45:45:45:45:45:45:00:00:01
                    service-carving
                        mode manual
                        manual
                            preference non-revertive create
                                value 50
                            exit
                            evi 101 to 200
                        exit
                    exit
                    multi-homing single-active
                    network-interconnect-vxlan 1
                    service-id
                        service-range 1 to 100
                        service-range 101 to 200
                    exit
                    no shutdown
                exit
```

In this example, VPLS 1 will be configured and associated with the preceding I-ESs. Figure 69 shows an example of VPLS 1 and how it is associated with the I-ESs.

Figure 69 **VPLS service and association with I-ESs**



The configuration of VPLS 1 for PE-1, PE-2, and PE-3 is as follows. VPLS 101 is also configured in all the PEs in a similar way as VPLS 1, but not shown here. Also, the VPLS 1 configuration on the rest of the PEs is equivalent to the one in PE-1, PE-2, and PE-3, as follows:

```
A:PE-1>config>service>vpls# info
----------------------------------------------
            vxlan vni 1 instance 1 create
            exit
            bgp
            exit
            bgp-evpn
                evi 1
                vxlan
                    no shutdown
                exit
                mpls
                    shutdown
                exit
            exit
            stp
                shutdown
            exit
            sap 1/1/1:1 create
                no shutdown
            exit
            no shutdown
----------------------------------------------


A:PE-2>config>service>vpls# info
----------------------------------------------
            vxlan vni 1 instance 1 create
            exit
            bgp
                route-distinguisher 192.0.2.2:1
            exit
            bgp 2
                route-distinguisher 192.0.2.2:2
            exit
            bgp-evpn
                evi 1
                vxlan
                    no shutdown
                exit
                mpls
                    ingress-replication-bum-label
                    ecmp 2
                    bgp-instance 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            stp
                shutdown
            exit
            no shutdown
----------------------------------------------


A:PE-3>config>service>vpls# info
----------------------------------------------
            vxlan vni 1 instance 1 create
            exit
            bgp
                route-distinguisher 192.0.2.3:1
            exit
            bgp 2
                route-distinguisher 192.0.2.3:2
            exit
            bgp-evpn
                evi 1
                vxlan
                    no shutdown
                exit
                mpls
                    ingress-replication-bum-label
                    ecmp 2
                    bgp-instance 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            stp
                shutdown
            exit
            no shutdown
----------------------------------------------
```

As in the case of any other ESs, the association of instance and service is based on the ES configuration and there is no extra configuration required at the service level to make that association. The existing **show** commands that are used to check the status of the ES can be used to check the I-ESs. For example, on I-ES231:

```
A:PE-2# show service system bgp-evpn ethernet-segment name "I-ES231" all

===============================================================================
Service Ethernet Segment
===============================================================================
Name                    : I-ES231
Eth Seg Type            : Virtual
Admin State             : Enabled            Oper State          : Up
ESI                     : 00:23:23:23:23:23:23:00:00:01
Multi-homing            : allActive          Oper Multi-homing   : allActive
ES SHG Label            : 524281
Source BMAC LSB         : <none>
VXLAN Instance Id       : 1
ES Activation Timer     : 3 secs (default)
Svc Carving             : manual             Oper Svc Carving    : manual
Cfg Range Type          : lowest-pref

-------------------------------------------------------------------------------
DF Pref Election Information
-------------------------------------------------------------------------------
Preference     Preference     Last Admin Change        Oper Pref      Do No
Mode           Value                                   Value          Preempt
-------------------------------------------------------------------------------
non-revertive  150            09/04/2017 15:05:42      150            Enabled
-------------------------------------------------------------------------------

-------------------------------------------------------------------------------
EVI Ranges
-------------------------------------------------------------------------------
From                                    To
```

```
-------------------------------------------------------------------------------
101                                     200
-------------------------------------------------------------------------------
ISID Ranges: <none>
===============================================================================

===============================================================================
EVI Information
===============================================================================
EVI                 SvcId               Actv Timer Rem      DF
-------------------------------------------------------------------------------
1                   1                   0                   yes
101                 101                 0                   no
-------------------------------------------------------------------------------
Number of entries: 2
===============================================================================

-------------------------------------------------------------------------------
DF Candidate list
-------------------------------------------------------------------------------
EVI                                     DF Address
-------------------------------------------------------------------------------
1                                       192.0.2.2
1                                       192.0.2.3
101                                     192.0.2.2
101                                     192.0.2.3
-------------------------------------------------------------------------------
Number of entries: 4
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
---snip---

===============================================================================
Vxlan Instance Service Ranges
===============================================================================
Svc Range Start          Svc Range End            Last Changed
-------------------------------------------------------------------------------
1                        100                      09/04/2017 15:05:42
101                      200                      09/04/2017 16:57:19
-------------------------------------------------------------------------------
Number of Entries: 2
===============================================================================
```

The **show service id 1 vxlan instance 1 oper-flags** command shows the status of a VXLAN instance in the service. A service VXLAN instance will raise the oper-flag **MhStandby** (multi-homing standby) due to any of the following reasons:

- The PE is (single-active) non-Designated Forwarder (NDF) for that I-ES.
- The VXLAN service is added to the I-ES and either the ES is **shutdown** or **bgp-evpn>mpls** is **shutdown** in all the services included in the ES.

For example, because PE-5 is an NDF in I-ES451, the MhStandby flag will show "true":

```
A:PE-5# show service id 1 vxlan instance 1 oper-flags
```

```
===============================================================================
VPLS VXLAN oper flags
===============================================================================
MhStandby                           : true
===============================================================================
```

EVPN Route Handling in Dual BGP-instance VPLSs with I-ES

The configuration of I-ESs on DCGWs with two BGP instances has the following impact on the advertisement and process of the BGP-EVPN routes:

- EVPN MAC/IP routes:
  - MAC/IP routes received on the EVPN-MPLS BGP instance will be re-advertised to the EVPN-VXLAN BGP instance with the ESI set to zero in release 15.0.R4.
  - EVPN-VXLAN PE/NVEs (Network Virtual Edge devices) in the DC will receive the same MAC from two (or more) different MAC/IP routes from the DCGWs. The EVPN-VXLAN PE/NVEs will perform regular EVPN MAC/IP route selection.
  - MAC/IP routes received on the EVPN-VXLAN BGP instance will be re-advertised to the EVPN-MPLS BGP instance with the configured non-zero I-ESI value, assuming the VXLAN instance is not in the MhStandby operational state. MAC/IP routes received on the EVPN-VXLAN BGP instance will be dropped if the VXLAN instance is in the MhStandby state.
  - EVPN-MPLS PEs in the WAN will receive the same MAC from two (or more) DCGWs, set with the same ESI. EVPN-MPLS PEs will perform regular aliasing and backup functions.
- ES routes are exchanged for the I-ES. They should be sent only to the MPLS network and not to the VXLAN side. This can be achieved by using router policies. In any case, because ES routes use an ES-import route-target extended community, they should not be imported by VXLAN PEs.
- Auto-discover per ES (AD per-ES) and AD per-EVI routes are also advertised for the I-ES. They should be sent only to the MPLS network and not to the VXLAN network. As for ES routes, router policies can be used to prevent AD routes being sent to VXLAN peers.

Required BGP Policies to Avoid Control Plane Loops

Usually, the use of router policies is required when I-ESs are used for redundancy, to avoid control plane loops with MAC/IP routes. The control plane loops to be avoided are as follows:

1. Loops created by remote MACs (learned on remote PE SAPs):
   a. Remote EVPN-MPLS MAC/IP routes are re-advertised into EVPN-VXLAN with a Site of Origin (SOO) extended community (added by a BGP peer or vsi-export policy) identifying the DCGW pair. The other DCGW in the pair will drop EVPN-VXLAN MAC routes tagged with the self SOO. Router policies to add SOO and drop routes received with self SOO are needed.
   b. Also, when remote EVPN-VXLAN MAC/IP routes are re-advertised into EVPN-MPLS, the DCGWs will automatically drop EVPN-MPLS MAC/IP routes received with their own non-zero I-ESI. No router policies are needed for this.
2. Loops created by local SAP MACs:
   a. Local SAP MACs are learned and MAC/IP routes are advertised into both BGP instances. The MAC/IP routes advertised in the EVPN-VXLAN instance will be dropped by the peer based on the SOO router policies, as described in (1a) above, and DCGW local MACs will always be learned over the EVPN-MPLS destinations between the DCGWs.
   b. Because only EVPN-MPLS destinations exist between the DCGWs, EVPN-VXLAN MAC/IP and IMET routes exchanged between the DCGWs will be discarded and EVPN-VXLAN destinations will not be created between them.

As an example, the following BGP peer policies on PE-2 and PE-3 achieve the goals described above (similar policies would be configured on PE-4 and PE-5) and summarized as follows:

- Avoid sending service VXLAN routes to MPLS peers, and service MPLS routes to VXLAN peers.
- Avoid sending AD and ES routes to VXLAN peers.
- Add SOO to VXLAN routes to be sent to the ES peer.
- Drop VXLAN routes received from the ES peer.

```
A:PE-2/PE-3>config>router>policy-options# info
----------------------------------------------
            community "mpls" members "bgp-tunnel-encap:MPLS"
            community "vxlan" members "bgp-tunnel-encap:VXLAN"
            community "SOO-DCGW-23" members "origin:64500:23"
```

The following policy prevents the router from sending service VXLAN routes to MPLS peers:

```
policy-statement "allow only mpls"
    entry 10
        from
            community "vxlan"
            family evpn
        exit
        action drop
        exit
    exit
exit
```

The following policy makes sure the router exports only routes that include the VXLAN encapsulation:

```
policy-statement "allow only vxlan"
    entry 10
        from
            community "vxlan"
            family evpn
        exit
        action accept
        exit
    exit
    default-action drop
    exit
exit
```

The following import policy avoids importing routes with self SOO:

```
policy-statement "drop SOO-DCGW-23"
    entry 10
        from
            community "SOO-DCGW-23"
            family evpn
        exit
        action drop
        exit
    exit
exit
```

The following export policy adds SOO but only to VXLAN routes. This allows the peer to drop routes based on the SOO, without affecting the MPLS routes.

```
policy-statement "add SOO to vxlan routes"
    entry 10
        from
            community "vxlan"
            family evpn
        exit
        action accept
            community add "SOO-DCGW-23"
        exit
```

```
                exit
                default-action accept
                exit
            exit
```

The BGP configuration for PE-2 and PE-3 is as follows:

```
A:PE-2>config>router>bgp# info
----------------------------------------------
            family evpn
            vpn-apply-import
            vpn-apply-export
            rapid-withdrawal
            rapid-update evpn
            group "dc"
                type internal
                export "allow only vxlan"
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.3
                    import "drop SOO-DCGW-23"
                    export "add SOO to vxlan routes"
                exit
            exit
            group "wan"
                type internal
                export "allow only mpls"
                neighbor 192.0.2.4
                exit
                neighbor 192.0.2.5
                exit
            exit
            no shutdown
----------------------------------------------
```

```
A:PE-3>config>router>bgp# info
----------------------------------------------
            family evpn
            vpn-apply-import
            vpn-apply-export
            rapid-withdrawal
            rapid-update evpn
            group "dc"
                type internal
                export "allow only vxlan"
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                    import "drop SOO-DCGW-23"
                    export "add SOO to vxlan routes"
                exit
            exit
            group "wan"
                type internal
                export "allow only mpls"
                neighbor 192.0.2.4
                exit
                neighbor 192.0.2.5
                exit
            exit
            no shutdown
----------------------------------------------
```

Single-active Multi-homing Operation

When the I-ES is configured as **single-active** and **no shutdown** (assuming at least one service is associated), the DCGWs will send ES and AD routes as usual for any ES, and run DF election based on the ES routes, with the candidate list being pruned by the AD routes.

In Figure 69, PE-4 and PE-5 are configured with I-ES451, which is a single-active ES. The NDF for a service (PE-5 for VPLS 1 in the example) will perform the following tasks:

- The VXLAN instance on the NDF will enter the MhStandby state and will block ingress and egress traffic on the VXLAN destinations associated with the I-ES.

```
A:PE-5# show service id 1 vxlan instance 1 oper-flags

===============================================================================
VPLS VXLAN oper flags
===============================================================================
MhStandby                              : true
===============================================================================
```

- MAC/IP routes and FDB process:
  - Advertised MAC/IP routes that are associated with the VXLAN instance are withdrawn.
  - Advertised MAC/IP routes corresponding to local SAP MACs or EVPN-MPLS binding MACs are withdrawn if they were advertised to the EVPN-VXLAN instance.
  - Received MAC/IP routes associated with the VXLAN instance are not installed in FDB. The MAC routes will show as "used" in the **show router bgp routes evpn mac** commands; however, only the MAC received from MPLS (in particular from the ES peer) will be programmed. As an example, the following CLI output shows how MAC 00:ca:fe:ca:fe:06 is learned on PE-4 (DF) and associated with the VXLAN destination to PE-6, whereas the MAC is installed associated with an MPLS destination (remote ES) on PE-5 (NDF).

```
A:PE-4# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
```

```
                                                     Age
-------------------------------------------------------------------------------
1         00:ca:fe:ca:fe:01 eES:                     Evpn     09/05/17 13:00:03
                             00:23:23:23:23:23:23:00:00:01
1         00:ca:fe:ca:fe:06 vxlan:                   Evpn     09/05/17 13:00:03
                             192.0.2.6:1
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================

A:PE-5# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC                Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         00:ca:fe:ca:fe:01 eES:                     Evpn     09/05/17 13:00:03
                             00:23:23:23:23:23:23:00:00:01
1         00:ca:fe:ca:fe:06 eES:                     Evpn     09/05/17 13:00:03
                             00:45:45:45:45:45:45:00:00:01
-------------------------------------------------------------------------------
No. of MAC Entries: 2
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
```

- Inclusive Multicast Ethernet Tag (IMET) routes process:
  - IMET-Assisted Replication with replicator role (IMET-AR-R) routes are withdrawn if the VXLAN instance enters the MhStandby state. Only the DF will advertise the IMET-AR-R routes. For more information on AR, see chapter Layer 2 Multicast Optimization for EVPN-VXLAN - Assisted Replication.
  - IMET-Ingress Replication advertisements (IMET-IR) routes, in case of NDF (or the MhStandby state), are controlled by the **config>service>vpls>bgp-evpn>vxlan# [no] send-imet-ir-on-ndf** command.
    - By default, the command is enabled and the router will advertise IMET-IR routes even if the PE is NDF (MhStandby). This will attract BUM traffic (even if the NDF ends up dropping it); however, attracting BUM traffic will also speed up convergence in case of DF switchover. The command works for single-active and all-active.
    - If disabled, the router will withdraw the IMET-IR routes when the PE is NDF and will not attract BUM traffic.

In spite of not sending BUM or unicast traffic, the NDF for a service still creates the VXLAN bindings; however, they are not associated with any MACs and they are flagged as non-multicast capable, or "-" in the Mcast column of the following command:

```
A:PE-5# show service id 1 vxlan
===============================================================================
Vxlan Src Vtep IP: N/A
===============================================================================
VPLS VXLAN, Ingress VXLAN Network Id: 1
Creation Origin: manual
Assisted-Replication: none
RestProtSrcMacAct: none
VXLAN Instance Id: 1

===============================================================================
VPLS VXLAN service Network Specifics
===============================================================================
Ing Net QoS Policy : none                              Vxlan VNI Id     : 1
Ingress FP QGrp    : (none)                            Ing FP QGrp Inst : (none)
===============================================================================
Egress VTEP, VNI
===============================================================================
VTEP Address                            Egress VNI  Num. MACs   Mcast Oper  L2
                                                                      State PBR
-------------------------------------------------------------------------------
192.0.2.6                               1           0           -     Up    No
-------------------------------------------------------------------------------
Number of Egress VTEP, VNI : 1
-------------------------------------------------------------------------------
===============================================================================
```

The I-ES DF PE for the service (PE-4) will continue advertising IMET and MAC/IP routes for the associated VXLAN instance. Forwarding will also happen as usual on the DF VXLAN bindings. When the DF PE receives BUM traffic from VXLAN, it will send it, adding the egress ESI label if needed.

All-active Multi-homing Operation

The same considerations as in single-active for ES and AD routes and DF election apply to all-active multi-homing. In Figure 69, PE-2 and PE-3 are configured with I-ES231, which is an all-active ES. The NDF PE for a service (PE-3 for VPLS 1, in the example) will show the following behavior:

- The VXLAN instance on the NDF will not enter the MhStandby state because it will still forward unicast traffic:

```
*A:PE-3# show service id 1 vxlan instance 1 oper-flags

===============================================================================
VPLS VXLAN oper flags
```

```
===============================================================================
MhStandby                              : false
===============================================================================
```

- MAC/IP routes and FDB process: MAC/IP routes are received, installed, and advertised as in the DF router.
- IMET routes process:
  - As in the single-active case, IMET-AR-R routes are withdrawn on the NDF. Only the DF will advertise the IMET-AR-R routes.
  - Also, as in the single-active case, IMET-IR advertisement from the NDF will be controlled by the **config>service>vpls>bgp-evpn>vxlan# [no] send-imet-ir-on-ndf** command. Advertising the IMET-IR route from the NDF will attract BUM traffic from the VXLAN PEs to the NDF, even though the unknown unicast traffic will be forwarded only when it is safe to do so. See section All-active Multi-homing and Unknown Unicast Forwarding on the NDF for more information about unknown unicast forwarding.

Contrary to the behavior in single-active multi-homing, in all-active, the NDF will forward unknown unicast to the VXLAN PEs as usual, but block broadcast and multicast in the upstream and downstream direction. In our example, the NDF for VPLS 1 (PE-3) will show the VXLAN destinations created as "U" (Unknown unicast) in the Mcast column of the **show service id 1 vxlan** command, as follows:

```
*A:PE-3# show service id 1 vxlan
===============================================================================
Vxlan Src Vtep IP: N/A
===============================================================================
VPLS VXLAN, Ingress VXLAN Network Id: 1
Creation Origin: manual
Assisted-Replication: none
RestProtSrcMacAct: none
VXLAN Instance Id: 1

===============================================================================
VPLS VXLAN service Network Specifics
===============================================================================
Ing Net QoS Policy : none                              Vxlan VNI Id     : 1
Ingress FP QGrp    : (none)                            Ing FP QGrp Inst : (none)
===============================================================================
Egress VTEP, VNI
===============================================================================
VTEP Address                            Egress VNI  Num. MACs   Mcast Oper  L2
                                                                      State PBR
-------------------------------------------------------------------------------
192.0.2.1                               1           1           U     Up    No
-------------------------------------------------------------------------------
Number of Egress VTEP, VNI : 1
-------------------------------------------------------------------------------
===============================================================================
```

All-active Multi-homing and Unknown Unicast Forwarding on the NDF

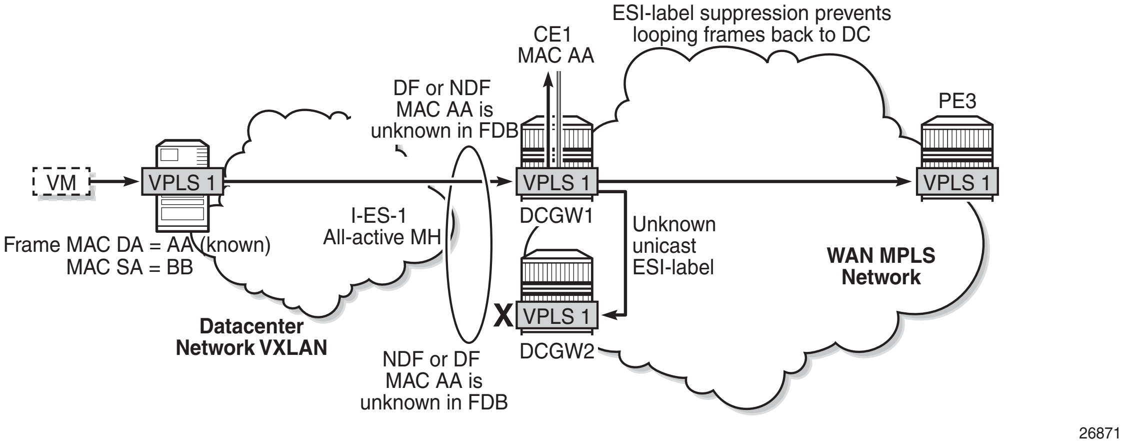
The unknown unicast traffic will be transmitted on the (all-active multi-homing) NDF in the upstream and downstream directions only in those cases where there is no risk of packet duplication. The router considers there is no risk when transmitting an unknown unicast packet on the NDF if:

- Unknown unicast packet arrives without an ESI label.
- Unknown unicast packet arrives without a BUM label (label advertised by an IMET route as opposed to a MAC/IP route).
- Unknown unicast packet passes a MAC Source Address (MAC SA) suppression (MAC SA lookup does not yield an entry associated with the I-ES).

The following examples show how unknown unicast traffic is handled in all-active I-ESs.

Figure 70 shows an example with two DCGWs where (all-active) I-ES-1 is defined.

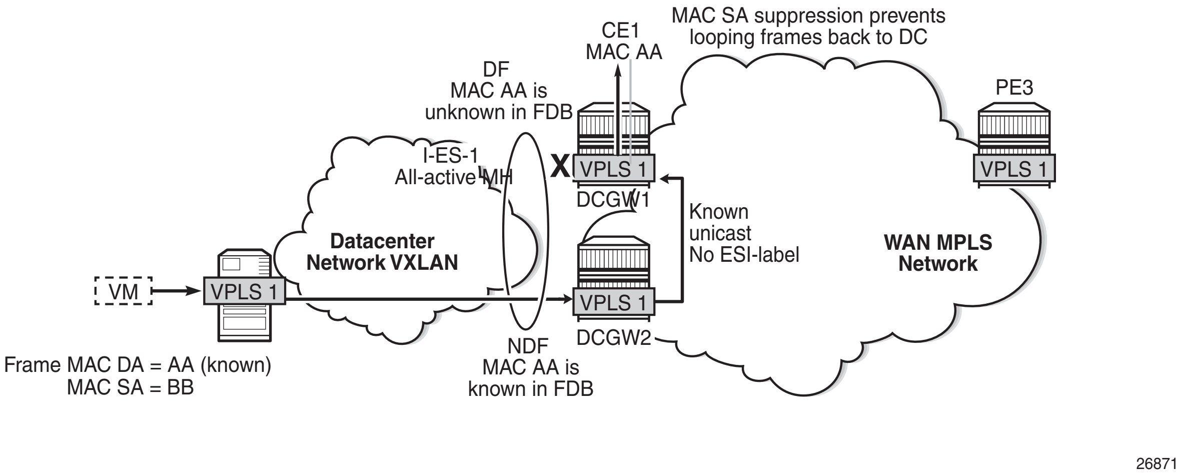
Figure 70 **All-active Multi-homing and Unknown Unicast Example 1**



The VXLAN PE/NVE transmits known unicast traffic, whereas DCGW1 has not learned the MAC yet. Regardless of the DCGW1 being DF or NDF, it will accept unknown unicast and will flood to local SAPs and EVPN destinations. When sending to DCGW2, the router will send the ESI label identifying the I-ES. DCGW2 will not send unknown traffic back to the DC due to the ESI-label suppression on the I-ES.

Figure 71 shows a similar example where the VXLAN node sends known unicast with MAC Destination Address (MAC DA) "AA" to DCGW2.

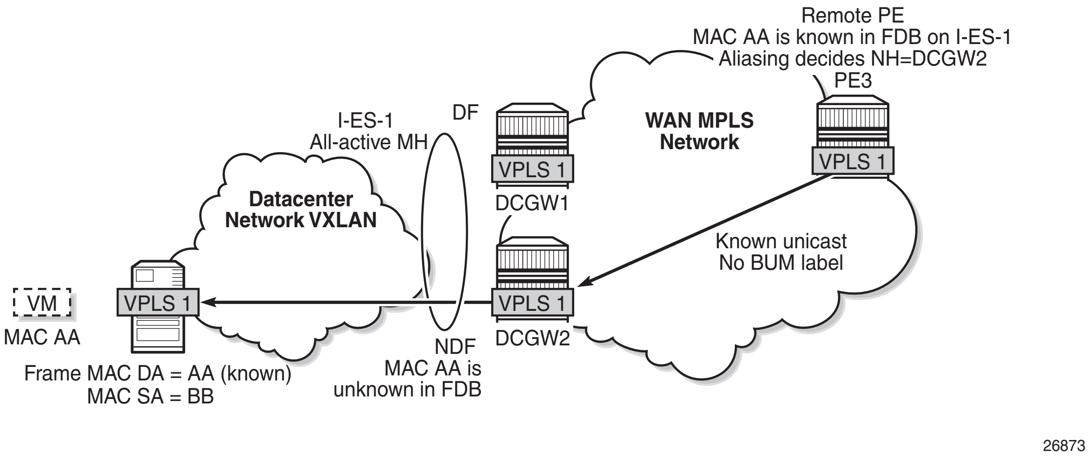
Figure 71 All-active Multi-homing and Unknown Unicast Example 2



DCGW2 does a MAC lookup and sends the frame as known unicast to DCGW1 via the EVPN-MPLS destination. However, MAC AA is unknown in DCGW1 for some reason (such as FDB limit exceeded, SAP failure, and so on). In this case, DCGW1 will flood the frame to CE1 and not to the VXLAN network. Even though the frame is not coming with an ESI label, the DCGW1 router does a MAC SA suppression and will not send unknown unicast frames to the I-ES. MAC SA suppression means that the router will do a MAC SA lookup on the FDB and will suppress the flooding to the I-ES if the MAC SA is learned on the I-ES (as in Figure 71).

Figure 72 shows an example in which the NDF forwards "no-risk" unknown unicast traffic to avoid black-holes.

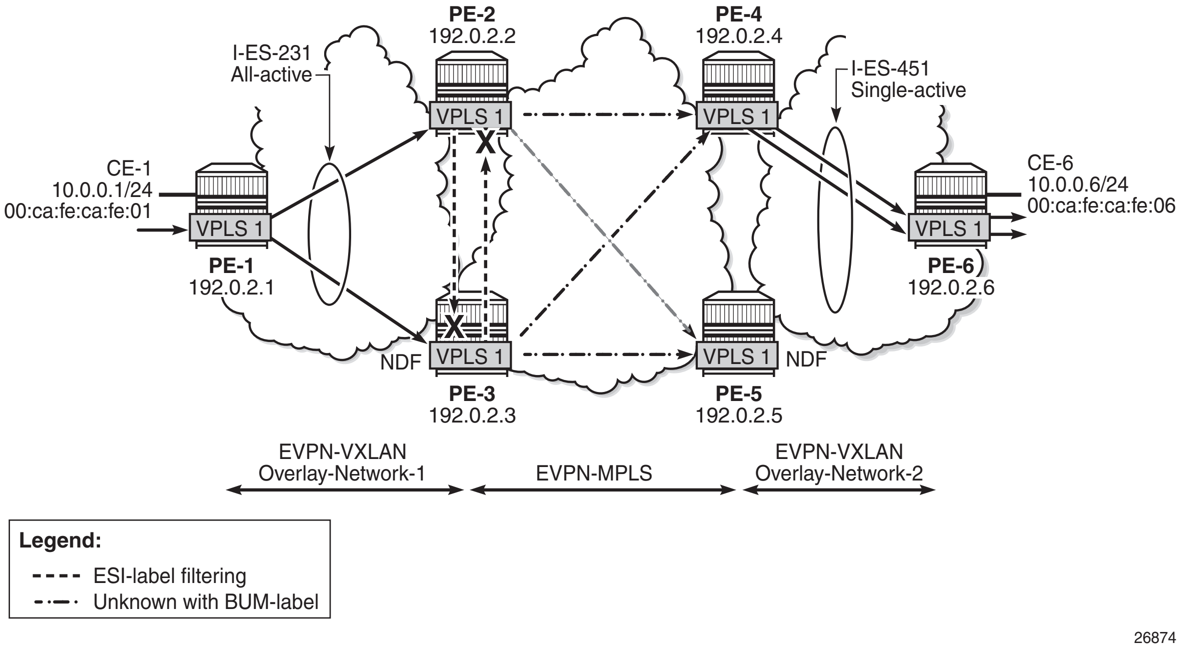
Figure 72 All-active Multi-homing and Unknown Unicast Example 3



PE3 receives unicast traffic with MAC DA = AA. The MAC address is known in the FDB and associated with I-ES-1; therefore, because PE3 is configured to do aliasing to DCGW1 and DCGW2 (bgp-evpn>mpls# ecmp 2), a packet hash determines that it has to be sent to DCGW2 (NDF). The packet arrives at DCGW2 with a unicast label. DCGW2 does a lookup and MAC AA is unknown for some reason (such as FDB limit exceeded, MAC not learned yet, and so on). In this case, DCGW2 will forward the packet to the I-ES VXLAN bindings, even if it is NDF. This behavior avoids black-hole periods in the network for unicast traffic.

Finally, in some cases, the unknown unicast forwarding behavior on the NDF may cause some transient packet duplication that can be avoided by configuring the **no send-imet-ir-on-ndf** command. The following example shows the use of this command to avoid transient packet duplication. Figure 73 shows how transient packet duplication may occur with the default setting **send-imet-ir-on-ndf**.

Figure 73 **All-active Multi-homing and send-imet-ir-on-ndf**



Transient packet duplication may occur when sending unknown unicast from CE-1 to CE-6, if **send-imet-ir-on-ndf** is configured in PE-3 and PE-2. To show this, we clear the FDBs in all the PEs in the example as well as the ARP caches on the CEs.

The following command is executed in all the PEs and CEs:

```
A:PE-1# clear service id 1 fdb all
A:PE-1#
A:PE-1# show service id 1 fdb detail
```

```
===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================
```

The following command clears the ARP table of the VPRN instance (defined in PE-1 using a loop) simulating CE-1:

```
A:PE-1# clear router 300 arp all
A:PE-1#
A:PE-1# show router 300 arp

===============================================================================
ARP Table (Service: 300)
===============================================================================
IP Address      MAC Address       Expiry    Type   Interface
-------------------------------------------------------------------------------
10.0.0.1        00:ca:fe:ca:fe:01 00h00m00s Oth[I] local
-------------------------------------------------------------------------------
No. of ARP Entries: 1
===============================================================================
```

When ICMP traffic is sent from CE-1 to CE-6, a duplicate entry occurs on CE-1:

```
A:PE-1# ping router 300 10.0.0.6
PING 10.0.0.6 56 data bytes
64 bytes from 10.0.0.6: icmp_seq=1 ttl=64 time=28.8ms.
64 bytes from 10.0.0.6: icmp_seq=1 ttl=64, duplicate.
64 bytes from 10.0.0.6: icmp_seq=2 ttl=64 time=2.94ms.
64 bytes from 10.0.0.6: icmp_seq=3 ttl=64 time=2.97ms.
64 bytes from 10.0.0.6: icmp_seq=4 ttl=64 time=3.01ms.
64 bytes from 10.0.0.6: icmp_seq=5 ttl=64 time=2.85ms.
---- 10.0.0.6 PING Statistics ----
5 packets transmitted, 5 packets received, 1 duplicate
round-trip min = 2.85ms, avg = 8.12ms, max = 28.8ms, stddev = 10.3ms
```

This duplicate entry occurs because the packet gets to CE-6 twice and CE-6 sends two unicast ICMP reply messages back. From the CE-1 packet walkthrough:

- PE-1 floods the packet to PE-2 and PE-3 because the CE-6 MAC DA is unknown and it has VXLAN multicast destinations to them.
- PE-2 floods the unknown unicast packet to all the remote PEs because it is DF for I-ES231. PE-2 will add an ESI label when sending to PE-3, and a BUM label when sending to all of them.

- PE-3 is NDF for I-ES231, but it floods the packet because the I-ES is all-active and the unknown unicast packet is considered low risk. The packet arrives with no ESI label, no BUM label (in VXLAN, VNIs are the same for unicast and BUM), and the MAC SA suppression passes because the packet is coming from the I-ES and not from MPLS. PE-3 uses a BUM label when flooding the packet and an ESI label when sending to PE-2.
- PE-4 receives two unknown unicast packets and forwards both to PE-6.
- PE-5 does not forward because it is NDF. This is true regardless of the I-ES being single-active or all-active (if all-active, the packet will not be forwarded because it arrives with a BUM label).

This packet duplication situation is transient and it will stop as soon as the two MAC addresses are learned on the PEs. However, if needed, this situation can be avoided by configuring **no send-imet-ir-on-ndf** (the BGP-EVPN VXLAN must be shutdown first):

```
*A:PE-2# configure service vpls 1 bgp-evpn vxlan no send-imet-ir-on-ndf
MINOR: SVCMGR #7886 cannot modify evpn - Evpn vxlan not shut
*A:PE-2# configure service vpls 1 bgp-evpn vxlan shutdown
*A:PE-2# configure service vpls 1 bgp-evpn vxlan no send-imet-ir-on-ndf
*A:PE-2# configure service vpls 1 bgp-evpn vxlan no shutdown
```

```
*A:PE-3# configure service vpls 1 bgp-evpn vxlan shutdown
*A:PE-3# configure service vpls 1 bgp-evpn vxlan no send-imet-ir-on-ndf
*A:PE-3# configure service vpls 1 bgp-evpn vxlan no shutdown
```

This command will make the NDF (PE-3) withdraw the IMET-IR route; therefore, PE-1 will only flood unknown unicast packets to the DF (PE-2). The following IMET-IR routes are received on PE-1: one route sent by DF PE-2 for VPLS 1 and two routes for VPLS 101.

```
*A:PE-1# show router bgp routes evpn inclusive-mcast
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN Inclusive-Mcast Routes
===============================================================================
Flag  Route Dist.         OrigAddr
      Tag                 NextHop
-------------------------------------------------------------------------------
u*>i  192.0.2.2:1         192.0.2.2
      0                   192.0.2.2

u*>i  192.0.2.2:101       192.0.2.2
      0                   192.0.2.2
```

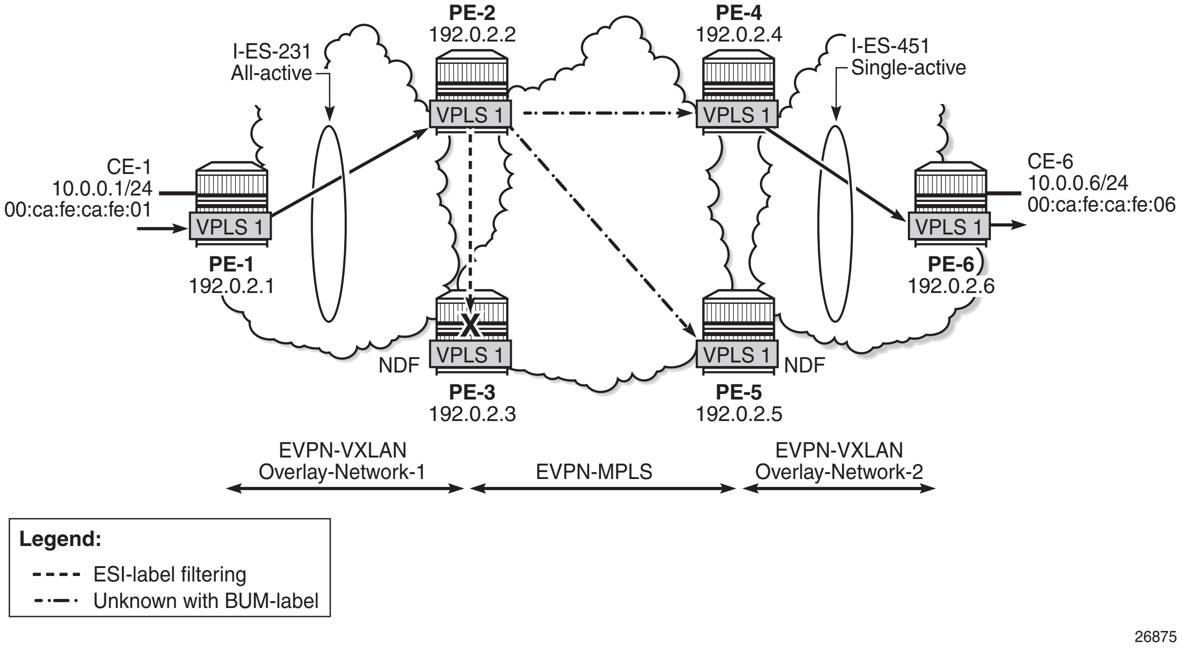
```
u*>i  192.0.2.3:101       192.0.2.3
      0                   192.0.2.3

-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:PE-1#
```

If a DF switchover occurs in the I-ES, the new DF would advertise the IMET-IR route and the new NDF would withdraw it.

After clearing FDBs and ARP caches again, the test is repeated with no packet duplication. Figure 74 shows how PE-1 does not send unknown unicast to PE-3 (NDF) anymore and, therefore, there is no duplication.

```
A:PE-1# ping router 300 10.0.0.6
PING 10.0.0.6 56 data bytes
64 bytes from 10.0.0.6: icmp_seq=1 ttl=64 time=46.6ms.
64 bytes from 10.0.0.6: icmp_seq=2 ttl=64 time=2.49ms.
64 bytes from 10.0.0.6: icmp_seq=3 ttl=64 time=2.58ms.
64 bytes from 10.0.0.6: icmp_seq=4 ttl=64 time=2.69ms.
64 bytes from 10.0.0.6: icmp_seq=5 ttl=64 time=3.03ms.

---- 10.0.0.6 PING Statistics ----
5 packets transmitted, 5 packets received, 0.00% packet loss
round-trip min = 2.49ms, avg = 11.5ms, max = 46.6ms, stddev = 17.6ms
```

Figure 74   All-active Multi-homing and no send-imet-ir-on-ndf



Local SAPs and Provider Tunnels along with I-ES

As described in the Overview section, the main advantages of the I-ES solution over the anycast redundant solution for dual BGP-instance services are the support of local SAPs and P2MP mLDP trees without packet duplication. This section shows the configuration of local SAPs and provider tunnels along with I-ES in VPLS services. The local SAPs can, at the same time, belong to an ES or a vES.

As an example, PE-2 VPLS 1 would be reconfigured as follows (similar configuration would exist in PE-3, with provider tunnel also configured in PE-4 and PE-5):

```
*A:PE-2# configure service vpls 1
*A:PE-2>config>service>vpls# info
----------------------------------------------
            vxlan vni 1 instance 1 create
            exit
            bgp
                route-distinguisher 192.0.2.2:1
            exit
            bgp 2
                route-distinguisher 192.0.2.2:2
            exit
            bgp-evpn
                evi 1
                vxlan
                    no send-imet-ir-on-ndf
                    no shutdown
                exit
                mpls
                    ingress-replication-bum-label
                    ecmp 2
                    bgp-instance 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            provider-tunnel
                inclusive
                    owner bgp-evpn-mpls
                    root-and-leaf
                    mldp
                    no shutdown
                exit
            exit
            stp
                shutdown
            exit
            sap lag-1:1 create
                no shutdown
            exit
            no shutdown
----------------------------------------------
```

To have EVPN multi-homing from a CE locally connected to PE-2 and PE-3, an additional ES is configured on PE-2 and PE-3 that will include the local SAPs in VPLS 1, as follows:

```
*A:PE-2>config>service>system>bgp-evpn# info
----------------------------------------------
                ethernet-segment "I-ES231" virtual create
                    esi 00:23:23:23:23:23:23:00:00:01
                    service-carving
                        mode manual
                        manual
                            preference non-revertive create
                                value 150
                            exit
                            evi 101 to 200
                        exit
                    exit
                    multi-homing all-active
                    network-interconnect-vxlan 1
                    service-id
                        service-range 1 to 100
                        service-range 101 to 200
                    exit
                    no shutdown
                exit
                ethernet-segment "vES232" virtual create
                    esi 00:23:23:23:23:23:23:00:00:02
                    service-carving
                        mode auto
                    exit
                    multi-homing all-active
                    lag 1
                    dot1q
                        q-tag-range 1
                    exit
                    no shutdown
                exit
----------------------------------------------
```

Troubleshooting and Debugging

Common troubleshooting commands to operate dual BGP-instance VPLS services are in the corresponding section of EVPN-MPLS Interconnect for EVPN-VXLAN VPLS Services. Also, ES and virtual ES can be troubleshot by using the commands described in chapter EVPN for MPLS Tunnels.

As well, the following **show** commands are specific to the use of I-ES in the router:

```
*A:PE-2# show service id 1 vxlan instance 1 oper-flags

===============================================================================
VPLS VXLAN oper flags
===============================================================================
MhStandby                              : false
===============================================================================

*A:PE-2# show service vxlan-instance-using ethernet-segment

===============================================================================
VXLAN Ethernet-Segment Information
===============================================================================
SvcId       VXLAN Instance      ES Name                          Status
-------------------------------------------------------------------------------
1           1                   I-ES231                          DF
101         1                   I-ES231                          NDF
===============================================================================

*A:PE-2# show service vxlan-instance-using ethernet-segment "I-ES231"

===============================================================================
VXLAN Ethernet-Segment Information
===============================================================================
SvcId                            VXLAN Instance                  Status
-------------------------------------------------------------------------------
1                                1                               DF
101                              1                               NDF
===============================================================================
```

Conclusion

Based on draft-ietf-bess-dci-evpn-overlay, SR OS supports the connectivity of Layer 2 EVPN-VXLAN services to an EVPN-MPLS network. This chapter complements the chapter EVPN-MPLS Interconnect for EVPN-VXLAN VPLS Services by describing how redundancy can be improved with the use of I-ES multi-homing, a concept standardized in draft-ietf-bess-dci-evpn-overlay.

EVPN-MPLS Interconnect for EVPN-VXLAN VPLS Services

This chapter provides information about EVPN-MPLS Interconnect for EVPN-VXLAN VPLS Services.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SR OS Release 14.0.R5, but the CLI in the current edition is based on SR OS release 15.0.R2.

Chapters EVPN for MPLS Tunnels and EVPN for VXLAN Tunnels (Layer 2) are prerequisite reading.

Overview

When EVPN-MPLS is deployed in the WAN, many service providers are looking for a way to integrate existing Layer 2 EVPN-VXLAN based data center services into the WAN, while keeping the end-to-end advantages of EVPN. The IETF draft document in http://tools.ietf.org/html/draft-ietf-bess-dci-evpn-overlay describes how to provide Layer 2 connectivity for EVPN-overlay data centers in different ways. This chapter follows section 3.4 of that document, in which EVPN-MPLS is used in the same VPLS service that terminates overlay (VXLAN) tunnels.

To provide EVPN-MPLS connectivity to VPLS services terminating EVPN-VXLAN, SR OS supports the configuration of BGP-EVPN MPLS and BGP-EVPN VXLAN at the same time by adding two BGP instances to the service. Two BGP instances are supported in the same VPLS at most. As a rule, BGP-EVPN MPLS can use BGP instance 1 or 2, and BGP-EVPN VXLAN can only use BGP instance 1.

In a service with EVPN-VXLAN and EVPN-MPLS, the **config>service>vpls>bgp-evpn>mpls>bgp-instance 2** command allows the user to associate BGP-EVPN MPLS to a different instance than BGP-EVPN VXLAN, and therefore, have both encapsulations simultaneously enabled in the same service. When the two BGP instances are successfully added to the same VPLS service, the service behaves as follows:

- MAC/IP routes received on one instance will be "consumed" (accepted, imported, and installed in FDB) and re-advertised in the other instance, as long as the route is the best route for a specific MAC or MAC/IP.
- Inclusive multicast routes are independently generated for each BGP instance.
- From a data plane perspective, EVPN-MPLS and EVPN-VXLAN destinations are instantiated in different implicit Split-Horizon-Groups (SHGs) so that traffic can be forwarded between the two SHGs, but not between destinations of the same kind. For example, traffic coming from EVPN-MPLS cannot be forwarded to other destinations in the EVPN-MPLS SHG.

The following example shows a VPLS service configured on PE-2 with two BGP instances and both encapsulations, VXLAN and MPLS, configured at the same time:

```
configure
    service
        vpls 1 customer 1 create
            description "evpn-mpls and evpn-vxlan in the same service"
            vxlan vni 1 create
            exit
            bgp
                route-distinguisher 10:1
                route-target export target:64500:1 import target:64500:1
            exit
            bgp 2
                route-distinguisher 10:2
                route-target export target:64500:1 import target:64500:1
            exit
            bgp-evpn
                evi 1
                vxlan
                    no shutdown
                exit
                mpls
                    bgp-instance 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            no shutdown
```

In the preceding example

- **bgp 1** or simply **bgp** is the default BGP instance.

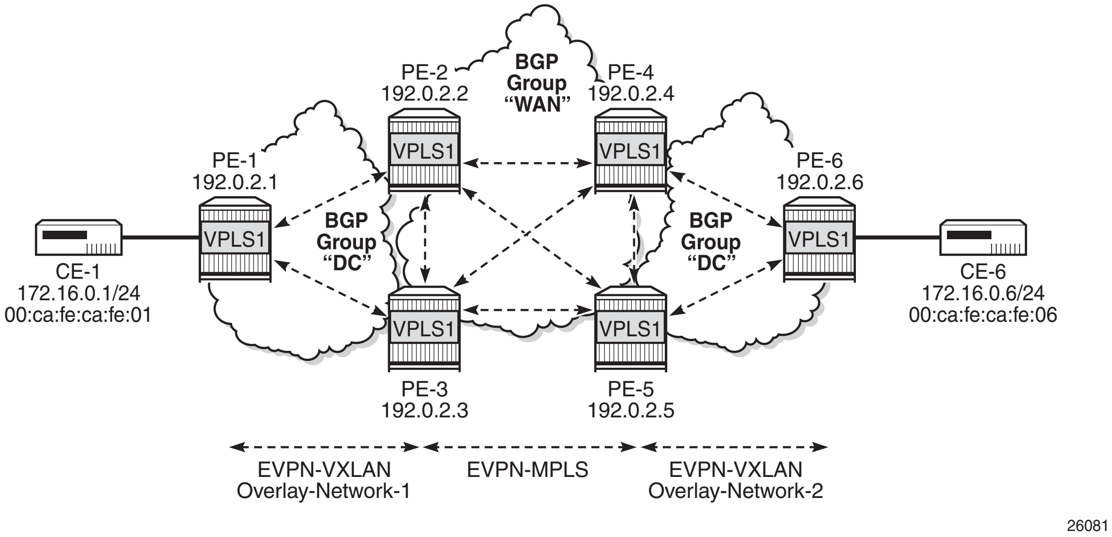
- **bgp 2** is the additional instance that is required when both BGP-EVPN VXLAN and BGP-EVPN MPLS are enabled in the service.
- The same commands supported under instance 1 exist for this second instance, with the following considerations:
  - **pw-template-binding** - the pseudowire (PW) template binding can only exist in instance 1; it is not supported in instance 2. Because no SAPs or SDP-bindings can exist in a VPLS service with two BGP instances, the **pw-template-binding** command is ineffective in this configuration.
  - **route-distinguisher** - the route distinguisher in both BGP instances must be different.
  - **route-target** - the route target in both instances can be the same or different.
  - **vsi-import** and **vsi-export** - import and export policies can also be defined for either BGP instance.
- The **mpls bgp-instance 2** command will assign the second BGP instance to MPLS. VXLAN always uses instance 1 (the default instance). The **bgp-evpn vxlan no shutdown** command will only be allowed if **bgp-evpn** is **shutdown** or if the BGP instance associated with MPLS has a different route distinguisher than the VXLAN instance (and vice versa).
- The **evi** can still be used for auto-derivation of RD/RT on **bgp-instance 1** and auto-derivation of RT (not RD) on **bgp-instance 2**. Auto-RD or an explicitly configured RD is needed in **bgp-instance 2**.

Configuration

Figure 75 shows the example topology that will be used throughout this chapter, as well as the BGP peering topology. PE-1, PE-2, and PE-3 simulate a data center, shown as Overlay-Network-1, where PE-2 and PE-3 are DC GWs. In the same way, PE-4, PE-5, and PE-6 simulate a remote data center, Overlay-Network-2. Inside each DC, EVPN-VXLAN is used.

The two DC GW pairs are connected by EVPN-MPLS; therefore, CE-1 and CE-6 are end-to-end connected by EVPN without any VLAN or PW hand-off, maintaining all the EVPN advantages across the DC Interconnect (DCI) network.

Figure 75 **EVPN-MPLS Interconnect for EVPN-VXLAN - Example Topology**



The example topology consists of six 7750 SR routers with the following initial configuration:

- Hybrid ports (they could have been network type too) are interconnecting the six PEs with configured router interfaces.
- The six PEs are running IS-IS and creating point-to-point adjacencies.
- Link LDP is configured in the core, among PE-2, PE-3, PE-4, and PE-5, while PE-1 and PE-6 are only running VXLAN.
- EVPN uses MP-BGP for exchanging reachability at service level. Therefore, BGP peering sessions must be established among the PEs for the EVPN family. Figure 75 shows the peering sessions established among the six PEs. Although usually a Route-Reflector (RR) is used in each DC and another RR in the WAN, in this example, there are direct peering sessions in each DC and in the WAN.

The following output shows the BGP configuration of PE-2. The BGP configuration on the rest of the DC GWs (PE-3, PE-4, and PE-5) would be similar:

```
configure
    router
        bgp
            family evpn
            vpn-apply-import
            vpn-apply-export
            rapid-withdrawal
            rapid-update evpn
            group "DC"
                type internal
                import "drop SOO-DCGW-23"
                export "allow only vxlan and add SOO"
```

```
                    neighbor 192.0.2.1
                    exit
                    neighbor 192.0.2.3
                    exit
                exit
                group "WAN"
                    type internal
                    import "drop SOO-DCGW-23"
                    export "allow only mpls and add SOO"
                    neighbor 192.0.2.4
                    exit
                    neighbor 192.0.2.5
                    exit
                exit
                no shutdown
```

Two different BGP groups are configured: DC and WAN. The DC group contains the DC neighbors (including the peer DC GW) and the WAN group contains the WAN neighbors. This grouping makes the use of policies easier. These policies will be explained in the section The Mandatory Use of BGP Policies in the Multi-Homed Anycast Solution.

The following output shows the BGP configuration of PE-1. PE-6 has a similar BGP configuration.

```
configure
    router
        bgp
            family evpn
            rapid-withdrawal
            rapid-update evpn
            group "DC"
                type internal
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
            exit
            no shutdown
```

VPLS Service Configuration

After the base infrastructure (interfaces, IGP, LDP in the core, and BGP) is configured, the services can be added. The configuration example in this section will use VPLS 1 as the service to be interconnected across the two DCs.

PE-1 and PE-6 have a regular EVPN-VXLAN configuration; DCI connectivity provided by EVPN-MPLS is completely transparent to them. The configuration of VPLS 1 in PE-1 is as follows:

```
configure
    service
        vpls 1 customer 1 create
            vxlan vni 1 create
            exit
            bgp
            exit
            bgp-evpn
                evi 1
                vxlan
                    no shutdown
                exit
                mpls
                    shutdown
                exit
            exit
            sap 1/2/1:1 create
                no shutdown
            exit
            no shutdown
        exit
```

See the EVPN for VXLAN Tunnels (Layer 2) chapter for a complete description of the EVPN-VXLAN commands.

The configuration on PE-2, PE-3, PE-4, and PE-5 (see Figure 75) enables EVPN-VXLAN and EVPN-MPLS in the same VPLS service. As an example, the VPLS 1 configuration on PE-2 is as follows:

```
configure
    service
        vpls 1 customer 1 create
            vxlan vni 1 create
            exit
            bgp
                route-distinguisher 64500:1
            exit
            bgp 2
                route-distinguisher 64500:2
            exit
            bgp-evpn
                incl-mcast-orig-ip 23.23.23.23
                evi 1
                vxlan
                    no shutdown
                exit
                mpls
                    ingress-replication-bum-label
                    bgp-instance 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            no shutdown
        exit
```

As described in the Overview section, the preceding configuration enables the router to create EVPN-VXLAN and EVPN-MPLS destinations in the same VPLS service, but in different SHGs. In addition to the **bgp 2** and **bgp-instance 2** commands already described in the Overview section, a new command, **incl-mcast-orig-ip**, is added in the configuration. If configured, this command will change the originating IP address in the inclusive multicast routes (from the default system IP) for both BGP instances. The section Multi-homed Anycast Configuration for Dual BGP-Instance VPLS Services describes why this command is added.

The following section provides a detailed description of the expected behavior for EVPN routes that are imported and exported on dual BGP-instance VPLS services.

EVPN Route Handling in Dual BGP-Instance VPLS Services

This section describes how the BGP-EVPN routes are processed in dual BGP-instance services.

Usually, the router validates the received tunnel encapsulation (from the RFC 5512 Extended Community) with the configured encapsulation of the service/BGP-instance. Therefore, an EVPN-VXLAN route will not get imported into the BGP-EVPN MPLS instance and vice-versa. This is also how the different EVPN route types are handled in dual BGP-instance services:

- **Route Type 1 - Auto-Discovery Routes**

  AD per-EVI routes are never generated by services with two BGP instances (because no Ethernet Segment (ES) can be associated with the dual BGP-instance service). However, AD per-EVI routes can still be received from the EVPN-MPLS peers and are processed as usual. Therefore, a VPLS service with two BGP instances will still support aliasing/backup and AD per-ES checking procedures for a remote multi-homed ES, as described in the EVPN for MPLS Tunnels chapter. However, in the example in Figure 75, PE-6 does not have any local multi-homed ES configured; therefore, no AD per-EVI routes are present in this example.
- **Route Type 2 - MAC/IP Routes**

No SAP/SDP-bindings are allowed in services with two BGP instances; therefore, all the MACs that are installed in the DC GWs FDBs are coming from EVPN. MAC/IP routes received on one of the two BGP instances will be imported and the MACs added to the FDB according to the existing selection rules. If the MAC is active, (therefore installed in the FDB) it will be re-advertised in the other BGP instance with the new BGP attributes of the other BGP instance (new route target if different, new route distinguisher, and so on). The **mac-advertisement** command will govern the advertisement of any MACs in either BGP instance.

The MAC/IP route redistribution across BGP instances is performed according to the following rules:

– A MAC route is redistributed only if it is the best route according to the EVPN selection rules in the EVPN for MPLS Tunnels chapter.

– Assuming a specific MAC route is the best one and has to be redistributed, the MAC/IP information along with the sticky bit is propagated in the redistribution.

– A change in the MAC/IP route sequence number or sticky bit in one instance is updated in the other instance, as long as that route is the best MAC route for the route key.

– When a MAC moves within the EVPN-VXLAN (or the EVPN-MPLS) network, the MAC route is received on the same BGP instance where it was previously received, but now with a higher sequence number. In this case, the MAC route will be redistributed with the new sequence number. However, a router with two BGP instances in the same service will not detect any duplicate MAC on the EVPN-VXLAN and EVPN-MPLS networks.

As an example, the following output shows the debug of a MAC/IP route received on PE-2, on the BGP instance for EVPN-VXLAN on VPLS 1, and how the route is re-advertised to the BGP instance used for MPLS (with a different next-hop, route distinguisher, label, and BGP tunnel encapsulation):

```
A:PE-2#
7 2017/05/10 08:13:44.60 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 88
    Flag: 0x90 Type: 14 Len: 44 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.1
        Type: EVPN-MAC Len: 33 RD: 192.0.2.1:1 ESI: ESI-0, tag: 0, mac len: 48
                       mac: 00:ca:fe:ca:fe:01, IP len: 0, IP: NULL, label1: 1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:64500:1
        bgp-tunnel-encap:VXLAN
"

8 2017/05/10 08:13:44.60 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.4
"Peer 1: 192.0.2.4: UPDATE
Peer 1: 192.0.2.4 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 96
    Flag: 0x90 Type: 14 Len: 44 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.2
        Type: EVPN-MAC Len: 33 RD: 64500:2 ESI: ESI-0, tag: 0, mac len: 48
                       mac: 00:ca:fe:ca:fe:01, IP len: 0, IP: NULL, label1: 4194272
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        origin:64500:23
        target:64500:1
        bgp-tunnel-encap:MPLS
"

9 2017/05/10 08:13:44.60 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.5
"Peer 1: 192.0.2.5: UPDATE
Peer 1: 192.0.2.5 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 96
    Flag: 0x90 Type: 14 Len: 44 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.2
        Type: EVPN-MAC Len: 33 RD: 64500:2 ESI: ESI-0, tag: 0, mac len: 48
                       mac: 00:ca:fe:ca:fe:01, IP len: 0, IP: NULL, label1: 4194272
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        origin:64500:23
        target:64500:1
        bgp-tunnel-encap:MPLS
"
```

- **Route Type 3 - Inclusive Multicast Routes**

  EVPN Inclusive Multicast Ethernet tag (IMET) routes are generated independently for each BGP instance with the correct BGP tunnel encapsulation extended community and the tunnel type associated to the BGP instance; for example, Ingress Replication (IR), P2MP mLDP, or Assisted Replication (AR):

  - On the EVPN-VXLAN BGP instance, IR or AR IMET routes are supported.

- When **assisted-replication replicator** is enabled and the received VXLAN broadcast and multicast packets contain an IP DA = AR-IP, the DC GW will send the packets back to VXLAN (but not to the VXLAN termination end-point (VTEP) from where the packet is received) in addition to the EVPN-MPLS destinations.
- If **assisted-replication replicator** is used on the DC GWs, the AR-IP (**configure>service>system>vxlan>assisted-replication-ip**) must be a loopback different from the router's system IP and the configured **bgp-evpn>incl-mcast-orig-ip**. The two AR-IPs in the DC GW pair do not need to be the same IP address.

– On the EVPN-MPLS BGP instance, IR, P2MP mLDP, or composite IMET routes are supported.

– Following is the behavior when the **incl-mcast-orig-ip** command is used:

- The configured IP in the **incl-mcast-orig-ip** command is encoded in the originating IP field of the IMET routes for IR, P2MP, and composite routes for both BGP instances.
- The originating IP field of the IMET AR routes is still derived from the configured **service>system>vxlan>assisted-replication-ip** value.

– The received IMET routes will be processed in the following way depending on their type:

- IMET-IR routes: the EVPN destination (MPLS or VXLAN) is set up based on the NLRI next-hop.
- IMET-P2MP routes: the Provider Multicast Service Interface (PMSI) Tunnel Attribute (PTA) tunnel ID will be used to join the mLDP tree (as mLDP FEC in the LDP mapping messages).
- IMET-P2MP-IR (composite) routes: the PTA tunnel ID is used to join the mLDP tree. The NLRI next-hop is used to build the EVPN destination.
- IMET-AR routes: the NLRI next-hop is used to build the EVPN-VXLAN destination.

– Upon reception of two IMET routes with similar information, the router behaves as follows:

- If the router receives two IMET routes with the same originating IP, different RDs, and different NLRI next-hops, it will set up two EVPN destinations, one to each next-hop.
- If the router gets two IMET routes with the same originating IP, different RDs, but the same next-hop, it will set up only one EVPN destination.

  - The router will not set up an EVPN destination to its DC GW peer if the received originating IP matches its own originating IP, regardless of whether the local RD and the remote RD are the same or different. This enables the use of the redundant anycast solution that is described in the following section: Multi-homed Anycast Configuration for Dual BGP-Instance VPLS Services.

- **Route Type 4 - ES Routes**

  ESs are supported in routers where dual BGP-instance services exist. However, since dual BGP-instance VPLS services do not support SAP/SDP-bindings, ESs and ES routes are not relevant to these types of services.

- **Route type 5 - IP-prefix routes**

  R-VPLS services are not supported along with dual BGP instances; therefore, IP-prefix routes are neither generated nor processed by the service.

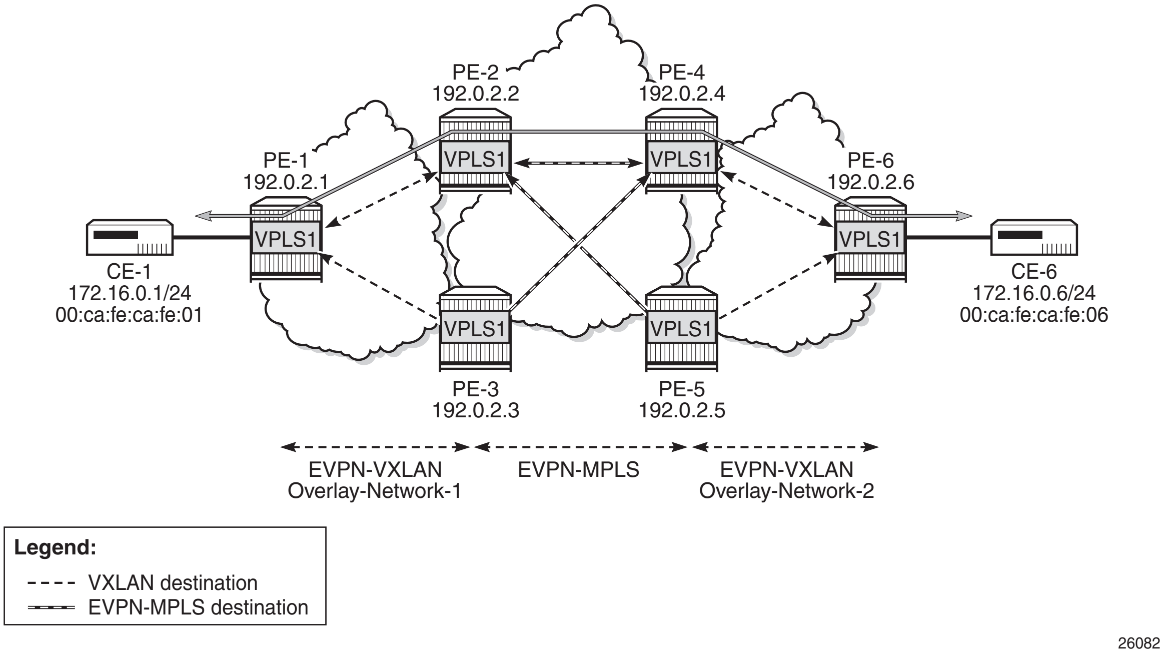
Multi-homed Anycast Configuration for Dual BGP-Instance VPLS Services

Services with EVPN-MPLS and EVPN-VXLAN SHGs are specified in *draft-ietf-bess-dci-evpn-overlay* and the associated multi-homing solution is also described in the same draft. That multi-homing solution is based on an interconnect ES that allows all-active and single-active multi-homed EVPN networks as well as local attachment circuits in the DC GWs (SAP/SDP-bindings).

In the used SR OS release, interconnect ESs are not supported. Therefore, an anycast solution is used to provide redundancy. This anycast solution is based on the two PE DC GWs in the redundant pair being configured to advertised MAC/IP and IMET routes with the same route key, so that the remote PEs will only pick up one of the two anycast DC GWs when sending unicast or BUM traffic, and no loop or packet duplication is created.

Figure 76 is an example of how multi-homing can be achieved for dual BGP-instance VPLS services. The figure also shows the EVPN destinations created and their direction (see the arrows). For instance, only one EVPN multicast destination is created for PE-1, PE-2, or PE-4. Therefore, BUM traffic sent by CE-1 will be sent via PE-2, PE-4, and PE-6 only, and no duplication or loops occur.

Figure 76 **EVPN Destinations Created on Multi-Homed Anycast DC GWs**



The following output shows the VPLS 1 configuration on PE-2 and PE-3 so that this anycast redundancy can be realized. The route distinguishers as well as the **incl-mcast-orig-ip** addresses must match between the two PEs in the redundant pair. VPLS 1 is configured on PE-2 as follows:

```
configure
    service
        vpls 1 customer 1 create
            vxlan vni 1 create
            exit
            bgp
                route-distinguisher 64500:1
            exit
            bgp 2
                route-distinguisher 64500:2
            exit
            bgp-evpn
                incl-mcast-orig-ip 23.23.23.23
                evi 1
                vxlan
                    no shutdown
                exit
                mpls
                    ingress-replication-bum-label
                    bgp-instance 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            no shutdown
```

The VPLS 1 configuration on PE-3 is as follows:

```
configure
    service
        vpls 1 customer 1 create
            vxlan vni 1 create
            exit
            bgp
                route-distinguisher 64500:1
            exit
            bgp 2
                route-distinguisher 64500:2
            exit
            bgp-evpn
                incl-mcast-orig-ip 23.23.23.23
                evi 1
                vxlan
                    no shutdown
                exit
                mpls
                    ingress-replication-bum-label
                    bgp-instance 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            no shutdown
```

The VPLS 1 configuration on PE-4 is as follows:

```
configure
    service
        vpls 1 customer 1 create
            vxlan vni 1 create
            exit
            bgp
                route-distinguisher 64501:1
            exit
            bgp 2
                route-distinguisher 64501:2
            exit
            bgp-evpn
                incl-mcast-orig-ip 45.45.45.45
                evi 1
                vxlan
                    no shutdown
                exit
                mpls
                    ingress-replication-bum-label
                    bgp-instance 2
                    auto-bind-tunnel
                    resolution any
                exit
                no shutdown
            exit
        exit
        no shutdown
```

The VPLS 1 configuration on PE-5 is as follows:

```
configure
    service
        vpls 1 customer 1 create
            vxlan vni 1 create
            exit
            bgp
                route-distinguisher 64501:1
            exit
            bgp 2
                route-distinguisher 64501:2
            exit
            bgp-evpn
                incl-mcast-orig-ip 45.45.45.45
                evi 1
                vxlan
                    no shutdown
                exit
                mpls
                    ingress-replication-bum-label
                    bgp-instance 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            no shutdown
```

Based on the preceding configuration example, the DC GWs behavior in this scenario is as follows:

- PE-2 and PE-3 both send IMET IR routes to the other PEs with the same route key but a different next-hop. The route key in IMET routes comprises [RD, Ethernet tag, originator-IP/length], which in this case will be [64500:1, 0, 23.23.23.23/32] for the EVPN-VXLAN IMET routes and [64500:2, 0, 23.23.23.23/32] for the EVPN-MPLS IMET routes.
- In the same way, PE-2 and PE-3 both send MAC/IP routes to the other PEs with the same route key but a different next-hop. The route key comprises [RD, Ethernet tag, MAC/MAC-length, IP/IP-length].

The configuration of the same **incl-mcast-orig-ip** address and RDs in both DC GWs enables the anycast solution due to the following:

- The configured originating IP (for example, 23.23.23.23 in PE-2 and PE-3) is not required to be a reachable IP address, which forces the remote PEs (or RRs if they exist) to select only one of the two DC GWs for BUM traffic (based on regular BGP selection). In this example, the remote PEs will select the PE-2 IMET route and create only one destination. The following output shows the IMET routes received by PE-1 (only the PE-2 route is used) and the created EVPN-VXLAN destination to PE-2. The same behavior could have been shown in the rest of the PEs.

```
A:PE-1# show router bgp routes evpn inclusive-mcast
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN Inclusive-Mcast Routes
===============================================================================
Flag  Route Dist.         OrigAddr
      Tag                 NextHop
-------------------------------------------------------------------------------
u*>i  64500:1             23.23.23.23
      0                   192.0.2.2

*>i   64500:1             23.23.23.23
      0                   192.0.2.3

-------------------------------------------------------------------------------
Routes : 2
===============================================================================


A:PE-1# show service id 1 vxlan
===============================================================================
Vxlan Src Vtep IP: N/A
===============================================================================
VPLS VXLAN, Ingress VXLAN Network Id: 1
Creation Origin: manual
Assisted-Replication: none
RestProtSrcMacAct: none

===============================================================================
VPLS VXLAN service Network Specifics
===============================================================================
Ing Net QoS Policy : none                            Vxlan VNI Id     : 1
Ingress FP QGrp    : (none)                          Ing FP QGrp Inst : (none)

===============================================================================
Egress VTEP, VNI
===============================================================================
VTEP Address                            Egress VNI  Num. MACs   Mcast Oper  L2
                                                                      State PBR
-------------------------------------------------------------------------------
192.0.2.2                               1           0           BUM   Up    No
-------------------------------------------------------------------------------
```

```
Number of Egress VTEP, VNI : 1
-------------------------------------------------------------------------------
===============================================================================
```

- Due to the same RD and originating IP configured on PE-2 and PE3 (similarly in PE-4 and PE-5), the DC GW redundant PEs will never establish an EVPN destination between each other. PE-2 only sets up EVPN multicast destinations to PE-1 and PE-4, as follows:

```
A:PE-2# show service id 1 vxlan
===============================================================================
Vxlan Src Vtep IP: N/A
===============================================================================
VPLS VXLAN, Ingress VXLAN Network Id: 1
Creation Origin: manual
Assisted-Replication: none
RestProtSrcMacAct: none

===============================================================================
VPLS VXLAN service Network Specifics
===============================================================================
Ing Net QoS Policy : none                            Vxlan VNI Id     : 1
Ingress FP QGrp    : (none)                          Ing FP QGrp Inst : (none)

===============================================================================
Egress VTEP, VNI
===============================================================================
VTEP Address                            Egress VNI  Num. MACs   Mcast Oper  L2
                                                                      State PBR
-------------------------------------------------------------------------------
192.0.2.1                               1           0           BUM   Up    No
-------------------------------------------------------------------------------
Number of Egress VTEP, VNI : 1
-------------------------------------------------------------------------------
===============================================================================


A:PE-2# show service id 1 evpn-mpls

===============================================================================
BGP EVPN-MPLS Dest
===============================================================================
TEP Address     Egr Label     Num. MACs   Mcast           Last Change
                 Transport
-------------------------------------------------------------------------------
192.0.2.4       262141        0           Yes             05/10/2017 08:10:15
                ldp
-------------------------------------------------------------------------------
Number of entries : 1
-------------------------------------------------------------------------------
===============================================================================
---snip---
```

- Likewise, when the two redundant PEs receive the same MAC/IP route, they will both re-advertise it with the same route key, forcing the remote PEs to pick up only one of the two (based on regular BGP selection) and create only one EVPN destination (if different from the multicast destination). In the following example, PE-6 advertised the CE-6 MAC address, that is, re-advertised by PE-4/PE-5 and then by PE-2/PE-3, but only one of the routes is selected at each hop. The following output shows that PE-1 selects the PE-2 MAC/IP route (see the "used" flag) and uses the existing EVPN destination to PE-2:

```
A:PE-1# show router bgp routes evpn mac
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN MAC Routes
===============================================================================
Flag  Route Dist.         MacAddr           ESI
      Tag                 Mac Mobility      Label1
                          Ip Address
                          NextHop
-------------------------------------------------------------------------------
u*>i  64500:1             00:ca:fe:ca:fe:06 ESI-0
      0                   Seq:0             VNI 1
                          N/A
                          192.0.2.2

*>i   64500:1             00:ca:fe:ca:fe:06 ESI-0
      0                   Seq:0             VNI 1
                          N/A
                          192.0.2.3

-------------------------------------------------------------------------------
Routes : 2
===============================================================================
```

```
A:PE-1# show service id 1 fdb detail
===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         00:ca:fe:ca:fe:06 vxlan:                   Evpn     05/10/17 09:09:36
                            192.0.2.2:1
-------------------------------------------------------------------------------
No. of MAC Entries: 1
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
```

```
A:PE-1# show service id 1 vxlan
```

```
===============================================================================
Vxlan Src Vtep IP: N/A
===============================================================================
VPLS VXLAN, Ingress VXLAN Network Id: 1
Creation Origin: manual
Assisted-Replication: none
RestProtSrcMacAct: none

===============================================================================
VPLS VXLAN service Network Specifics
===============================================================================
Ing Net QoS Policy : none                            Vxlan VNI Id     : 1
Ingress FP QGrp    : (none)                          Ing FP QGrp Inst : (none)

===============================================================================
Egress VTEP, VNI
===============================================================================
VTEP Address                            Egress VNI  Num. MACs   Mcast Oper  L2
                                                                      State PBR
-------------------------------------------------------------------------------
192.0.2.2                               1           1           BUM   Up    No
-------------------------------------------------------------------------------
Number of Egress VTEP, VNI : 1
-------------------------------------------------------------------------------
===============================================================================
```

- As shown in the preceding outputs, the EVPN destinations are always created to the IMET or MAC/IP route's BGP next-hops, which are still the system IP of the routers (they could have also been a loopback address). The BGP next-hops need to be reachable in their respective network: DC or WAN.

The Mandatory Use of BGP Policies in the Multi-Homed Anycast Solution

BGP policies must be configured in a multi-homed anycast solution, such as the one described in the previous section. Without policies, the following undesired behavior would happen:

- IMET routes with VXLAN encapsulation would be sent to the BGP peers in the MPLS network and IMET routes with MPLS encapsulation sent to BGP peers in the DC. The configured BGP policies will avoid that and make sure that the VXLAN routes are only sent to the DC and MPLS routes only to the WAN.

- MAC/IP routes received in the VXLAN BGP instance of a DC GW would be re-advertised to the redundant DC GW in the MPLS BGP instance and the redundant DC GW would re-advertise the same MAC again into the VXLAN instance, creating a control plane loop. The same thing would happen for MAC/IP routes received in an MPLS BGP instance. The configured BGP policies will prevent a DC GW from re-advertising MAC/IP routes received from the redundant DC GW.

While service-level BGP policies (**config>service>vpls>bgp>vsi-import/export**) may have been configured to prevent these loops and misbehavior, the use of BGP peer-level policies (**config>router>bgp>group>import/export**) is recommended due to the following reasons:

- Simplicity - BGP peer-level policies do not require any extra configuration at the service level, only at the BGP level.
- Scalability - BGP peer-level policies scale better than VSI-level policies, because the number of services where the VSI policies should be configured may be significant.

The following policies are configured in the example used in this chapter. No policies are needed in PE-1 and PE-6; only the DC GWs must be configured.

Following are the policies and how they are applied in PE-2:

```
configure
    router
        policy-options
            begin
            community "mpls" members "bgp-tunnel-encap:MPLS"
            community "vxlan" members "bgp-tunnel-encap:VXLAN"
            community "SOO-DCGW-23" members "origin:64500:23"

/* "drop SOO-DCGW-23" will drop any EVPN route that is received from PE-3,
the other DC GW in the pair. */

            policy-statement "drop SOO-DCGW-23"
                entry 10
                    from
                        community "SOO-DCGW-23"
                        family evpn
                    exit
                    action drop
                    exit
                exit
            exit

/* "allow only mpls and add SOO" has a twofold objective: avoids sending EVPN-VXLAN
routes to the MPLS network and marks the advertised EVPN routes with a Site-Of-
Origin extended community that identifies the DC GW pair. */

            policy-statement "allow only mpls and add SOO"
                entry 10
                    from
```

```
                        community "vxlan"
                        family evpn
                    exit
                    action drop
                    exit
                exit
                entry 20
                    from
                        family evpn
                    exit
                    action accept
                        community add "SOO-DCGW-23"
                    exit
                exit
            exit
```

/* In the same way, "allow only vxlan and add SOO" avoids sending EVPN-MPLS routes to the VXLAN network and marks the EVPN routes with a Site-Of-Origin extended community that identifies the DC GW pair. */

```
            policy-statement "allow only vxlan and add SOO"
                entry 10
                    from
                        community "mpls"
                        family evpn
                    exit
                    action drop
                    exit
                exit
                entry 20
                    from
                        family evpn
                    exit
                    action accept
                        community add "SOO-DCGW-23"
                    exit
                exit
            exit
            commit
        exit
```

/* The policies are properly applied at group level. */

```
        bgp
            family evpn
            vpn-apply-import
            vpn-apply-export
            rapid-withdrawal
            rapid-update evpn
            group "DC"
                type internal
                import "drop SOO-DCGW-23"
                export "allow only vxlan and add SOO"
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.3
                exit
            exit
            group "WAN"
```

```
                type internal
                import "drop SOO-DCGW-23"
                export "allow only mpls and add SOO"
                neighbor 192.0.2.4
                exit
                neighbor 192.0.2.5
                exit
            exit
            no shutdown
        exit
```

The same policies are configured on PE-3 (including the addition and filtering of the same Site-Of-Origin because PE-3 is part of the same DC GW pair):

```
configure
    router
        policy-options
            begin
            community "mpls" members "bgp-tunnel-encap:MPLS"
            community "vxlan" members "bgp-tunnel-encap:VXLAN"
            community "SOO-DCGW-23" members "origin:64500:23"
            policy-statement "drop SOO-DCGW-23"
                entry 10
                    from
                        community "SOO-DCGW-23"
                        family evpn
                    exit
                    action drop
                    exit
                exit
            exit
            policy-statement "allow only mpls and add SOO"
                entry 10
                    from
                        community "vxlan"
                        family evpn
                    exit
                    action drop
                    exit
                exit
                entry 20
                    from
                        family evpn
                    exit
                    action accept
                        community add "SOO-DCGW-23"
                    exit
                exit
            exit
            policy-statement "allow only vxlan and add SOO"
                entry 10
                    from
                        community "mpls"
                        family evpn
                    exit
                    action drop
                    exit
                exit
                entry 20
                    from
                        family evpn
                    exit
                    action accept
                        community add "SOO-DCGW-23"
                    exit
                exit
            exit
            commit
        exit
        bgp
            family evpn
            vpn-apply-import
            vpn-apply-export
            rapid-withdrawal
            rapid-update evpn
            group "DC"
                type internal
                import "drop SOO-DCGW-23"
                export "allow only vxlan and add SOO"
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
            exit
            group "WAN"
                type internal
                import "drop SOO-DCGW-23"
                export "allow only mpls and add SOO"
                neighbor 192.0.2.4
                exit
                neighbor 192.0.2.5
                exit
            exit
            no shutdown
        exit
```

PE-4 and PE-5 use the same BGP peer policies, but using a Site Of Origin extended community identifying the PE-4/PE-5 pair instead of the PE-2/PE-3 pair:

```
configure
    router
        policy-options
            begin
            community "mpls" members "bgp-tunnel-encap:MPLS"
            community "vxlan" members "bgp-tunnel-encap:VXLAN"
            community "SOO-DCGW-45" members "origin:64500:45"
---snip---
```

Dual BGP Instance VPLS Service Caveats

When two BGP instances are enabled on the same VPLS service, the following considerations apply:

- SAPs or SDP-bindings are not supported (therefore, no pw-template-binding is needed in the service). Any attempt to add a SAP or SDP-binding to a service with two BGP instances will be blocked by the CLI. For example:

```
configure service vpls 1 sap 1/2/1:1 create
MINOR: SVCMGR #7888 Cannot be configured/enabled with EVPN -
 saps not allowed when both vxlan and evpn-mpls are enabled
```

- Services that are not supported: R-VPLS, M-VPLS, I-VPLS, B-VPLS, or Etree VPLS
  - A consequence of not supporting R-VPLS is that no routes type 5 (IP-Prefix routes) are supported on dual BGP-instance services.
- Proxy-ARP/ND is not supported. ARP request or neighbor solicitation messages are usually only expected on SAP/SDP-bindings though.
- BGP multi-homing is not supported.
- Although the Assisted-Replication feature is supported on dual BGP-instance VPLS services, the Assisted-Replication configuration is only relevant to the VXLAN destinations. See section EVPN Route Handling in Dual BGP-Instance VPLS Services for some considerations about how EVPN handles IMET AR routes.
- The configuration of ESs is not supported for services with dual BGP instances (they cannot be configured because these services do not support SAP/SDP-bindings).

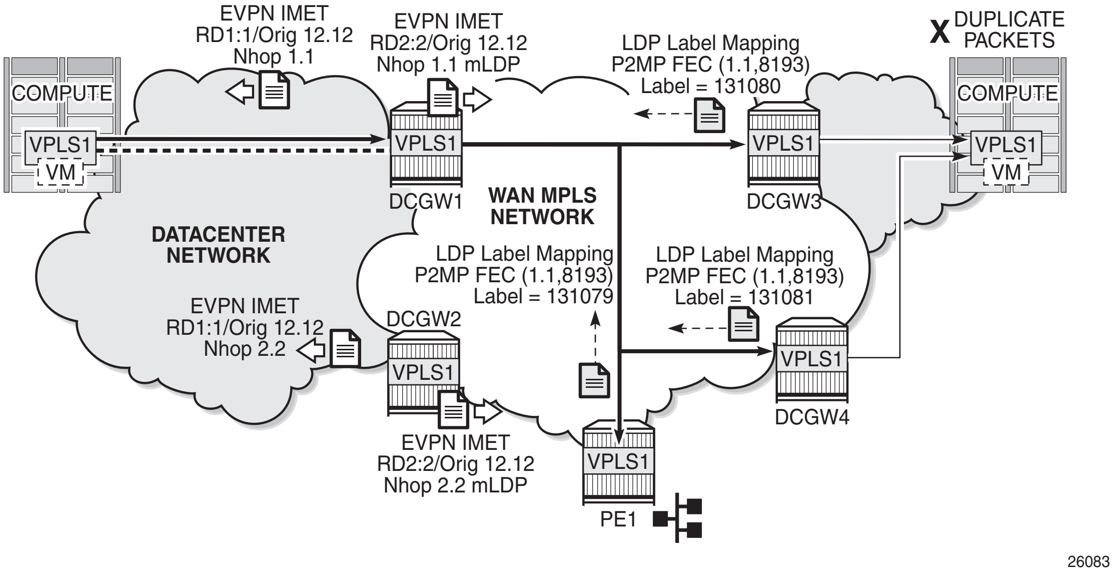
In addition to the preceding restrictions, some commands have a specific behavior when two BGP instances are configured:

- **config>service>vpls>bgp-evpn>[no] mac-advertisement** enables/disables the re-advertisement of MAC/IP routes in a BGP instance for MACs that have been learned in the other BGP instance in the service.
- **config>service>vpls>bgp-evpn>[no] unknown-mac-route** enables/disables the advertisement of the unknown MAC route (MAC 00:..:00) on the BGP-EVPN VXLAN instance. The unknown-mac-route is never sent to the BGP-EVPN MPLS instance.

The Use of Provider-tunnels on Multi-Homed Anycast Solutions

The use of provider-tunnels in dual BGP-instance VPLS services connecting multiple DCs is not recommended. Figure 77 shows the case where the same BGP-EVPN service is configured in redundant anycast DC GWs and mLDP is used in the MPLS instance. In this case, packet duplication may occur if the configuration is not done carefully.

Figure 77 **Use of Provider-tunnels Between Anycast DC GWs Create Packet Duplication**



When mLDP is used along with multiple anycast multi-homing DC GWs to send BUM traffic to remote PEs, but no BUM traffic between DCs is needed, the same originating IP must be used on all the DC GWs; otherwise, packet duplication may happen. In the example in Figure 77, each pair of DC GWs, DCGW1/DCGW2 and DCGW3/DCGW4, is configured with a different originating IP (**config>service>vpls>bgp-evpn> incl-mcast-orig-ip**):

- DCGW3 and DCGW4 will receive the IMET route with the same route key from DCGW1 and DCGW2.
- DCGW3 and DCGW4 will select only one route, which will usually be the same; for example, the DCGW1 IMET route.
- Because of that, both DCGW3 and DCGW4 will join the mLDP tree with root in DCGW1, creating packet duplication when DCGW1 sends BUM traffic.
- Remote PE nodes with a single MPLS instance will join the mLDP tree without any issue.

To avoid the packet duplication shown by the example of Figure 77, the same originating IP may be configured in the four DCGWs, while the RD is still different per pair. By doing that:

- In the example of Figure 77, DCGW3 and DCGW4 will never join any mLDP tree sourced from DCGW1 or DCGW2. This will prevent any packet duplication because a router will ignore IMET routes received with its own originating IP, regardless of the RD.
- PE-1 (a remote EVPN-MPLS PE) will still join the mLDP trees from the two DCs.
- The preceding configuration allows the use of mLDP as long as no BUM traffic is required between the two DCs. If BUM traffic is required between DCs, IR must be used.

Troubleshooting and Debugging

The following show and debug commands can be used in dual BGP-instance VPLS services:

- show router bgp routes evpn (and filters)
- show service evpn-mpls [<TEP ip-address>]
- show service vxlan [<TEP ip-address>]
- show service id bgp-evpn
- show service id evpn-mpls (and modifiers)
- show service id vxlan
- debug router bgp update
- log-id 99

See chapter EVPN for MPLS Tunnels and EVPN for VXLAN Tunnels (Layer 2) for a detailed description of these commands.

Also, in dual BGP-instance VPLS services, the **show service id bgp <bgp-instance>** command may help see the BGP parameters of each individual BGP instance (where instance 1 is always associated with VXLAN):

```
A:PE-2# show service id 1 bgp
  - bgp [<bgp-instance>]

 <bgp-instance>       : [1..2]


A:PE-2# show service id 1 bgp 1
===============================================================================
BGP Information
===============================================================================
Vsi-Import           : None
Vsi-Export           : None
Route Dist           : 64500:1
Oper Route Dist      : 64500:1
Oper RD Type         : configured
Rte-Target Import    : None                 Rte-Target Export: None
Oper RT Imp Origin   : derivedEvi           Oper RT Import   : 64500:1
Oper RT Exp Origin   : derivedEvi           Oper RT Export   : 64500:1
PW-Template Id       : None
-------------------------------------------------------------------------------
===============================================================================

A:PE-2# show service id 1 bgp 2
===============================================================================
BGP Information
===============================================================================
Vsi-Import           : None
Vsi-Export           : None
Route Dist           : 64500:2
Oper Route Dist      : 64500:2
Oper RD Type         : configured
Rte-Target Import    : None                 Rte-Target Export: None
Oper RT Imp Origin   : derivedEvi           Oper RT Import   : 64500:1
Oper RT Exp Origin   : derivedEvi           Oper RT Export   : 64500:1
-------------------------------------------------------------------------------
===============================================================================
```

Conclusion

As service providers deploy EVPN-MPLS in the network for Ethernet local area network (E-LAN) and Ethernet point-to-point (E-Line) services, the use of EVPN-MPLS to interconnect data centers is becoming a popular option. Based on *draft-ietf-bess-dci-evpn-overlay*, SR OS supports the connectivity of Layer 2 EVPN-VXLAN services to an EVPN-MPLS network. To implement that EVPN-MPLS Data Center Interconnect (DCI) solution, VPLS services support dual BGP instances, where EVPN-VXLAN and EVPN-MPLS can coexist simultaneously in the same VPLS service. This chapter describes the configuration of such dual BGP-instance VPLS services and how to deploy them in a redundant anycast DC GW configuration.

Fully Dynamic VSD Integration Model

This chapter provides information about fully dynamic virtualized service directory (VSD) integration model.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

Software requirements for this feature are SR OS release 13.0.R4 or later and Nuage Virtualized Services Platform (VSP) release 3.2.R1 or later. This configuration was tested on SR OS release 13.0.R4 and Nuage VSP release 3.2.R3.

➡ **Note:** Fully dynamic extensible messaging and presence protocol (XMPP) provisioning is not supported along with the dynamic business services feature in release 13.0. Both features are mutually exclusive.

➡ **Note:** Provisioning of filter entries from Virtualized Services Directory (VSD) is not supported in SR OS 13.0.

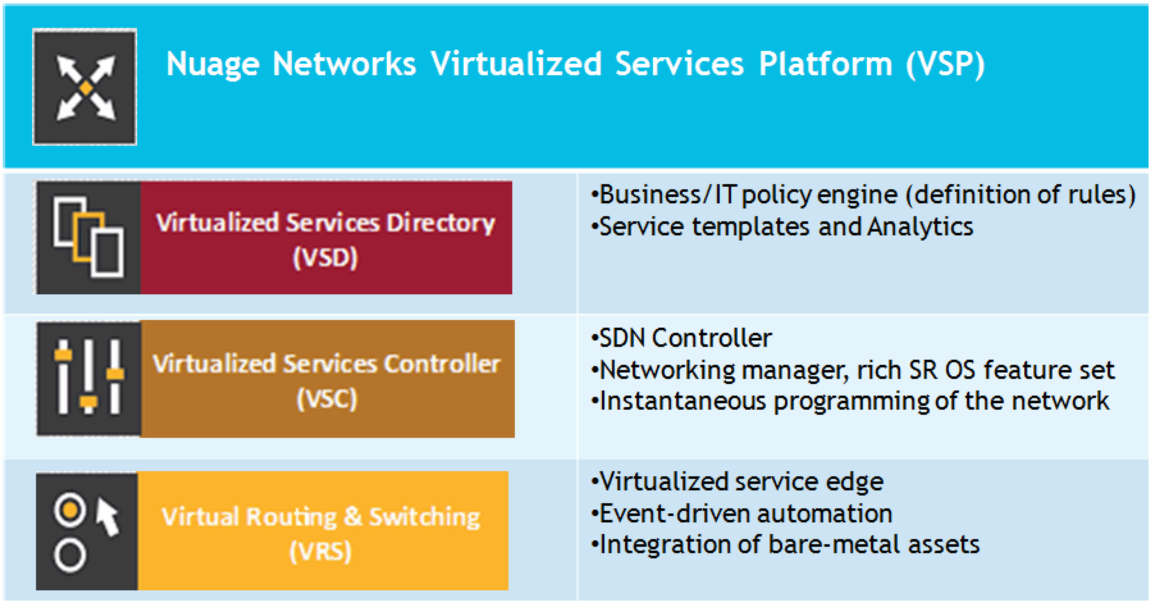
Nuage VSP conceptual knowledge and Nuage VSD operational knowledge are prerequisites. See the Nuage VSP user documentation for more information.

Overview

The Nuage VSP is a Software-Defined Networking (SDN) solution that provides data center (DC) network virtualization and automatically establishes connectivity between compute resources upon their creation. Leveraging programmable business logic and a powerful policy engine, the Nuage VSP provides an open and highly responsive solution that scales to meet the stringent needs of massive multi-tenant DCs. The Nuage VSP can be deployed over an existing DC IP network fabric, and has three main components:

Virtualized Services Directory (VSD), Virtualized Services Controller (VSC), and Virtual Routing and Switching (VRS), as displayed in Figure 78

Figure 78 **Nuage VSP overview**



Nuage Networks Virtualized Services Platform (VSP)

Virtualized Services Directory (VSD)

•Business/IT policy engine (definition of rules)
•Service templates and Analytics

Virtualized Services Controller (VSC)

•SDN Controller
•Networking manager, rich SR OS feature set
•Instantaneous programming of the network

Virtual Routing & Switching (VRS)

•Virtualized service edge
•Event-driven automation
•Integration of bare-metal assets

Virtualized Services Directory (VSD)

The Nuage VSD is a programmable policy and analytics engine. It provides a flexible and hierarchical network policy framework that enables IT administrators to define and enforce resource policies in a user-friendly manner.

The VSD contains a multi-tenant service directory, which supports role-based administration of users, computing, and network resources. The VSD also manages network resource assignments such as IP addresses and ACLs.

For the purpose of service assurance, the VSD allows the definition of sophisticated statistics rules, such as collection frequencies, rolling averages, and samples, as well as Threshold Crossing Alerts (TCA). When a TCA occurs, it will trigger an event that can be exported to external systems through a generic messaging bus. Statistics are aggregated over hours, days, and months, and stored in a Hadoop® analytics cluster to facilitate data mining and performance reporting.

The VSD runs as a number of processes in a virtual machine (VM) environment.

Virtualized Services Controller (VSC)

The Nuage VSC is an SDN controller. It functions as the robust network control plane for DCs, maintaining a full view of per-tenant network and service topologies. Through the VSC, virtual routing and switching constructs are established to program the network forwarding plane, the Nuage VRS, using the OpenFlow protocol.

The VSC communicates with the VSD policy engine using Extensible Messaging and Presence Protocol (XMPP). An ejabberd XMPP server/cluster is used to distribute messages between the VSD and VSC entities. Multiple VSC instances can be interconnected within and across DCs by leveraging Multi-Protocol Border Gateway Protocol (MP-BGP).

The VSC is based on the Service Router Operating System (SR OS) and runs in a virtual machine environment.

Virtual Routing and Switching (VRS)

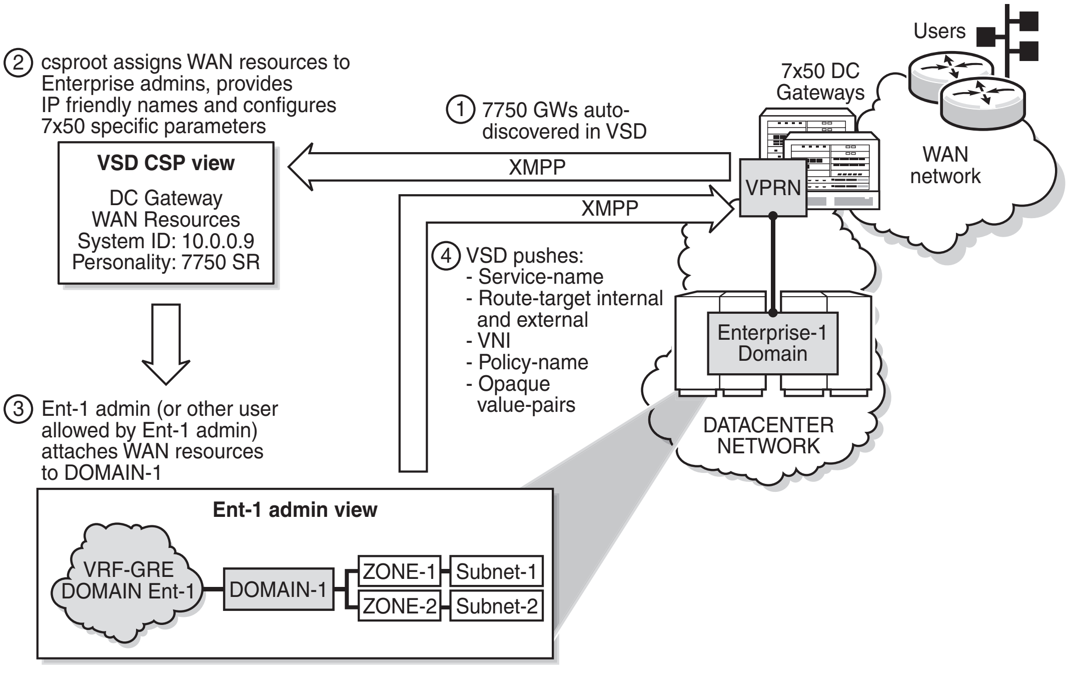
The Nuage VRS component is an enhanced Open vSwitch (OVS) implementation that constitutes the network forwarding plane. It encapsulates and de-encapsulates user traffic, enforcing L2 to L4 traffic policies as defined by the VSD. The VRS tracks Virtual Machine (VM) creation, migration, and deletion events in order to dynamically adjust network connectivity.

DC Gateway automated service provisioning

The first phase of VSD-7x50 integration was introduced in SR OS 12.0.R4. This phase included the development of an XMPP interface on the 7x50 SR and the integration in the Nuage XMPP architecture. This so-called Static + Dynamic (S-D) provisioning model allows the auto-provisioning of VPLS and VPRN route targets, as well as VPLS VNI (VXLAN Network Identifiers) on the 7x50 SR through the XMPP interface and the VSD interaction. The prerequisite in this model is the pre-configuration of the VPLS and VPRN services on the 7x50 through CLI or SNMP. This model is intended to be used in DC Gateways where the WAN and the DC are managed by different administrative entities. The DC administrator will use VSD to “attach” the already configured VPLS or VPRN service to the L2 or L3 domain in the DC.

The second phase of VSD-7x50 integration was introduced in SR OS 13.0R4. This phase supports the Fully Dynamic (F-D) provisioning model. The goal of this model is to avoid the prerequisite of pre-configuring the services on the 7x50 SR existing in the S-D provisioning model, since this model assumes that the service is completely owned by the DC administrator. The entire service will be auto-generated on the 7x50 SR as a result of the interaction with the VSD. Figure 79 shows the workflow of the F-D provisioning model.

Figure 79 **DC Gateway fully dynamic provisioning workflow**



1. As soon as the XMPP server is configured on the 7x50 DC Gateway, it is auto-discovered by the VSD.
2. The Cloud Service Provider (CSP) root user creates WAN services and assigns these resources to Enterprise administrators; for example, to the Ent-1 admin.
3. The Ent-1 admin sees the WAN service in the infrastructure resources and assigns permissions to certain user groups in Ent-1, who can consume these WAN resources by connecting to an L2/L3 domain.
4. As soon as the WAN service is added to an L2/L3 domain, the VSD pushes a list of parameters to the 7x50 DC Gateway, which uses a python script to construct the configuration of the WAN service. The list of parameters sent to the 7750 routers can include:
   - **service-name** (Service ID field in the WAN Service GUI) - used as VSD domain in the CLI
   - **config-type** (Config Type field in the WAN Service GUI) - DYNAMIC for F-D XMPP provisioning
   - **service-type** (based on combination of the Service Type field and IRB check box in the WAN Service GUI) - possible values: L2DOMAIN, L2DOMAIN-IRB, VRF-GRE, or VRF-VXLAN
   - **name** (name of the L2 domain in the VSD to which the WAN service is assigned, or BackHaulSubnet in the case of service-type VRF-VXLAN)
   - **service-policy** (service Policy in the WAN Service GUI field) - should match the python policy configured on the 7x50 DC Gateway
   - **vn-id** (VNI used for the Nuage overlay service) - dynamically supplied for VXLAN WAN services
   - **RT-I** (internal Route Target used for the Nuage overlay service)
   - **RT-E** (ext. Route Target in the WAN Service GUI field)
   - **metadata** (list of opaque parameters supplied in the Metadata section of the WAN Service GUI)

The dynamic provisioning of parameters is provided for the following VSD domain types (configured in the 7x50 DC Gateway):

l2-domain To attach a service at the gateway to an L2 (Ethernet) domain in the data center with no routing at the gateway, a VPLS service must be associated with a vsd-domain of type l2-domain.

l2-domain-irb To attach a service at the gateway to an L2 (Ethernet) domain in the data center with routing at the gateway, an R-VPLS service should be associated with a vsd-domain of type l2-domain-irb.

vrf-gre To attach a service at the gateway to an L3 domain (with GRE transport) in the data center, a VPRN service should be associated with a vsd-domain of type vrf-gre.

vrf-vxlan To attach a service at the gateway to an L3 domain (with VXLAN transport) in the data center, an R-VPLS service (with ip-route-advertisement enabled and linked to an EVPN-tunnel) should be associated with a vsd-domain of type vrf-vxlan.

This chapter will show examples of l2-domain, l2-domain-irb, and vrf-vxlan service type F-D provisioning, and focuses mostly on the 7x50 DC Gateway configuration. For a more detailed F-D provisioning workflow on the VSD UI, refer to the VSP User Guide.

Python script

The XMPP parameters supplied by the VSD are parsed by a python script on the 7x50 DC Gateway that dynamically provisions the VPLS and/or VPRN services provided for the Nuage overlay services.

The python script generates an executable CLI script based on the information received in the XMPP attributes. Three dynamic data service functions can be specified: **setup**, **modify**, or **teardown**. A fourth action, **revert**, is automatically invoked when the modify action fails:

- **setup** function: output = CLI to create a new dynamic data service.
- **teardown** function: output = CLI to delete an existing dynamic data service.
- **modify** function: output = CLI to change the parameters of an existing dynamic data service
- **revert** function: output = CLI to rollback the dynamic data service modify function actions in case of a modify failure

The python script uses the alc.dyn python module that contains a number of functions required to set up dynamic data services. To use the alc.dyn module, it must be imported into the python script:

```
from alc import dyn
```

The alc.dyn module contains a number of functions. Relevant alc.dyn functions for F-D XMPP provisioning are listed here:

- **dyn.action**(dictionary)
- **dyn.add_cli**(string)
- **dyn.select_free-id**(service-id)

The next sections provide a basic description of these functions. The alc.dyn module contains other functions that are not relevant for this feature. For a full list of the alc.dyn functions together with an extensive explanation of each function, refer to the RADIUS-Triggered Dynamic Data Service Provisioning chapter.

The trigger in the python script to execute a specific function is by calling the internal function **dyn.action**(d), where "d" is a python dictionary:

- d = { key : value, key : value, key : value, … , key : value }

For F-D XMPP provisioning, only 1 key:value pair is used and the key string must be set to "script".

The value is a tuple with the following comma separated values:

- (setup-function, modify-function, revert-function, teardown-function)

Setup and teardown functions are mandatory. Modify and revert functions are optional. If a modify function is defined, the revert function must also be defined. If no modify/revert function is required, the keyword "None" should be used instead.

The following two combinations are supported for F-D Dynamic XMPP provisioning:

1. without modify function:

```
d = {"script" : (setup_script, None, None, teardown_script)}
dyn.action(d)
```

2. with modify function (allows for changes in the WAN service on the VSD while the service is assigned to a domain):

```
d = {"script" : (setup_script, modify_script, revert_script, teardown_script)}
dyn.action(d)
```

When the configuration for a new service-name is received from the VSD, the vsd parameters and the opaque parameters string are concatenated into a single dictionary. The setup_script() is called and the dictionary is passed to the function. In this chapter, the dictionary will be named "vsdParams", but any other name would do. Within the python script:

- The VSD UI parameters are referenced as vsdParams['rt'], vsdParams['vni'], vsdParams['servicetype'], and so on.
- The metadata parameters are defined in an opaque string. For example, when the metadata string "rd=1:1,sap=1/1/1:1000" is supplied to the VSD WAN Service GUI, the format in the dictionary will be in the following format: "'metadata': 'rd=1:1,sap=1/1/1:1000 '".

To reference the metadata, the format is changed (trailing space is removed and parameters split up):

```
metadata = vsdParams['metadata']
    metadata = metadata.rstrip()
    metadata = dict(e.split('=') for e in metadata.split(','))
```

The individual metadata parameters can then be referenced in a similar way as the vsd parameters; for example, metadata['rd'], metadata['sap'], and so on.

When the startup script is executed, the *config>service>vsd>domain* is created outside the script context before running the actual script. The teardown script will remove the vsd domain. The domain-name is taken from the service-name supplied by the VSD ("Service ID" field in the WAN Service GUI - used as VSD domain in the CLI). When testing the script with the *tools perform python-script* command, the domain-name is taken from the domain-name command parameter (see Testing the python script section).

When subsequent configuration messages are received from the VSD, the new parameter list is again generated from the VSD message and compared to the last parameter list that was successfully executed.

- If the two strings are identical, no action is taken.
- If there is a difference between the strings, the *modify_script()* function is called. For example, the *modify_script()* function is set up to handle a change in the service-mtu.

If a configuration message is received from the VSD for an existing service-name with no VSD parameters, the *teardown_script()* is called.

If a *setup_script()* fails, the *teardown_script()* is called.

To generate CLI output in the python script, an internal function, **dyn.add_cli**(output-string), is available. It adds the specified output-string to the CLI script. Python enables the use of triple quotes to specify strings that span multiple lines. For example:

```
from alc import dyn
    dyn.add_cli("""
configure
    service
        ies %(svc_id)s customer 1 create
            service-name "%(inst)s"
            description "%(inst)s"
            no shutdown
        exit
    exit
exit
""" % d)
```

An internal function, **dyn.select_free_id**("service-id"), is available to select a free (unused) service identifier in the service-range specified in the dynamic-services context (see the Configuration section). If no service-range is configured, the python script fails when **dyn.select_free_id**("service-id") is called. The service-id is made available again after a successful teardown (removal) of the service.

XMPP

The Extensible Messaging and Presence Protocol (XMPP) is an open technology for real-time communication, using XML (Extensible Markup Language) as the base format for exchanging information. XMPP provides a way to send small pieces of XML from one entity to another in near real time. Although initially intended for Instant Messaging applications, it can be easily extended to be used in a DC environment.

In the Nuage solution, each XMPP client, including the 7x50 SR, is referred to with a JID(JabberID) in the following format: username@xmppserver.domain. The xmppserver.domainpoints to the XMPP server.

The Nuage VSP/7x50 DC Gateway solution uses the XMPP PubSub (Publish Subscribe) extension. This extension allows a user to subscribe to a node so that it can be notified whenever there is new or updated information available. The mechanism is used in this feature to auto-discover the username of the VSD JID. Additionally, the 7x50 will subscribe to a separate PubSub for each DC Gateway, to discover updates on specific domains. Subscriptions are confirmed periodically (every 15 min).

The 7x50 DC Gateway will periodically audit the VSD and request a DIFF list of F-D VSD domains. The VSD keeps a DIFF list of domains, which contains the F-D domain names for which the VSD has not received an info/query (IQ) request from the 7x50 for a long time. The DC Gateway periodically checks the info for each of its deployed dynamic services with an IQ request (every 16-24 min). A DIFF or FULL domain list audit can also be triggered with the *tools perform service vsd fd-domain-sync <full|diff> command.*

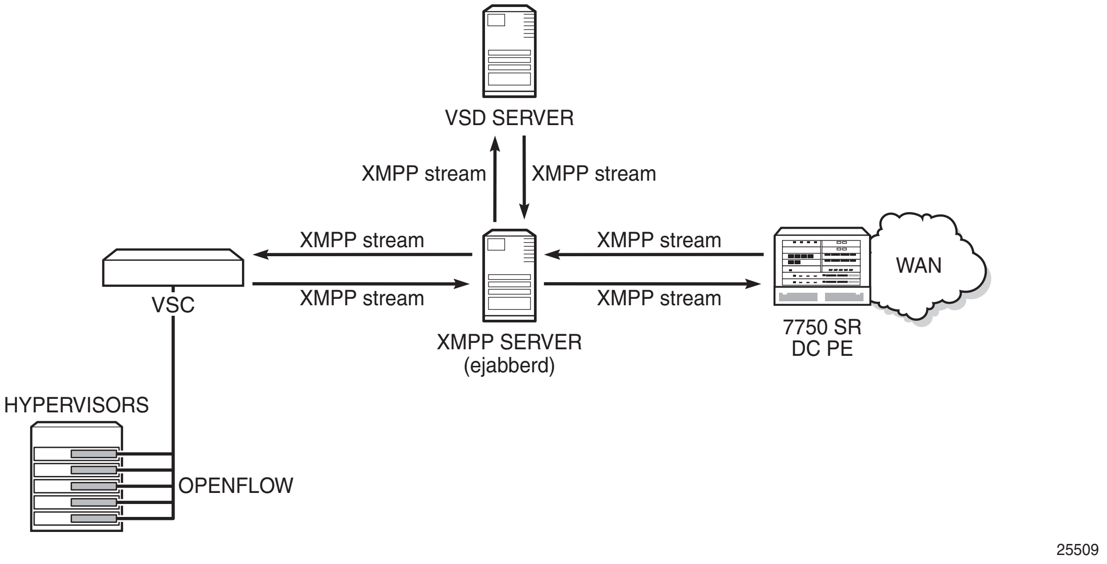
Configuration

This section describes the configuration that is required on the 7x50 DC Gateway for F-D XMPP provisioning.

The following figure shows the basic setup used and also illustrates the XMPP architecture in the data center. Although the VSD and XMPP servers are represented by a single server, a cluster of VSD servers (using the same database) and/or XMPP servers will be a very common configuration in a data center.

It is assumed that underlying IP connectivity and an IGP has already been configured in this setup.

Figure 80 **F-D XMPP provisioning setup**



XMPP configuration

To receive configuration parameters from the VSD, the 7x50 DC Gateway has to establish an XMPP client session with the XMPP server. Only one XMPP server can be configured.

When the XMPP server is properly configured, with no shutdown, the 7750 will try to establish a TCP session with the XMPP server through the management interface first. If it fails to establish communication, the 7750 will use in-band communication and will use its system IP as the source IP address.

To resolve the XMPP server fully qualified domain name (FQDN), provide a DNS server in the boot option file (bof) configuration and configure a dns-domain:

```
*A:pe-9>config>system# show bof
===============================================================================
BOF (Memory)
```

```
================================================================================
--- snipped ---
    primary-dns      138.203.39.47
    dns-domain       nuage.net
--- snipped ---
================================================================================
```

Then, configure the system-id of the DC Gateway that will be communicated to the VSD:

```
*A:pe-9>config>system# info
----------------------------------------------
#--------------------------------------------------
echo "vsd Configuration"
#--------------------------------------------------
        vsd
            system-id "pe9"
        exit
```

The next step is to configure the VSD server. The domain-name is the domain portion of the JID. The username is the username portion of the JID acting as an XMPP client. Ensure that the username uses all letters in lowercase (see SR OS 13.0 release notes). If no username is provided, an in-band registration will be provided, using the chassis MAC as username. The use of a password is optional:

```
*A:pe-9>config>system# info
#--------------------------------------------------
echo "Xmpp Configuration"
#--------------------------------------------------
        xmpp
            server vsd domain-name vsd1.nuage.net create username pe9
                no shutdown
            exit
        exit
----------------------------------------------
```

When the XMPP server has been configured, the state should move to "Functional":

```
*A:pe-9# show system xmpp server
================================================================================
XMPP Server Table
================================================================================
Name                            User Name           State
 XMPP FQDN                       Last State chgd     Admin State
--------------------------------------------------------------------------------
vsd                             pe9                 Functional
 vsd1.nuage.net                  0d 00:37:01         inService
--------------------------------------------------------------------------------
No. of XMPP server's: 1
================================================================================
```

XMPP Tx/Rx counters and other details can be obtained with the following command:

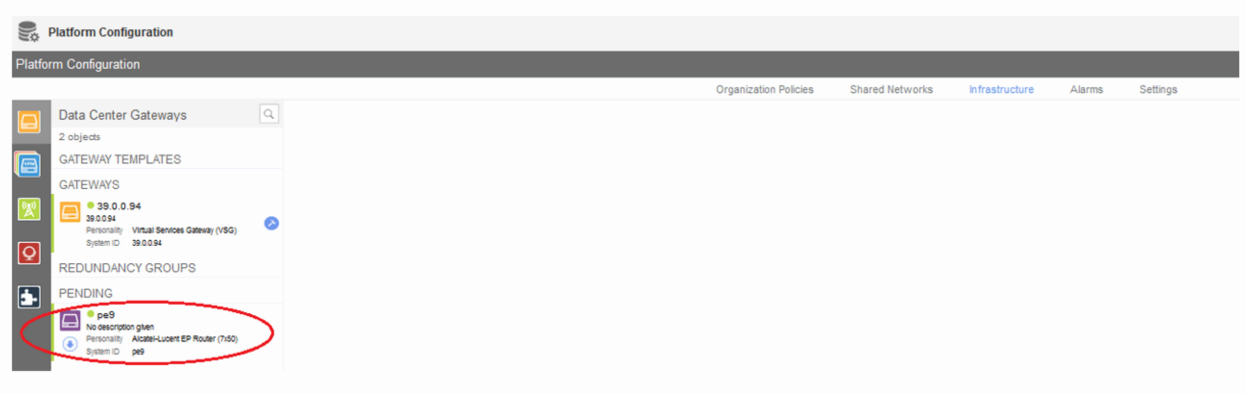
```
*A:pe-9# show system xmpp server "vsd"
===============================================================================
XMPP Server Table
===============================================================================
XMPP FQDN          : vsd1.nuage.net
XMPP Admin User    : pe9
XMPP Oper User     : pe9
State Lst Chg Since: 0d 00:07:43        State              : Functional
Admin State        : Up                 Connection Mode    : outOfBand
Auth Type          : md5
IQ Tx.             : 10                 IQ Rx.             : 10
IQ Error           : 0                  IQ Timed Out       : 0
IQ Min. Rtt        : 20 ms              IQ Max. Rtt        : 80 ms
IQ Ack Rcvd.       : 10
Push Updates Rcvd  : 1                  VSD list Upd Rcvd  : 1
Msg Tx.            : 3                  Msg Rx.            : 3
Msg Ack. Rx.       : 3                  Msg Error          : 0
Msg Min. Rtt       : 0 ms               Msg Max. Rtt       : 80 ms
Sub Tx.            : 1                  UnSub Tx.          : 0
Msg Timed Out      : 0
```

F-D XMPP provisioning uses the PubSub XMPP extension that allows each user to subscribe to a node, to be notified whenever that node gets new pieces of information or updated information.

The DC Gateway PubSub subscription state and subscriber name can be shown:

```
*A:pe-9# show system vsd
===============================================================================
VSD Information
===============================================================================
System Id               : pe9
GW Last Audit Tx Time   : 10/13/2015 13:56:34
Gateway Publish-Subscribe Information
-------------------------------------------------------------------------------
Subscribed              : True
Subscriber Name         : nuage_gateway_id_pe9
Last Subscription Time  : 10/13/2015 13:56:34
===============================================================================
```

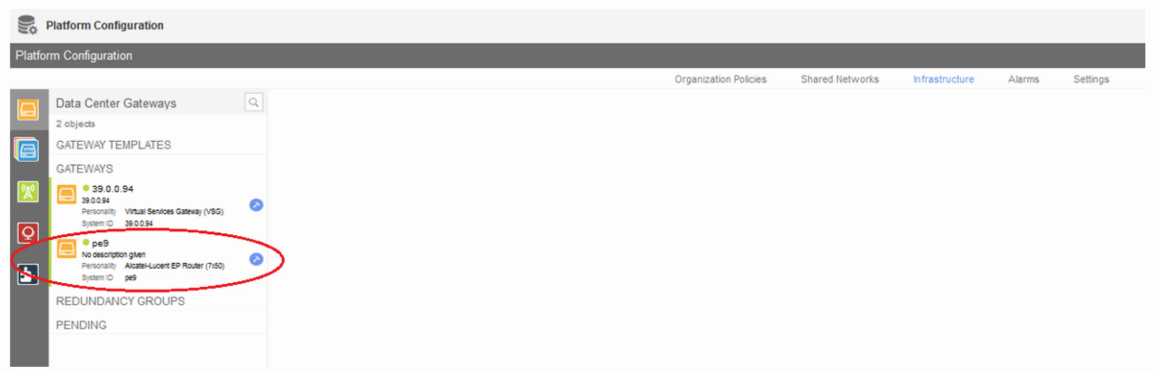
At the same time, the 7x50 DC Gateway will be announced as a pending gateway in the VSD:



39.0.0.94
39.0.0.94
Personality Virtual Services Gateway (VSG)
System ID 39.0.0.94

pe9
No description given
Personality Alcatel-Lucent EP Router (7x50)
System ID pe9

The gateway will be promoted to the available gateways group by clicking on the arrow below the pending gateway icon:



39.0.0.94
39.0.0.94
Personality Virtual Services Gateway (VSG)
System ID 39.0.0.94

pe9
No description given
Personality Alcatel-Lucent EP Router (7x50)
System ID pe9

BGP configuration

In the Nuage VSP solution, MP-BGP is used in the control plane to distribute MAC/IP information about the VMs. This information is distributed between the different VSCs, VSGs, and 7x50 DC Gateways. Configure MP-BGP on the 7x50 DC Gateway and the VSC/VSG (in this case, a VSG was used, but VSC is similar):

```
*A:pe-9>config>router# info
----------------------------------------------
#--------------------------------------------------
echo "IP Configuration"
#--------------------------------------------------
        interface "system"
            address 10.0.0.9/32
            no shutdown
        exit
```

```
--- snipped ---
        autonomous-system 65000
--- snipped ---
#--------------------------------------------------
echo "BGP Configuration"
#--------------------------------------------------
            min-route-advertisement 5
            rapid-withdrawal
            rapid-update evpn
            group "Nuage"
                family route-target evpn
                type internal
                neighbor 39.0.0.94
                exit
            exit
            no shutdown


*A:vsc1.nuage.net>config>router# info
----------------------------------------------
#--------------------------------------------------
echo "IP Configuration"
#--------------------------------------------------
--- snipped ---
        interface "system"
            address 39.0.0.94/32
            no shutdown
        exit
        autonomous-system 65000
--- snipped ---
#--------------------------------------------------
echo "BGP Configuration"
#--------------------------------------------------
        bgp
            family route-target evpn
            min-route-advertisement 5
            rapid-withdrawal
            rapid-update evpn
            group "internal"
                type internal
                neighbor 10.0.0.9
                    family evpn
                exit
            exit
            no shutdown
        exit
----------------------------------------------
```

In this setup, the family type "evpn" and "route-target" is used. The former is used to learn the EVPN route updates while the latter is restricting the 7x50 to only learn those MB-BGP routes for which it has a route target configured.

Note: To use vrf-gre domains, configure BGP family "vpn-ipv4" as well. Similarly, to use BGP-MH (for example, in case of redundant 7x50 DC Gateways with L2-domains), the use of BGP family "l2vpn" is required.

Verify that BGP peering is in the operational state:

```
*A:pe-9# show router bgp summary
===============================================================================
 BGP Router ID:10.0.0.9         AS:65000       Local AS:65000
===============================================================================
BGP Admin State         : Up          BGP Oper State              : Up
--- snipped ---
                   AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
39.0.0.94
                65000   76287    0 00h04m14s 0/0/0 (RouteTarget)
                         2634    0           0/0/0 (Evpn)
-------------------------------------------------------------------------------
```

Dynamic VSD Services range

F-D XMPP provisioning requires a reserved range of Service-IDs that can be used for dynamic data services. This configured range is no longer available for regular services configured via CLI/SNMP:

```
*A:pe-9>config>service# info
----------------------------------------------
        vsd
            service-range 64000 to 64999
        exit

*A:pe-9# show service vsd summary
===============================================================================
VSD Information
===============================================================================
Service Range
Start              : 64000                        End           : 64999
===============================================================================
No domain entries found
```

Python script

The python script that will build the dynamic services based on the VSD parameters obtained via XMPP can be stored locally on the CF or on a remote FTP server:

```
*A:pe-9>config>python# info
----------------------------------------------
        python-script "l2domain_services" create
            primary-url "ftp://*:*@138.203.15.48/./l2domain_service.py"
            no shutdown
        exit
```

The script is loaded into memory as soon as a no shutdown is performed and is reloaded with each *shutdown/no shutdown* action. Alternatively, a tools command can be used to reload the script:

```
*A:pe-9# tools perform python-script reload "l2domain_services"
```

In case incorrect python syntax is used in the script, an error message is displayed after a *no shut*down of the script (or a reload with the tools command), with an indication of the line where the error is located.

The details of the python script can be inspected:

```
*A:pe-9# show python python-script "l2domain_services"
===============================================================================
Python script "l2domain_services"
===============================================================================
Description    : (Not Specified)
Admin state    : inService
Oper state     : inService
Action on fail: drop
Protection     : none
Primary URL    : ftp://*:*@138.203.15.48/./l2domain_service.py
Secondary URL : (Not Specified)
Tertiary URL  : (Not Specified)
Active URL     : primary
Last changed  : 10/13/2015 09:54:26
===============================================================================
```

The contents of the python script can also be viewed. The contents of the script are shown in the "Test the python script" section:

```
*A:pe-9# show python python-script "l2domain_services" source-in-use
===============================================================================
Python script "l2domain_services"
===============================================================================
Admin state    : inService
Oper state     : inService
Primary URL    : ftp://*:*@138.203.15.48/./l2domain_service.py
Secondary URL : (Not Specified)
Tertiary URL  : (Not Specified)
Active URL     : primary
-------------------------------------------------------------------------------
Source (dumped from memory)
-------------------------------------------------------------------------------
      1 from alc import dyn
      2
      3 # example of metadata to be added in VSD WAN Service: "rd=1:1,sap=1/1/1:1000
"
      4
--- snipped ---
      8 def setup_script(vsdParams):
--- snipped ---
     70 def modify_script(vsdParams,setup_result):
--- snipped ---
```

```
    def revert_script(vsdParams,setup_result):
--- snipped ---
    def teardown_script(setupParams):
--- snipped ---
    d = {"script" : (setup_script, modify_script, revert_script, teardown_script
)}

    dyn.action(d)
===============================================================================
```

A list of CLI command nodes that can be used with python script function dyn.add_cli is provided with the *tools dump service vsd-services command-list* command. In general, all the 'leaf' commands under the nodes shown in the tools dump command, can be used with the python script.

Further restriction of CLI commands is possible by creating a separate CLI user for the XMPP interface and associate that user with a profile where the commands are limited.

The CLI user for the XMPP interface is configurable:

```
config>system>security>cli-script>authorization>
        vsd
[no] cli-user <username>
```

Python policy

Python scripts are called by a python policy that will be referred to in the VSD WAN services GUI in the Service Policy field.

Create a python policy (that will be referred to in the VSD GUI) and link the python policy to the python script:

```
*A:pe-9>config>python# info
----------------------------------------------
        python-policy "py-l2" create
            description "Python script to create L2 domains"
            vsd script "l2domain_services"
        exit


*A:pe-9# show python python-script "l2domain_services" association
===============================================================================
Python Script Association
===============================================================================
Policy                           Type                                       Dir
-------------------------------------------------------------------------------
py-l2                            vsdAccessRequest                       ingress
-------------------------------------------------------------------------------


*A:pe-9# show python python-policy "py-l2"
```

```
===============================================================================
Python policy "py-l2"
===============================================================================
Description               : Python script to create L2 domains
-------------------------------------------------------------------------------
Messages
-------------------------------------------------------------------------------
Type                                   Dir     Script
-------------------------------------------------------------------------------
vsdAccessRequest                       ingress l2domain_services
-------------------------------------------------------------------------------
```

Test the python script

The python script can be tested separately on the 7x50 DC Gateway; even before connecting it to the Nuage setup.

Some notes about the python script and creating dynamic services:

- The VSD (and *tools evaluate-script* command) will provide some compulsory parameters like:
  - domain-name
  - domain type (l2-domain|vrf-gre|vrf-vxlan|l2-domain-irb)
  - action (setup or teardown)
  - vni
  - RT (internal and external route-targets are provided)
- The VSD (and *tools evaluate-script* command) can provide extra metadata that is supplied in a text string and comma separated. For example, metadata "rd=1:1,sap=1/1/1:1000".
- The RT format supplied by the VSD though XMPP (that is, "x:x") differs from the RT format that can be used with the *tools evaluate-script* command (for example, "target:x:x"). For that reason, it can be useful to add the following check in the script:

```
if not rt.startswith ('target'):
     rt = "target:"+rt
```

- The VSD metadata string includes an empty space at the end. This can be removed with the following python command:

```
metadata = metadata.rstrip()
```

- If a configuration message is received from the VSD for an existing service-name with no VSD parameters, or if a setup_script() fails, the teardown_script() is called.

- At any point in the script, you can add *print* commands to check the status/content of various parameters.

The python policy and script can then be tested with the tools evaluate-script command. Example syntax has been added into the scripts for convenience.

Before testing the script, it will be useful to enable the following debugging:

```
debug
    python
        python-script "l2domain_services"
            script-all-info
        exit
    exit
    vsd
        scripts
            event
                cli
                errors
                executed-cmd
                warnings
                state-change
            exit
        exit
    exit
exit
```

In this section, a python example script for an l2-domain is used. The contents of the script are shown in the following text format. Examples for l2-domain-irb and vrf-vxlan type domains are shown in dedicated sections:

```
from alc import dyn
# example of metadata to be added in VSD WAN Service: "rd=1:1,sap=1/1/1:1000"
# example of tools cli to test this script: tools perform service vsd evaluate-
script domain-name "l2dom1" type l2-domain action setup policy "py-l2" vni 1234 rt-
i target:1:1 rt-e target:1:1 metadata "rd=1:1,sap=1/1/1:1000"

# teardown example cli: tools perform service vsd evaluate-script domain-
name "l2dom1" type l2-domain action teardown policy "py-l2" vni 1234 rt-
i target:1:1 rt-e target:1:1

def setup_script(vsdParams):

    print ("These are the VSD params: " + str(vsdParams))
    servicetype = vsdParams['servicetype']
    vni = vsdParams['vni']
    rt = vsdParams['rt']

# add "target:" if provisioned by VSD (VSD uses x:x format whereas tools command use
s target:x:x format)
    if not rt.startswith ('target'):
       rt = "target:"+rt
    metadata = vsdParams['metadata']

# remove trailing space at the end of the metadata
    metadata = metadata.rstrip()
    print ("VSD metadata" + str(metadata))


    metadata = dict(e.split('=') for e in metadata.split(','))
    print ("Modified metadata" + str(metadata))
    vplsSvc_id = dyn.select_free_id("service-id")
    print ("this is the free svc id picked up by the system: " + vplsSvc_id)
    if servicetype == "L2DOMAIN":

      rd = metadata['rd']
      sap_id = metadata['sap']
      print ('servicetype, VPLS id, rt, vni, rd, sap:', servicetype, vplsSvc_id, rt,
 vni, rd, sap_id)
      dyn.add_cli("""
        configure service
           vpls %(vplsSvc_id)s customer 1 create
              description vpls%(vplsSvc_id)s
              proxy-arp
                 dynamic-arp-populate
                 no shutdown
                 exit
              bgp
                 route-distinguisher %(rd)s
                 route-target %(rt)s
              exit
              vxlan vni %(vni)s create
              exit
              bgp-evpn
                  evi %(vplsSvc_id)s
                  vxlan
                      no shutdown
                  exit
              exit
              service-name evi%(vplsSvc_id)s
              sap %(sap_id)s create
              exit
              no shutdown
              exit
            exit
        exit
      """ % {'vplsSvc_id' : vplsSvc_id, 'vni' : vsdParams['vni'], 'rt' : rt, 'rd' :
metadata['rd'], 'sap_id' : sap_id})
      # L2DOMAIN returns setupParams: vplsSvc_id, servicetype, vni, sap
      return {'vplsSvc_id' : vplsSvc_id, 'servicetype' : servicetype, 'vni' : vni, '
sap_id' : sap_id}
#-----------------------------------------------------------------------------------
-------------
def modify_script(vsdParams,setup_result):

    print ("These are the setup_result params for modify_script: " + str(setup_resul
t))
    print ("These are the VSD params for modify_script: " + str(vsdParams))

    # remove trailing space at the end of the metadata
    metadata = vsdParams['metadata'].rstrip()

    print ("VSD metadata" + str(metadata))
    metadata = dict(e.split('=') for e in metadata.split(','))
    print ("Modified metadata" + str(metadata))
    # updating the setup_result dict
    setup_result.update(metadata)
    params = setup_result

    print ("The updated params from metadata and return from the setup result: " + s
tr(params))

    svc_mtu = params['svc-mtu']

    dyn.add_cli("""
      configure service
          vpls %(vplsSvc_id)s
            service-mtu %(svc-mtu)s
             exit
          exit
      exit
    """ %params )


    # Result is passed to teardown_script
    return params
#--------------------------------------------------------------------------------------
-------------
def revert_script(vsdParams,setup_result):
    print ("These are the setup_result params for revert_script: " + str(setup_resul
t))
    print ("These are the VSD params for revert_script: " + str(vsdParams))

    # When modify fails, the revert is called and then the teardown is called.
    # It is recommended to revert to same value as used in setup for the attributes
modified in modify_script.

    params = setup_result

    dyn.add_cli("""
      configure service
          vpls %(vplsSvc_id)s
            service-mtu 2000
             exit
          exit
      exit
    """ %params )

    # Result is passed to teardown_script
    return params

#--------------------------------------------------------------------------------------
-------------
def teardown_script(setupParams):
    print ("These are the teardown_script setupParams: " + str(setupParams))
    servicetype = setupParams['servicetype']
    if servicetype == "L2DOMAIN":
      dyn.add_cli("""
        configure service
            vpls %(vplsSvc_id)s
               no description
               proxy-arp shut
```

```
                no proxy-arp
                bgp-evpn
                    vxlan
                        shut
                    exit
                    no evi
                    exit
                no vxlan vni %(vni)s
                bgp
                   no route-distinguisher
                   no route-target
                exit
                no bgp
                no bgp-evpn
                sap %(sap_id)s
                   shutdown
                   exit
                no sap %(sap_id)s
                shutdown
                exit
                no vpls %(vplsSvc_id)s
             exit
         exit
      """ % {'vplsSvc_id' : setupParams['vplsSvc_id'], 'vni' : setupParams['vni'], '
sap_id' : setupParams['sap_id']})
      return setupParams

d = {"script" : (setup_script, modify_script, revert_script, teardown_script)}

dyn.action(d)
```

The script can be tested with the following command:

```
*A:pe-9# tools perform service vsd evaluate-script domain-name "l2dom1" type l2-
domain action setup policy "py-l2" vni 1234 rt-i target:1:1 rt-
e target:1:1 metadata "rd=1:1,sap=1/1/1:1000"
1 2015/10/15 09:51:16.08 UTC MINOR: DEBUG #2001 Base dyn-script req=setup
"dyn-script req=setup: l2dom1
  state=init->waiting-for-setup
"
2 2015/10/15 09:51:16.08 UTC MINOR: DEBUG #2001 Base dyn-script req=setup
"dyn-script req=setup: l2dom1
  state=waiting-for-setup->generating-setup
"
3 2015/10/15 09:51:16.08 UTC MINOR: DEBUG #2001 Base Python Output
"Python Output: l2domain_services
These are the VSD params: {'rt': 'target:1:1', 'rte': 'target:1:1', 'domain': ''
, 'servicetype': 'L2DOMAIN', 'vni': '1234', 'metadata': 'rd=1:1,sap=1/1/1:1000 '
}
VSD metadatard=1:1,sap=1/1/1:1000
Modified metadata{'rd': '1:1', 'sap': '1/1/1:1000'}
this is the free svc id picked up by the system: 64000
('servicetype, VPLS id, rt, vni, rd, sap:', 'L2DOMAIN', '64000', 'target:1:1', '
1234', '1:1', '1/1/1:1000')
"
4 2015/10/15 09:51:16.08 UTC MINOR: DEBUG #2001 Base Python Result
"Python Result: l2domain_services
"
```

```
5 2015/10/15 09:51:16.08 UTC MINOR: DEBUG #2001 Base dyn-script req=setup
"dyn-script req=setup: l2dom1
  state=generating-setup->executing-setup
"
6 2015/10/15 09:51:16.08 UTC MINOR: DEBUG #2001 Base dyn-script cli 1/1
"dyn-script cli 1/1: script:l2dom1(cli 705 dict 0->123)
        configure service
           vpls 64000 customer 1 create
              description vpls64000
              proxy-arp
                 dynamic-arp-populate
                 no shut
                 exit
              bgp
                 route-distinguisher 1:1
                 route-target target:1:1
              exit
              vxlan vni 1234 create
              exit
              bgp-evpn
                  evi 64000
                  vxlan
                      no shut
                  exit
              exit
              service-name evi64000
              sap 1/1/1:1000 create
              exit
              no shutdown
              exit
            exit
        exit
      "
7 2015/10/15 09:51:16.08 UTC MINOR: DEBUG #2001 Base dyn-script setup
"dyn-script setup: l2dom1 script:l2dom1 line 2
 configure service"
Success
--- snipped ---
24 2015/10/15 09:51:16.08 UTC MINOR: DEBUG #2001 Base dyn-script req=setup
"dyn-script req=setup: l2dom1
  state=executing-setup->established
"
```

At this moment a new VSD domain has been created as well as a new service:

```
*A:pe-9# show service vsd domain
===============================================================================
VSD Domain Table
===============================================================================
Name                                    Type           Origin    Admin
-------------------------------------------------------------------------------
l2dom1                                  l2Domain       vsd       inService
-------------------------------------------------------------------------------


*A:pe-9# show service vsd domain "l2dom1" association
============================================================
Service VSD Domain
============================================================
Svc Id      Svc Type  Domain Type    Domain Admin   Origin
------------------------------------------------------------
64000       vpls      l2Domain       inService      vsd
------------------------------------------------------------


*A:pe-9# show service vsd domain "l2dom1"
===============================================================================
VSD Information
===============================================================================
Name               : l2dom1
Description        : l2dom1
Type               : l2Domain                    Admin State   : inService
Last Error To Vsd  : (Not Specified)
Last Error From Vsd: (Not Specified)
Statistics
-------------------------------------------------------------------------------
Last Cfg Chg Evt   : 10/14/2015 16:02:54         Cfg Chg Evts  : 1
Last Cfg Update    : 10/14/2015 16:02:54         Cfg Upd Rcvd  : 1
Last Cfg Done      : 10/14/2015 16:02:54Cfg Success        : 1
      Cfg Failed    : 0
Last Recd Params   : script = {'domain' : '', 'vn
                   : i' : '1234', 'rt' : 'target:
                   : 1:1', 'rte' : 'target:1:1',
                   : 'servicetype' : 'L2DOMAIN',
                   : 'metadata' : 'rd=1:1,sap=1/1
                   : /1:1000 '}
Last Exec Params   : script = {'domain' : '', 'vn
                   : i' : '1234', 'rt' : 'target:
                   : 1:1', 'rte' : 'target:1:1',
                   : 'servicetype' : 'L2DOMAIN',
                   : 'metadata' : 'rd=1:1,sap=1/1
                   : /1:1000 '}
===============================================================================


*A:pe-9# show service service-using
===============================================================================
Services
===============================================================================
ServiceId    Type      Adm  Opr  CustomerId Service Name
-------------------------------------------------------------------------------
64000        VPLS      Up   Up   1          evi64000
2147483648   IES       Up   Down 1          _tmnx_InternalIesService
2147483649   intVpls   Up   Down 1          _tmnx_InternalVplsService
-------------------------------------------------------------------------------
```

Note: Service ID 2147483648 and 2147483649 are internal services that are always present on the 7x50. They are not relevant for this feature and will be truncated in other output examples in this document.

```
*A:pe-9# show service id 64000 all
===============================================================================
Service Detailed Information
===============================================================================
Service Id        : 64000               Vpn Id            : 0
Service Type      : VPLS
```

```
Name              : evi64000
Description       : vpls64000
Customer Id       : 1                   Creation Origin   : vsd
Last Status Change: 10/14/2015 16:02:54
Last Mgmt Change  : 10/14/2015 16:02:54
Etree Mode        : Disabled
Admin State       : Up                  Oper State        : Up
MTU               : 1514                Def. Mesh VC Id   : 64000
SAP Count         : 1                   SDP Bind Count    : 0
--- snipped ---
VSD Domain        : l2dom1
--- snipped ---
-------------------------------------------------------------------------------
BGP Information
-------------------------------------------------------------------------------
Vsi-Import        : None
Vsi-Export        : None
Route Dist        : 1:1
Oper Route Dist   : 1:1
Oper RD Type      : configured
Rte-Target Import : 1:1                 Rte-Target Export : 1:1
Oper RT Imp Origin: configured          Oper RT Import    : 1:1
Oper RT Exp Origin: configured          Oper RT Export    : 1:1
PW-Template Id    : None
-------------------------------------------------------------------------------
--- snipped ---
-------------------------------------------------------------------------------
SAP 1/1/1:1000
-------------------------------------------------------------------------------
Service Id         : 64000
SAP                : 1/1/1:1000               Encap             : q-tag
Description        : (Not Specified)
Admin State        : Up                       Oper State        : Up
--- snipped ---
```

Note: You cannot see the dynamic VSD services in the configuration nor can you edit their configuration under normal circumstances (this is discussed further in the next section).

```
*A:pe-9>config>service# info
----------------------------------------------
        customer 1 create
            description "Default customer"
        exit
        vsd
            service-range 64000 to 64999
        exit
----------------------------------------------
```

```
*A:pe-9# configure service vpls 64000
MINOR: CLI Modification of services created by a dynamic script is not allowed.
```

The service can be modified by adding/changing a service-mtu to the metadata. This will trigger the modify-script function in the python script. The following basic script is only an example of how a modify-script function operates. The script could be extended to modify other parameters as well; however this is out of the scope of this chapter:

```
*A:pe-9# tools perform service vsd evaluate-script domain-name "l2dom1" type l2-
domain action modify policy "py-l2" vni 1234 rt-i target:1:1 rt-
e target:1:1 metadata "rd=1:1,sap=1/1/1:1000,svc-mtu=2222"
25 2015/10/15 09:51:22.44 UTC MINOR: DEBUG #2001 Base dyn-script req=modify
"dyn-script req=modify: l2dom1
  state=established->waiting-for-modify
"
26 2015/10/15 09:51:22.44 UTC MINOR: DEBUG #2001 Base dyn-script req=modify
"dyn-script req=modify: l2dom1
  state=waiting-for-modify->generating-modify
"
27 2015/10/15 09:51:22.44 UTC MINOR: DEBUG #2001 Base Python Output
"Python Output: l2domain_services
These are the setup_result params for modify_script: {'servicetype': 'L2DOMAIN',
 'vplsSvc_id': '64000', 'vni': '1234', 'sap_id': '1/1/1:1000'}
These are the VSD params for modify_script: {'rt': 'target:1:1', 'rte': 'target:
1:1', 'domain': '', 'servicetype': 'L2DOMAIN', 'vni': '1234', 'metadata': 'rd=1:
1,sap=1/1/1:1000,svc-mtu=2222 '}
VSD metadatard=1:1,sap=1/1/1:1000,svc-mtu=2222
Modified metadata{'rd': '1:1', 'sap': '1/1/1:1000', 'svc-mtu': '2222'}
The updated params from metadata and return from the setup result: {'rd': '1:1',
 'servicetype': 'L2DOMAIN', 'svc-mtu': '2222', 'sap_id': '1/1/1:1000', 'sap': '1
/1/1:1000', 'vplsSvc_id': '64000', 'vni': '1234'}
"
Success
28 2015/10/15 09:51:22.44 UTC MINOR: DEBUG #2001 Base Python Result
"Python Result: l2domain_services
"
29 2015/10/15 09:51:22.44 UTC MINOR: DEBUG #2001 Base dyn-script req=modify
"dyn-script req=modify: l2dom1
  state=generating-modify->executing-modify
"
*A:pe-9#
30 2015/10/15 09:51:22.44 UTC MINOR: DEBUG #2001 Base dyn-script cli 1/1
"dyn-script cli 1/1: script:l2dom1(cli 123 dict 123->203)
      configure service
          vpls 64000
            service-mtu 2222
             exit
          exit
      exit
    "
31 2015/10/15 09:51:22.44 UTC MINOR: DEBUG #2001 Base dyn-script modify
"dyn-script modify: l2dom1 script:l2dom1 line 2
 configure service"
--- snipped ---
36 2015/10/15 09:51:22.44 UTC MINOR: DEBUG #2001 Base dyn-script req=commit
"dyn-script req=commit: l2dom1
  state=waiting-for-commit->established
"
```

The service-mtu has now been changed to 2222:

```
show service id 64000 all | match MTU
MTU               : 2222               Def. Mesh VC Id   : 64000
Admin MTU          : 9212                    Oper MTU          : 9212
```

The service can be removed with the teardown script:

```
*A:pe-9# tools perform service vsd evaluate-script domain-name "l2dom1" type l2-
domain action teardown policy "py-l2" vni 1234 rt-i target:1:1 rt-e target:1:1
37 2015/10/15 09:51:29.80 UTC MINOR: DEBUG #2001 Base dyn-script req=teardown
"dyn-script req=teardown: l2dom1
  state=established->waiting-for-teardown
"
38 2015/10/15 09:51:29.80 UTC MINOR: DEBUG #2001 Base dyn-script req=teardown
"dyn-script req=teardown: l2dom1
  state=waiting-for-teardown->generating-teardown
"
39 2015/10/15 09:51:29.80 UTC MINOR: DEBUG #2001 Base Python Output
"Python Output: l2domain_services
These are the teardown_script setupParams: {'servicetype': 'L2DOMAIN', 'svc-mtu'
: '2222', 'sap_id': '1/1/1:1000', 'vplsSvc_id': '64000', 'vni': '1234', 'rd': '1
:1', 'sap': '1/1/1:1000'}
"
40 2015/10/15 09:51:29.80 UTC MINOR: DEBUG #2001 Base Python Result
"Python Result: l2domain_services
"
41 2015/10/15 09:51:29.80 UTC MINOR: DEBUG #2001 Base dyn-script req=teardown
"dyn-script req=teardown: l2dom1
  state=generating-teardown->executing-teardown
"
42 2015/10/15 09:51:29.80 UTC MINOR: DEBUG #2001 Base dyn-script cli 1/1
"dyn-script cli 1/1: script:l2dom1(cli 709 dict 203->0)
        configure service
            vpls 64000
               no description
               proxy-arp shut
               no proxy-arp
               bgp-evpn
                   vxlan
                       shut
                   exit
                   no evi
                   exit
               no vxlan vni 1234
               bgp
                  no route-distinguisher
                  no route-target
               exit
               no bgp
               no bgp-evpn
               sap 1/1/1:1000
                  shutdown
                  exit
               no sap 1/1/1:1000
               shutdown
               exit
               no vpls 64000
            exit
        exit
      "
43 2015/10/15 09:51:29.80 UTC MINOR: DEBUG #2001 Base dyn-script teardown
"dyn-script teardown: l2dom1 script:l2dom1 line 2
 configure service"
--- snipped ---
63 2015/10/15 09:51:29.81 UTC MINOR: DEBUG #2001 Base dyn-script req=teardown
"dyn-script req=teardown: l2dom1
  state=executing-teardown->stopped
"
```

After the dynamic service has been torn down, the dynamic service and the VSD domain should not be present on the 7x50 DC Gateway:

```
*A:pe-9# show service service-using
===============================================================================
Services
===============================================================================
ServiceId    Type      Adm  Opr  CustomerId Service Name
-------------------------------------------------------------------------------
2147483648   IES       Up   Down 1          _tmnx_InternalIesService
2147483649   intVpls   Up   Down 1          _tmnx_InternalVplsService
-------------------------------------------------------------------------------
Matching Services : 2
-------------------------------------------------------------------------------
===============================================================================


*A:pe-9# show service vsd domain
No domain entries found
```

Editing dynamic VSD services

As indicated in the previous section, the dynamic VSD services CLI configuration cannot be shown or edited normally. However, under certain circumstances, it might be necessary to inspect/change/remove the configuration of a dynamic VSD service; for example, when the python VSD script was not using the correct syntax and the creation/deletion of the dynamic VSD service failed.

It is possible to edit the dynamic VSD services configuration by entering the *enable-vsd-config* mode. First, create a password, which is required to enter this mode:

```
*A:pe-9# configure system security password
*A:pe-9>config>system>security>password# vsd-password *****
```

Then, enter the enable-vsd-config mode. You will be asked for the previously configured password:

```
*A:pe-9# enable-vsd-config
Password:
```

Now you can edit the dynamic VSD services configuration and change/add/remove configuration:

```
*A:pe-9# configure service
*A:pe-9>config>service# info
----------------------------------------------
        customer 1 create
            description "Default customer"
        exit
        vsd
            domain l2dom1 type l2-domain create
                description "l2dom1"
                no shutdown
            exit
            service-range 64000 to 64999
        exit
        vpls 64000 customer 1 create
            description "vpls64000"
            vxlan vni 1234 create
            exit
            bgp
                route-distinguisher 1:1
            exit
            bgp-evpn
                evi 64000
                vxlan
                    no shutdown
                exit
                mpls
                    shutdown
                exit
            exit
            proxy-arp
                dynamic-arp-populate
                no shutdown
            exit
            stp
                shutdown
            exit
            service-name "evi64000"
            sap 1/1/1:1000 create
            exit
            vsd-domain "l2dom1"
            no shutdown
        exit
----------------------------------------------
```

In the enable-vsd-config mode, only dynamic VSD services can be edited, not regular CLI-based services:

```
*A:pe-9# configure service vpls 100 customer 1 create
MINOR: SVCMGR #1201 Invalid service-id - not reserved
```

After inspecting/editing the dynamic VSD services configuration, you should exit this mode again:

```
*A:pe-9# no enable-vsd-config
```

L2 VXLAN

An example python script for F-D XMPP provisioning of an L2 VXLAN type service (l2-domain) was provided in the "Testing the python script" section. In this section, the same script is used for provisioning via the VSD.

To dynamically provision this type of service, a few things must be configured on the VSD: (screenshots of this workflow are available in the VSP User Guide)

Create an L2 WAN service in the VSD:

- under Platform Config/Infrastructure, select the DC Gateway to add a WAN service
- select Service Type "layer 2" (no IRB)
- select "Dynamic" configuration type to allow Fully Dynamic provisioning
- under Service Policy, provide the python policy configured on the DC Gateway
- provide a Name and Service-ID (the Service-ID will be the name of the dynamically created service domain on the DC Gateway)

Add the metadata to the WAN service:

- right-click the WAN service and select "inspect"
- a dialog box appears, select the "Metadata" tag and add the metadata info "rd=1:1,sap=1/1/1:1000"

Add permissions to Enterprise1. The WAN service should now be visible in Enterprise1. Add permissions for a group of users to use this WAN service. Instantiate the L2 domain and attach the WAN service.

As soon as the WAN service is attached to the L2 domain, the VSD will send a notification via XMPP to the DC Gateway about the new Service-ID. The VSD will send an XMPP IQ request to the VSD to obtain the VSD service parameters.

Note: There is an 8 to 12 s delay. The command *tools perform service vsd domain refresh-config* can be used to expedite the request.

As soon as the DC Gateway receives this information, the python policy mentioned in the VSD service parameters is triggered and the VSD parameters are passed to the associated python script. The python script will then construct and execute the same configuration CLI and trigger similar debug information as is shown in the section "Testing the python script".

After the python script has completed successfully, the service can be inspected in a similar way as before.

- show service service-using
- show service id <*service-id*> all
- enter enable-vsd-config mode if required and inspect the service config

MAC addresses can now be learned in the Nuage VSC/VSG and sent via MP-BGP to the DC Gateway:

```
*A:pe-9# show service id 64000 fdb detail
===============================================================================
Forwarding Database, Service 64000
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
64000     1e:50:01:01:00:01 vxlan:                   Evpn     10/14/15 16:10:44
                            39.0.0.94:1006636
-------------------------------------------------------------------------------
```

In case HyperVisors (HVs) with VRS are deployed or a Host vPORT has been connected to a VSG, the EVPN MAC/Route (type 2) will also include the IP address of the VM/Host, in which case the DC Gateway can perform a proxy-arp function. For more information about EVPN-VXLAN features, refer to the chapters EVPN for VXLAN Tunnels (Layer 2) and EVPN for VXLAN Tunnels (Layer 3).

```
*A:pe-9# show service id 64000 proxy-arp detail
-------------------------------------------------------------------------------
Proxy Arp
-------------------------------------------------------------------------------
Admin State       : enabled
Dyn Populate      : enabled
Age Time          : disabled            Send Refresh      : disabled
Table Size        : 250                 Total             : 1
Static Count      : 0                   EVPN Count        : 1
Dynamic Count     : 0                   Duplicate Count   : 0
Dup Detect
-------------------------------------------------------------------------------
Detect Window     : 3 mins              Num Moves         : 5
Hold down         : 9 mins
Anti Spoof MAC    : None
EVPN
-------------------------------------------------------------------------------
Garp Flood        : enabled             Req Flood         : enabled
-------------------------------------------------------------------------------
```

```
===============================================================================
VPLS Proxy Arp Entries
===============================================================================
IP Address          Mac Address         Type      Status    Last Update
-------------------------------------------------------------------------------
10.32.78.100        1e:50:01:01:00:01   evpn      active    07/22/2015 10:05:25
-------------------------------------------------------------------------------
Number of entries : 1
===============================================================================
```

The svc-ID belonging to the Service-ID configured on the VSD can be easily obtained:

```
*A:pe-9# show service vsd domain "L2-service-1" association
============================================================
Service VSD Domain
============================================================
Svc Id      Svc Type  Domain Type    Domain Admin   Origin
------------------------------------------------------------
64000       vpls      l2Domain       inService      vsd
------------------------------------------------------------
Number of entries: 1
```

The associated RT/RD/VNI information can then be displayed with the *show service id <service-id> all* command, as shown previously.

An overview of the VTEPs that the DC Gateway shares this service with, and the corresponding VNIs is available:

```
*A:pe-9# show service id 64000 vxlan
===============================================================================
VPLS VXLAN, Ingress VXLAN Network Id: 1006636
===============================================================================
Egress VTEP, VNI
===============================================================================
VTEP Address           Egress VNI     Num. MACs    Mcast   Oper State   L2 PBR
-------------------------------------------------------------------------------
39.0.0.94              1006636        0            Yes     Up           No
-------------------------------------------------------------------------------
```

An overview of all the Service-IDs and associated VNIs that the DC Gateway has in common with a VSC or VSG can be shown with the following command:

```
*A:pe-9# show service vxlan 39.0.0.94
===============================================================================
VXLAN Tunnel Endpoint: 39.0.0.94
===============================================================================
Egress VNI                    Service Id               Oper State
-------------------------------------------------------------------------------
1006636                       64000                    Up
-------------------------------------------------------------------------------
===============================================================================
```

Relevant MAC/IP/VNI/RT/NH information is also in the EVPN BGP RIB:

```
*A:pe-9# show router bgp routes evpn mac
===============================================================================
 BGP Router ID:10.0.0.9         AS:65000       Local AS:65000
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP EVPN MAC Routes
===============================================================================
Flag  Route Dist.         MacAddr           ESI
      Tag                 Mac Mobility      Ip Address
                                            NextHop
                                            Label1
-------------------------------------------------------------------------------
u*>i  65534:39790         1e:50:01:01:00:01 ESI-0
      1006636             Static            10.32.78.100
                                            39.0.0.94
                                            VNI 1006636
-------------------------------------------------------------------------------
```

The WAN service can be detached from the L2 domain in the VSD GUI, if required.

This triggers similar debug information on the DC Gateway as the *tools perform service vsd evaluate teardown* command shown in the "Testing the python script" section.

After deleting the WAN service in the VSD GUI, the VPLS service and the service domain is removed from the DC Gateway.

L2 VXLAN IRB

An example python script for F-D XMPP provisioning of an L2 VXLAN IRB type service (l2-domain-irb) is as follows:

```
from alc import dyn
# example of metadata to be added in VSD WAN Service: "rd=2:2,sap=1/1/1:1000,vprnAS=
65000,vprnRD=65000:1,vprnRT=target:65000:1,vprnLo=1.1.1.1,irbGW=10.32.78.1/24"
# example of tools cli to test this script: tools perform service vsd evaluate-
script domain-name "l2domIRB1" type l2-domain-irb action setup policy "py-l2-
irb" vni  1234 rt-i target:2:2 rt-
e target:2:2 metadata "rd=2:2,sap=1/1/1:1000,vprnAS=65000,vprnRD=65000:1,vprnRT=targ
et:65000:1,vprnLo=1.1.1.1,irbGW=10.32.78.1/24"
# teardown example cli: tools perform service vsd evaluate-script domain-
name "l2domIRB1" type l2-domain-irb action teardown policy "py-l2-irb" vni 1234 rt-
i target:2:2  rt-e target:2:2
def setup_script(vsdParams):

    print ("These are the VSD params: " + str(vsdParams))
    servicetype = vsdParams['servicetype']
    vni = vsdParams['vni']
    rt = vsdParams['rt']
# add "target:" if provisioned by VSD (VSD uses x:x format whereas tools command use
s target:x:x format)
    if not rt.startswith ('target'):
       rt = "target:"+rt
    metadata = vsdParams['metadata']
# remove trailing space at the end of the metadata
    metadata = metadata.rstrip()
    print ("VSD metadata" + str(metadata))
    metadata = dict(e.split('=') for e in metadata.split(','))
    print ("Modified metadata" + str(metadata))
    vplsSvc_id = dyn.select_free_id("service-id")
    vprnSvc_id = dyn.select_free_id("service-id")
    print ("this are the free svc ids picked up by the system: VPLS:" + vplsSvc_id +
 " + VPRN:" + vprnSvc_id)
    if servicetype == "L2DOMAIN-IRB":

      rd = metadata['rd']
      sap_id = metadata['sap']
      vprn_AS = metadata ['vprnAS']
      vprn_RD = metadata ['vprnRD']
      vprn_RT = metadata ['vprnRT']
      vprn_Lo = metadata ['vprnLo']
      irb_GW = metadata ['irbGW']
      print ('servicetype, VPLS id, rt, vni, rd, sap, VPRN id, vprn_AS, vprn_RD, vpr
n_RT, vprn_Lo, irb_GW:', servicetype, vplsSvc_id, rt, vni, rd, sap_id, vprnSvc_id, v
prn_AS, vprn_RD, vprn_RT, vprn_Lo, irb_GW)
      dyn.add_cli("""
        configure service
           vpls %(vplsSvc_id)s customer 1 create
              allow-ip-int-bind
              exit
              description vpls%(vplsSvc_id)s
              bgp
                 route-distinguisher %(rd)s
                 route-target %(rt)s
              exit
              vxlan vni %(vni)s create
              exit
              bgp-evpn
                  evi %(vplsSvc_id)s
                  vxlan
                      no shut
                  exit
              exit
              service-name vpls%(vplsSvc_id)s
              sap %(sap_id)s create
              exit
              no shutdown
              exit
            exit
        exit
        configure service
           vprn %(vprnSvc_id)s customer 1 create
              autonomous-system %(vprn_AS)s
              route-distinguisher %(vprn_RD)s
              vrf-target %(vprn_RT)s
              interface "irbvpls-%(vplsSvc_id)s" create
                 address %(irb_GW)s
                 vpls "vpls%(vplsSvc_id)s"
                 exit
              exit
              interface "lo1" create
                 address %(vprn_Lo)s/32
                 loopback
              exit
              no shutdown
           exit
        exit
      """ % {'vplsSvc_id' : vplsSvc_id, 'vprnSvc_id' : vprnSvc_id, 'vni' : vsdParams
['vni'], 'rt' : rt, 'rd' : metadata['rd'], 'sap_id' : sap_id, 'vprn_AS' : vprn_AS, '
vprn_RD' : vprn_RD, 'vprn_RT' : vprn_RT, 'vprn_Lo' : vprn_Lo, 'irb_GW' : irb_GW})
      # L2DOMAIN-
IRB returns setupParams: vplsSvc_id, vprnSvc_id, servicetype, vni, sap, vprn_AS, vpr
n_RD, vprn_RT, vprn_Lo
      return {'vplsSvc_id' : vplsSvc_id, 'vprnSvc_id' : vprnSvc_id, 'servicetype' :
servicetype, 'vni' : vni, 'sap_id' : sap_id, 'vprn_AS' : vprn_AS, 'vprn_RD' : vprn_R
D, 'vprn_RT' : vprn_RT, 'vprn_Lo' : vprn_Lo, 'irb_GW': irb_GW}
#-----------------------------------------------------------------------------------
-------------
def teardown_script(setupParams):
    print ("These are the teardown_script setupParams: " + str(setupParams))
    servicetype = setupParams['servicetype']
    if servicetype == "L2DOMAIN-IRB":
      dyn.add_cli("""
        configure service
            vpls %(vplsSvc_id)s
               no description
               bgp-evpn
                   vxlan
                       shut
                   exit
                   no evi
                   exit
               no vxlan vni %(vni)s
               bgp
                  no route-distinguisher
                  no route-target
               exit
               no bgp
               no bgp-evpn
               sap %(sap_id)s
                  shutdown
                  exit
               no sap %(sap_id)s
               shutdown
               exit
            no vpls %(vplsSvc_id)s
            vprn %(vprnSvc_id)s
               interface lo1 shutdown
               no interface lo1
               interface "irbvpls-%(vplsSvc_id)s"
                  no vpls
                  shutdown
                  exit
               no interface "irbvpls-%(vplsSvc_id)s"
               shutdown
            exit
        no vprn %(vprnSvc_id)s
        exit
      """ % {'vplsSvc_id' : setupParams['vplsSvc_id'], 'vprnSvc_id' : setupParams['v
prnSvc_id'], 'vni' : setupParams['vni'], 'sap_id' : setupParams['sap_id']})
      return setupParams
d = {"script" : (setup_script, None, None, teardown_script)}
dyn.action(d)
```

The python script and policy are configured in a similar way as the previous example:

```
*A:pe-9# configure python
*A:pe-9>config>python# info
----------------------------------------------
        python-script "l2domain-irb_services" create
            primary-url "ftp://*:*@138.203.15.48/./l2domainIRB_service.py"
            no shutdown
        exit
        python-policy "py-l2-irb" create
            description "Python script to create L2-IRB domains"
            vsd script "l2domain-irb_services"
        exit
----------------------------------------------
```

It is also possible to create a python script that covers different domain types. The relevant part of the script is then addressed by using the following if-statement in the script:

```
    servicetype = vsdParams.get('servicetype')
        if servicetype == "L2DOMAIN-IRB":
           --- snipped ---
```

On the VSD, the following has to be provided:

(screenshots of this workflow are available in the VSP User Guide)

Create an L2 WAN service in the VSD:

- under Platform Config/Infrastructure, select the DC Gateway to add a WAN service
- select Service Type "layer 2" and select "IRB"
- select "Dynamic" configuration type to allow Fully Dynamic provisioning
- under Service Policy, provide the python policy configured on the DC Gateway
- provide a Name and Service-ID (the Service-ID will be the name of the dynamically created service domain on the DC Gateway)

Add the metadata to the WAN service:

- right-click the WAN service and select "inspect"
- a pop-up dialog box appears; select the "Metadata" tag and add the metadata info; for example,

"rd=2:2,sap=1/1/1:1000,vprnAS=65000,vprnRD=65000:1,vprnRT=target:65000:1,vprnLo=1.1.1.1,irbGW=10.32.78.1/24"

Add permissions to Enterprise1. The WAN service should now be visible in Enterprise1.

Add permissions for a group of users to use this WAN service. Instantiate an L2 domain and attach the WAN service.

After the script has completed, there should be two new services created:

```
*A:pe-9# show service service-using
===============================================================================
Services
===============================================================================
ServiceId    Type      Adm  Opr  CustomerId Service Name
-------------------------------------------------------------------------------
64000        VPLS      Up   Up   1          vpls64000
64001        VPRN      Up   Up   1

*A:pe-9# show service id 64000 all
===============================================================================
Service Detailed Information
===============================================================================
Service Id        : 64000               Vpn Id            : 0
Service Type      : VPLS
Name              : vpls64000
Description       : vpls64000
Customer Id       : 1                   Creation Origin   : vsd
Last Status Change: 07/22/2015 11:15:36
Last Mgmt Change  : 07/22/2015 11:15:36
Etree Mode        : Disabled
Admin State       : Up                  Oper State        : Up
MTU               : 1514                Def. Mesh VC Id   : 64000
SAP Count         : 1                   SDP Bind Count    : 0
--- snipped ---
VSD Domain        : L2-IRB-Service-1
--- snipped ---

-------------------------------------------------------------------------------
BGP Information
-------------------------------------------------------------------------------
Vsi-Import        : None
Vsi-Export        : None
Route Dist        : 2:2
Oper Route Dist   : 2:2
Oper RD Type      : configured
Rte-Target Import : 65534:6985          Rte-Target Export : 65534:6985
Oper RT Imp Origin: configured          Oper RT Import    : 65534:6985
Oper RT Exp Origin: configured          Oper RT Export    : 65534:6985
PW-Template Id    : None
-------------------------------------------------------------------------------
--- snipped ---
-------------------------------------------------------------------------------
SAP 1/1/1:1000
-------------------------------------------------------------------------------
```

```
Service Id         : 64000
SAP                : 1/1/1:1000              Encap            : q-tag
Description        : (Not Specified)
Admin State        : Up                      Oper State       : Up
--- snipped ---
===============================================================================
VPLS VXLAN, Ingress VXLAN Network Id: 1006636
===============================================================================
Egress VTEP, VNI
===============================================================================
VTEP Address           Egress VNI     Num. MACs    Mcast   Oper State   L2 PBR
-------------------------------------------------------------------------------
39.0.0.94              1006636        0            Yes     Up           No
-------------------------------------------------------------------------------
--- snipped ---


*A:pe-9# show service id 64001 all
===============================================================================
Service Detailed Information
===============================================================================
Service Id        : 64001               Vpn Id           : 0
Service Type      : VPRN
Name              : (Not Specified)
Description       : (Not Specified)
Customer Id       : 1                   Creation Origin  : vsd
Last Status Change: 07/22/2015 11:15:36
Last Mgmt Change  : 07/22/2015 11:15:36
Admin State       : Up                  Oper State       : Up

Route Dist.       : 65000:1             VPRN Type        : regular
Oper Route Dist   : 65000:1
Oper RD Type      : configured
AS Number         : 65000               Router Id        : 10.0.0.9
ECMP              : Enabled             ECMP Max Routes  : 1
--- snipped ---
-------------------------------------------------------------------------------
Interface
-------------------------------------------------------------------------------
If Name           : irbvpls-64000
Admin State       : Up                  Oper (v4/v6)     : Up/Down
Protocols         : None
IP Addr/mask      : 10.32.78.1/24       Address Type     : Primary
--- snipped ---
Routed VPLS Details
VPLS Name        : vpls64000
Binding Status   : Up
--- snipped ---
-------------------------------------------------------------------------------
Interface
-------------------------------------------------------------------------------
If Name           : lo1
Admin State       : Up                  Oper (v4/v6)     : Up/Down
Protocols         : None
IP Addr/mask      : 1.1.1.1/32          Address Type     : Primary
```

The dynamically created configuration can be inspected in enable-vsd-config mode (only enter the enable-vsd-config mode when absolutely required):

```
*A:pe-9>config>service# info
----------------------------------------------
        customer 1 create
            description "Default customer"
        exit
        vsd
            domain L2-IRB-Service-1 type l2-domain-irb create
                description "L2-IRB-Service-1"
                no shutdown
            exit
            service-range 64000 to 64999
        exit
        vpls 64000 customer 1 create
            description "vpls64000"
            allow-ip-int-bind
            exit
            vxlan vni 1006636 create
            exit
            bgp
                route-distinguisher 2:2
            exit
            bgp-evpn
                evi 64000
                vxlan
                    no shutdown
                exit
                mpls
                    shutdown
                exit
            exit
            stp
                shutdown
            exit
            service-name "vpls64000"
            sap 1/1/1:1000 create
            exit
            vsd-domain "L2-IRB-Service-1"
            no shutdown
        exit
        vprn 64001 customer 1 create
            autonomous-system 65000
            route-distinguisher 65000:1
            auto-bind-tunnel
                resolution any
            exit
            vrf-target target:65000:1
            interface "irbvpls-64000" create
                address 10.32.78.1/24
                vpls "vpls64000"
                exit
            exit
            interface "lo1" create
                address 1.1.1.1/32
                loopback
            exit
            vsd-domain "L2-IRB-Service-1"
            no shutdown
        exit
----------------------------------------------
```

MAC addresses can now be learned in the Nuage VSP/VSG and sent via MP-BGP to the DC Gateway:

```
*A:pe-9# show service id 64000 fdb detail
===============================================================================
Forwarding Database, Service 64000
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
64000     1e:50:01:01:00:01 vxlan:                   Evpn     07/22/15 11:26:15
                            39.0.0.94:1006636
64000     1e:e2:ff:00:f9:3d cpm                      Intf     07/22/15 11:15:36
-------------------------------------------------------------------------------
```

The second MAC address is the GW-MAC that the VM or VSG-connected host will use to reach the interface on the VPRN:

```
*A:pe-9# show router 64001 route-table
===============================================================================
Route Table (Service: 64001)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
1.1.1.1/32                                    Local   Local     00h17m10s  0
       lo1                                                          0
10.32.78.0/24                                 Local   Local     00h17m10s  0
       irbvpls-64000                                                0

*A:pe-9# show router 64001 arp
===============================================================================
ARP Table (Service: 64001)
===============================================================================
IP Address      MAC Address       Expiry    Type   Interface
-------------------------------------------------------------------------------
10.32.78.1      1e:e2:ff:00:f9:3d 00h00m00s Oth[I] irbvpls-64000
10.32.78.100    1e:50:01:01:00:01 03h53m22s Dyn[I] irbvpls-64000
1.1.1.1         1e:e2:ff:00:00:00 00h00m00s Oth    lo1
```

Similar commands as shown in the previous section are available to obtain relevant information:

- *show service vsd domain "L2-IRB-Service-1" association* to obtain svc-IDs
- *show service id <id> all* to obtain RT/RD/VNI values
- *show service id <id> vxlan* to obtain VTEPs in the VPLS service
- *show service vxlan <vtep-ip>* to obtain svc-ID and VNI information
- *show router bgp routes evpn mac* to obtain MAC/IP/VNI/RT/NH information

The VM or VSG-connected Host should be able to ping the loopback interface of the VPRN service:

```
*A:ce1# ping 1.1.1.1 source 10.32.78.100
PING 1.1.1.1 56 data bytes
64 bytes from 1.1.1.1: icmp_seq=1 ttl=64 time=1.67ms.
64 bytes from 1.1.1.1: icmp_seq=2 ttl=64 time=1.83ms.
```

L3 VXLAN

An example python script for F-D XMPP provisioning of an L3 VXLAN type service (vrf-vxlan) is:

```
from alc import dyn

# example of metadata to be added in VSD WAN Service: "rd=3:3,vprnAS=65000,vprnRD=65
000:1,vprnRT=target:65000:1,vprnLo=1.1.1.1"

# example of tools cli to test this script: tools perform service vsd evaluate-
script domain-name "l3dom1" type vrf-vxlan action setup policy "py-vrf-
vxlan" vni 1234 rt-i target:3:3 rt-
e target:3:3 metadata "rd=3:3,vprnAS=65000,vprnRD=65000:1,vprnRT=target:65000:1,vprn
Lo=1.1.1.1"

# teardown example cli: tools perform service vsd evaluate-script domain-
name "l3dom1" type vrf-vxlan action teardown policy "py-vrf-vxlan" vni 1234 rt-
i target:3:3 rt-e target:3:3

def setup_script(vsdParams):

    print ("These are the VSD params: " + str(vsdParams))
    servicetype = vsdParams['servicetype']
    vni = vsdParams['vni']
    rt = vsdParams['rt']

# add "target:" if provisioned by VSD (VSD uses x:x format whereas tools command use
s target:x:x format)
    if not rt.startswith ('target'):
       rt = "target:"+rt
    metadata = vsdParams['metadata']

# remove trailing space at the end of the metadata
    metadata = metadata.rstrip()
    print ("VSD metadata" + str(metadata))
    metadata = dict(e.split('=') for e in metadata.split(','))
    print ("Modified metadata" + str(metadata))
    vplsSvc_id = dyn.select_free_id("service-id")
    vprnSvc_id = dyn.select_free_id("service-id")
    print ("this are the free svc ids picked up by the system: VPLS:" + vplsSvc_id +
 " + VPRN:" + vprnSvc_id)

    if servicetype == "VRF-VXLAN":

      rd = metadata['rd']
      vprn_AS = metadata ['vprnAS']
      vprn_RD = metadata ['vprnRD']
      vprn_RT = metadata ['vprnRT']
      vprn_Lo = metadata ['vprnLo']
      print ('servicetype, VPLS id, rt, vni, rd, VPRN id, vprn_AS, vprn_RD, vprn_RT,
 vprn_Lo:', servicetype, vplsSvc_id, rt, vni, rd, vprnSvc_id, vprn_AS, vprn_RD, vprn
_RT, vprn_Lo)
      dyn.add_cli("""
        configure router policy-options
           begin
             community _VSD_%(vplsSvc_id)s members %(rt)s
             policy-statement vsi_import_%(vplsSvc_id)s
                entry 10
                   from
                      family evpn
                      community _VSD_%(vplsSvc_id)s
                      exit
                   action accept
                   exit
                exit
             exit
             policy-statement vsi_export_%(vplsSvc_id)s
                entry 10
                   from
                      family evpn
                      exit
                      action accept
                      community add _VSD_%(vplsSvc_id)s
                      exit
                   exit
              exit
           commit
        exit

        configure service
           vpls %(vplsSvc_id)s customer 1 create
              allow-ip-int-bind
                 exit
              description vpls%(vplsSvc_id)s
              bgp
                 route-distinguisher %(rd)s
                 vsi-import vsi_import_%(vplsSvc_id)s
                 vsi-export vsi_export_%(vplsSvc_id)s
                 exit
              vxlan vni %(vni)s create
                 exit
              bgp-evpn
                 ip-route-advertisement
                 vxlan
                     no shut
                     exit
                 exit
              service-name vpls%(vplsSvc_id)s
              no shutdown
              exit
            exit
        exit

        configure service
           vprn %(vprnSvc_id)s customer 1 create
              autonomous-system %(vprn_AS)s
              route-distinguisher %(vprn_RD)s
              vrf-target %(vprn_RT)s
              interface "vpls-%(vplsSvc_id)s" create
                 vpls "vpls%(vplsSvc_id)s" evpn-tunnel
                 exit
              interface "lo1" create
                 address %(vprn_Lo)s/32
                 loopback
                 exit
              no shutdown
              exit
        exit

      """ % {'vplsSvc_id' : vplsSvc_id, 'vprnSvc_id' : vprnSvc_id, 'vni' : vsdParams
['vni'], 'rt' : rt, 'rd' : metadata['rd'], 'vprn_AS' : vprn_AS, 'vprn_RD' : vprn_RD,
 'vprn_RT' : vprn_RT, 'vprn_Lo' : vprn_Lo})
      # VRF-
VXLAN returns setupParams: vplsSvc_id, vprnSvc_id, servicetype, vni, vprn_AS, vprn_R
D, vprn_RT, vprn_Lo
      return {'vplsSvc_id' : vplsSvc_id, 'vprnSvc_id' : vprnSvc_id, 'servicetype' :
servicetype, 'vni' : vni, 'vprn_AS' : vprn_AS, 'vprn_RD' : vprn_RD, 'vprn_RT' : vprn
_RT, 'vprn_Lo' : vprn_Lo}
#----------------------------------------------------------------------------------
-------------

def teardown_script(setupParams):
    print ("These are the teardown_script setupParams: " + str(setupParams))
    servicetype = setupParams['servicetype']
    if servicetype == "VRF-VXLAN":
      dyn.add_cli("""
        configure service
            vpls %(vplsSvc_id)s
               no description
               bgp-evpn
                   vxlan
                       shut
                       exit
                   no evi
                   exit
               no vxlan vni %(vni)s
               bgp
                  no route-distinguisher
                  no route-target
                  exit
               no bgp
               no bgp-evpn
               shutdown
               exit
            no vpls %(vplsSvc_id)s
            vprn %(vprnSvc_id)s
               interface lo1 shutdown
               no interface lo1
               interface "vpls-%(vplsSvc_id)s"
                  vpls "vpls%(vplsSvc_id)s"
                     no evpn-tunnel
                     exit
                  no vpls
                  shutdown
                  exit
               no interface "vpls-%(vplsSvc_id)s"
               shutdown
```

```
                exit
            no vprn %(vprnSvc_id)s
            exit
            configure router policy-options
                begin
                no community _VSD_%(vplsSvc_id)s
                no policy-statement vsi_import_%(vplsSvc_id)s
                no policy-statement vsi_export_%(vplsSvc_id)s
                commit
            exit

      """ % {'vplsSvc_id' : setupParams['vplsSvc_id'], 'vprnSvc_id' : setupParams['v
prnSvc_id'], 'vni' : setupParams['vni']})
      return setupParams

d = {"script" : (setup_script, None, None, teardown_script)}

dyn.action(d)
```

The python script and policy are configured in a similar way as the previous example:

```
*A:pe-9# configure python
*A:pe-9>config>python# info
----------------------------------------------
        python-script "vrf-vxlan_services" create
            primary-url "ftp://*:*@138.203.15.48/./vrf-vxlan_service.py"
            no shutdown
        exit
        python-policy "py-vrf-vxlan" create
            description "Python script to create vrf-vxlan domains"
            vsd script "l3vxlan"
        exit
----------------------------------------------
```

The following steps are required on the VSD to provision the WAN service:

(screenshots of this workflow are available in the VSP User Guide)

Create an L3 WAN service in the VSD:

- under Platform Config / Infrastructure, select the DC Gateway to add a WAN service
- select Service Type "layer 3"
- select "Dynamic" configuration type to allow Fully Dynamic provisioning
- under Service Policy, provide the python policy configured on the DC Gateway
- provide a Name and Service-ID (the Service-ID will be the name of the dynamically created service domain on the DC Gateway)

Add the metadata to the WAN service:

- right-click the WAN service and select "inspect"

- a pop-up dialog box appears; select the "Metadata" tag and add the metadata info; for example,

  "rd=3:3,vprnAS=65000,vprnRD=65000:1,vprnRT=target:65000:1,vprnLo=1.1.1.1"

Add permissions to Enterprise1. The WAN service should now be visible in Enterprise1.

Add permissions for a group of users to use this WAN service. Instantiate an L3 domain and attach the WAN service.

After the script has completed, there should be two new services:

```
*A:pe-9# show service service-using
===============================================================================
Services
===============================================================================
ServiceId    Type      Adm  Opr  CustomerId Service Name
-------------------------------------------------------------------------------
64000        VPLS      Up   Up   1          vpls64000
64001        VPRN      Up   Up   1

*A:pe-9# show service id 64000 all
===============================================================================
Service Detailed Information
===============================================================================
Service Id        : 64000               Vpn Id            : 0
Service Type      : VPLS
Name              : vpls64000
Description       : vpls64000
Customer Id       : 1                   Creation Origin   : vsd
Last Status Change: 10/14/2015 18:38:44
Last Mgmt Change  : 10/14/2015 18:38:44
Etree Mode        : Disabled
Admin State       : Up                  Oper State        : Up
MTU               : 1514                Def. Mesh VC Id   : 64000
SAP Count         : 0                   SDP Bind Count    : 0
--- snipped ---
VSD Domain        : L3-service-1
--- snipped ---

-------------------------------------------------------------------------------
BGP Information
-------------------------------------------------------------------------------
Vsi-Import        : vsi_import_64000
Vsi-Export        : vsi_export_64000
Route Dist        : 3:3
Oper Route Dist   : 3:3
Oper RD Type      : configured
Rte-Target Import : None                Rte-Target Export : None
Oper RT Imp Origin: vsi                 Oper RT Import    : None
Oper RT Exp Origin: vsi                 Oper RT Export    : None
PW-Template Id    : None
-------------------------------------------------------------------------------
--- snipped ---
```

```
===============================================================================
VPLS VXLAN, Ingress VXLAN Network Id: 119281

===============================================================================
Egress VTEP, VNI
===============================================================================
VTEP Address           Egress VNI     Num. MACs    Mcast   Oper State   L2 PBR
-------------------------------------------------------------------------------
39.0.0.94              119281         1            No      Up           No
-------------------------------------------------------------------------------
--- snipped ---
```

The script dynamically creates a VSI-import and VSI-export policy and links it to an RT that was dynamically created by the VSC/VSG:

```
*A:pe-9# show router policy
===============================================================================
Route Policies
===============================================================================
Policy                           Description
-------------------------------------------------------------------------------
vsi_export_64000
vsi_import_64000
-------------------------------------------------------------------------------
Policies : 2
===============================================================================


*A:pe-9# show router policy "vsi_import_64000"
    entry 10
        from
            community "_VSD_64000"
            family evpn
        exit
        action accept
        exit
    exit


*A:pe-9# show router policy "vsi_export_64000"
    entry 10
        from
            family evpn
        exit
        action accept
            community add "_VSD_64000"
        exit
    exit


*A:pe-9# show router policy community "_VSD_64000"
community "_VSD_64000" members "65534:38619"


*A:pe-9# show service id 64001 all
===============================================================================
Service Detailed Information
===============================================================================
Service Id        : 64001               Vpn Id            : 0
Service Type      : VPRN
```

```
Name              : (Not Specified)
Description       : (Not Specified)
Customer Id       : 1                    Creation Origin   : vsd
Last Status Change: 10/14/2015 18:38:44
Last Mgmt Change  : 10/14/2015 18:38:44
Admin State       : Up                   Oper State        : Up

Route Dist.       : 65000:1              VPRN Type         : regular
Oper Route Dist   : 65000:1
Oper RD Type      : configured
AS Number         : 65000                Router Id         : 10.0.0.9
ECMP              : Enabled              ECMP Max Routes   : 1
Auto Bind Tunnel
Resolution        : any
--- snipped ---
Vrf Target        : target:65000:1
--- snipped ---
-------------------------------------------------------------------------------
Interface
-------------------------------------------------------------------------------
If Name           : vpls-64000
Admin State       : Up                   Oper (v4/v6)      : Up/Down
Protocols         : None

IP Addr/mask      : Not Assigned
--- snipped ---
Routed VPLS Details
VPLS Name        : vpls64000
Binding Status   : Up
--- snipped ---
-------------------------------------------------------------------------------
Interface
-------------------------------------------------------------------------------
If Name           : lo1
Admin State       : Up                   Oper (v4/v6)      : Up/Down
Protocols         : None
IP Addr/mask      : 1.1.1.1/32           Address Type      : Primary
```

The dynamically created configuration can be inspected in enable-vsd-config mode (only enter the enable-vsd-config mode when required):

```
*A:pe-9>config>service# info
----------------------------------------------
        customer 1 create
            description "Default customer"
        exit
        vsd
            domain L3-service-1 type vrf-vxlan create
                description "L3-service-1"
                no shutdown
            exit
            service-range 64000 to 64999
        exit
        vpls 64000 customer 1 create
            description "vpls64000"
            allow-ip-int-bind
            exit
            vxlan vni 119281 create
            exit
            bgp
                route-distinguisher 3:3
                vsi-export "vsi_export_64000"
                vsi-import "vsi_import_64000"
            exit
            bgp-evpn
                ip-route-advertisement
                vxlan
                    no shutdown
                exit
                mpls
                    shutdown
                exit
            exit
            stp
                shutdown
            exit
            service-name "vpls64000"
            vsd-domain "L3-service-1"
            no shutdown
        exit
        vprn 64001 customer 1 create
            autonomous-system 65000
            route-distinguisher 65000:1
            auto-bind-tunnel
                resolution any
            exit
            vrf-target target:65000:1
            interface "vpls-64000" create
                vpls "vpls64000"
                    evpn-tunnel
                exit
            exit
            interface "lo1" create
                address 1.1.1.1/32
                loopback
            exit
            vsd-domain "L3-service-1"
            no shutdown
        exit
----------------------------------------------
```

The EVPN tunnel NH-MAC addresses can now be learned in the Nuage VSP/VSG and sent via MP-BGP to the DC Gateway:

```
*A:pe-9# show service id 64000 fdb detail
Forwarding Database, Service 64000
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
64000     00:00:27:00:00:5e vxlan:                   Evpn     07/22/15 13:38:47
                            39.0.0.94:119281
64000     1e:e2:ff:00:f9:3d cpm                      Intf     07/22/15 13:38:44
-------------------------------------------------------------------------------
```

The first MAC entry address is the tunnel NH-MAC for the VSG and the second is the address for the DC Gateway for EVPN-tunnel service 64000:

```
*A:pe-9# show router 64001 route-table
===============================================================================
1.1.1.1/32                                    Local   Local     00h09m08s  0
       lo1                                                          0
10.32.78.0/24                                 Remote  BGP EVPN  00h09m05s  169
       vpls-64000 (ET-00:00:27:00:00:5e)                            0
10.32.78.100/32                               Remote  BGP EVPN  00h00m04s  169
       vpls-64000 (ET-00:00:27:00:00:5e)                            0
-------------------------------------------------------------------------------
```

The following commands are useful to obtain relevant information:

- *show service vsd domain "L3-service-1" association* to obtain svc-IDs
- *show service id <id> all* to obtain RT/RD/VNI values
- *show service id <id> vxlan* to obtain VTEPs in the VPLS service
- *show service vxlan <vtep-ip>* to obtain svc-id and VNI information

Relevant EVPN tunnel NH-MAC/VNI/RT/NH information is also in the EVPN BGP RIB:

```
*A:pe-9# show router bgp routes evpn mac detail
--- snipped ---

Modified Attributes

Network        : N/A
Nexthop        : 39.0.0.94
From           : 39.0.0.94
Res. Nexthop   : 192.168.39.94
Local Pref.    : 200                    Interface Name : toNuage
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : target:65534:38619 bgp-tunnel-encap:VXLAN
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 39.0.0.94
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
EVPN type      : MAC
ESI            : ESI-0
Tag            : 119281
IP Address     : N/A
Route Dist.    : 65534:29625
Mac Address    : 00:00:27:00:00:5e
MPLS Label1    : VNI 119281             MPLS Label2    : N/A
Route Tag      : 0
--- snipped ---
```

VM/VSG-connected Host and network information is also in the EVPN BGP RIB:

```
*A:pe-9# show router bgp routes evpn ip-prefix detail
--- snipped ---
Modified Attributes

Network        : N/A
Nexthop        : 39.0.0.94
From           : 39.0.0.94
Res. Nexthop   : 192.168.39.94
Local Pref.    : 200                    Interface Name : toNuage
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : target:65534:38619 ext:30b:220000000000
                 ext:30b:100b00b00000 bgp-tunnel-encap:VXLAN
                 mac-nh:00:00:27:00:00:5e
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 39.0.0.94
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
EVPN type      : IP-PREFIX
ESI            : N/A
Tag            : 119281
Gateway Address: 00:00:27:00:00:5e
Prefix         : 10.32.78.100/32
Route Dist.    : 39.0.0.94:10269
MPLS Label     : VNI 119281
Route Tag      : 0
--- snipped ---

Modified Attributes

Network        : N/A
Nexthop        : 39.0.0.94
From           : 39.0.0.94
Res. Nexthop   : 192.168.39.94
Local Pref.    : 200                    Interface Name : toNuage
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : target:65534:38619 ext:30b:220000000000
                 ext:30b:100b00b00000 bgp-tunnel-encap:VXLAN
                 mac-nh:00:00:27:00:00:5e
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 39.0.0.94
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
EVPN type      : IP-PREFIX
ESI            : N/A
Tag            : 119281
Gateway Address: 00:00:27:00:00:5e
Prefix         : 10.32.78.0/24
Route Dist.    : 39.0.0.94:10269
MPLS Label     : VNI 119281
```

```
Route Tag      : 0
--- snipped ---
```

The VM or VSG-connected Host should be able to ping the loopback interface of the VPRN service:

```
*A:ce1# ping 1.1.1.1 source 10.32.78.100
PING 1.1.1.1 56 data bytes
64 bytes from 1.1.1.1: icmp_seq=1 ttl=64 time=1.67ms.
64 bytes from 1.1.1.1: icmp_seq=2 ttl=64 time=1.83ms.
```

Troubleshooting and debug commands

When testing/troubleshooting F-D XMPP provisioning, the following show/tools/debug commands can be useful:

- tools perform service vsd evaluate-script
- tools perform service vsd fd-domain-sync <full|diff>
- tools perform service vsd domain refresh-config
- tools perform python-script reload
- tools dump service vsd-services command-list

- debug python python-script
- debug vsd scripts event/instance
- debug system xmpp
- debug router bgp update

- show service vsd domain
- show service vsd script
- show service vsd summary
- show system vsd
- show xmpp vsd
- show python python-policy <name> {association}
- show python python-script <name> {association|source-in-use}

- show service vxlan [<vtep-ip>]

- show service service-using {<service-type>}
- show service id route-table
- show service id fdb detail
- show service id proxy-arp detail
- show router [<router-instance>] route-table
- show router [<router-instance>] arp
- show router bgp routes bgp <mac|ip-prefix|inclusive-mcast> {detail}

- log-id 99

Conclusion

The fully dynamic VSD integration model allows for automated provisioning of breakout services on the 7x50 DC Gateway. Different domain types (l2-domain/l2-domain-irb/vrf-vxlan/vrf-gre) are supported. This chapter has shown how to construct, load, and test python scripts for this feature. It also described how to configure WAN services on the VSD and how to verify the dynamically created services.

Inter-AS Model C for VLL

This chapter describes advanced inter-AS model C for VLL configurations.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SR OS Release 8.0.R4. The CLI in the current edition corresponds to SR OS Release 15.0.R2.

Overview

SR OS supports RFC 3107, *Carrying Label Information in BGP-4*, including VLL/VPLS. BGP SDPs can also be used with PBB-VPLS services.

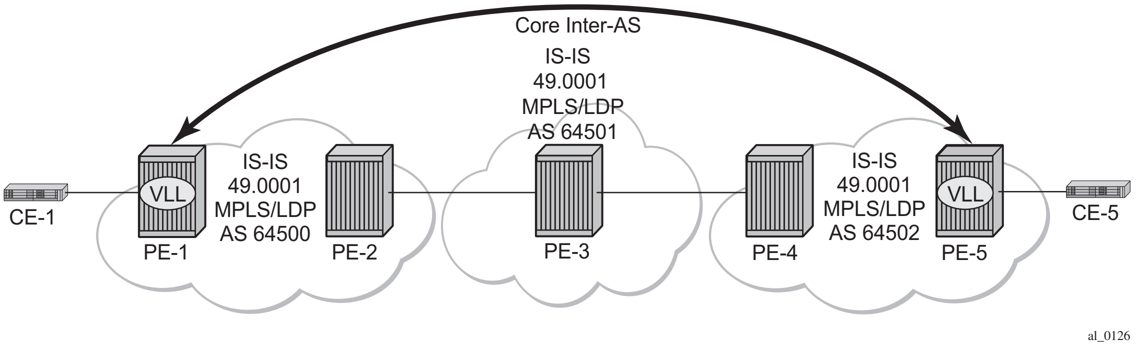
ISPs are looking for mechanisms to implement the VLL and VPLS services across autonomous systems (ASs). Service providers may have inter-AS operation as a consequence of delivering inter-provider VLL/VPLS or because they use multiple ASs as a result of acquisitions and mergers.

The objective of this chapter is to describe the interconnection of VLL services across multiple ASs, using inter-AS model C. Inter-AS Model C involves eBGP redistribution of internal system addresses to the neighboring AS using labeled IPv4 routes.

Example Topology

Figure 81 shows the example topology used for Inter-AS Model C.

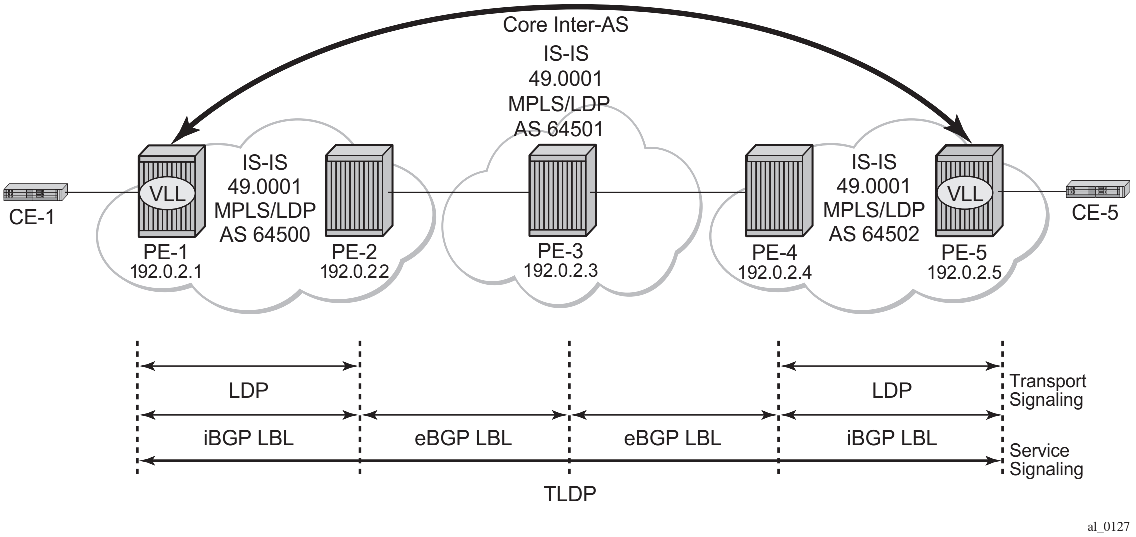
Figure 81 **Example Topology- Inter-AS Model C for VLL**



The example topology shown in Figure 81 consists of three sites in different ASs with each site using 7750 SRs.

AS 64500 contains PE-1 and PE-2, AS 64501 contains PE-3, and AS 64502 contains PE-4 and PE-5. There is a business customer with two remote locations, Site A and Site B, with Customer Edge (CE) devices CE-1 connected to the AS 64500 via PE-1 and CE-5 connected to the AS 64502 via PE-5. A VLL Epipe service is configured between PE-1 and PE-5 to connect site A and site B.

Figure 82 **Inter-AS Model C for VLL**

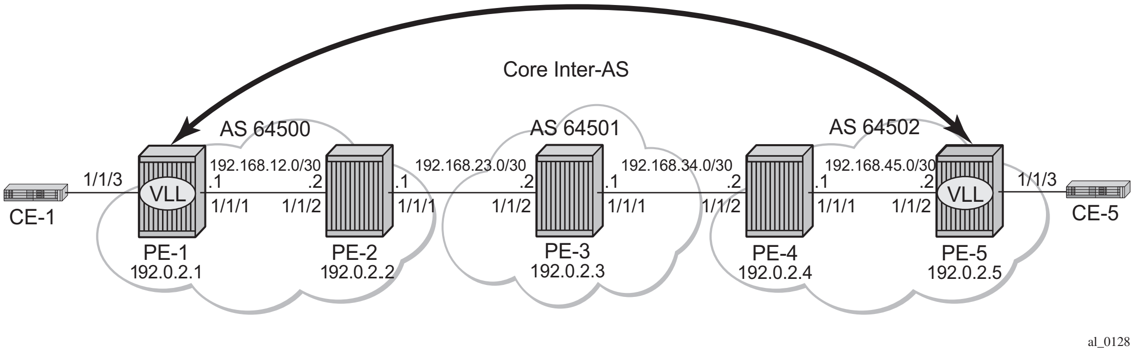


Configuration

This section describes all of the relevant configuration tasks for the detailed setup shown in Figure 83. In this particular example, the following protocols are assumed to be already configured.

- IS-IS as the IGP with all the nodes being level Level1/Level 2.
- LDP as the MPLS protocol to signal the transport tunnels within AS 64500 and AS 64502.

Figure 83 **Network Setup Configuration**



BGP Configuration

A BGP tunnel must be established between PE-1 and PE-5, therefore, labeled BGP routes must be exchanged for prefixes 192.0.2.1/32 and 192.0.2.5/32 across the ASs. The following shows the BGP configuration — iBGP and eBGP — required for the PE routers to implement an Inter-AS VLL.

The BGP configuration on PE-3 in AS 64501 is as follows:

```
*A:PE-3# configure
    router
        autonomous-system 64501
        bgp
            min-route-advertisement 1
            rapid-withdrawal
            split-horizon
            group "EBGP"
                local-as 64501
                neighbor 192.168.23.1
                    family label-ipv4
                    peer-as 64500
                exit
                neighbor 192.168.34.2
                    family label-ipv4
                    peer-as 64502
                exit
            exit
        exit
```

The address family **label-ipv4** must be configured so that MPLS labels are carried along with MP-BGP Network Layer Reachability Information (NLRIs), see chapter *Separate BGP RIBs for Labeled Routes*. The setting **split-horizon** is optional and prevents that a received route is sent back to the originator, which might result in multiple routes for a certain prefix.

To export the prefixes of the nodes where the Epipe is configured (PE-1 and PE-5) to another AS, a common scenario is to advertise the prefix to be exported within the AS as labeled BGP. Therefore, an export policy is defined for prefix192.0.2.1/32 on PE-1 and this prefix will be advertised to the ASBR in AS 64500, in this case to PE-2. On PE-2, the labeled BGP route for prefix 192.0.2.1/32 is inactive, because the IGP route for that prefix is preferred. No export policy needs to be configured in the Autonomous System Border Router (ASBR) PE-2 for the EBGP session with PE-3 in AS 64501. Rather, the setting **advertise-inactive** will allow the inactive labeled BGP routes from AS 64500 to be advertised to PE-3 in AS 64501. Likewise, an export policy will be configured on PE-5 to advertise prefix 192.0.2.5/32 to ASBR PE-4 in AS 64502. On PE-4, BGP is configured with **advertise-inactive** to advertise the labeled BGP route to its EBGP peer, PE-3. The advantage of this approach is that labeled BGP is used end-to-end between PE-1 and PE-5 and no IGP routes are to be redistributed into BGP, which would be the case if no local BGP labeled routes were advertised within AS 64500 or AS 64502 and only IGP routes were defined within these ASs. The ASBRs PE-2, PE-3, and PE-4 will swap the BGP labels. PE-3 will advertise the labeled BGP routes learned from AS 64500 to AS 64502 and vice versa and the ASBRs will advertise these labeled routes for remote PE prefixes to their BGP peers. Eventually, PE-1 will have learned a labeled BGP route for prefix 192.0.2.5/32 and PE-5 will have learned a labeled BGP route for prefix 192.0.2.1/32 and a VLL Epipe can be established between PE-1 and PE-5.

The BGP configuration of ASBR PE-2 in AS 64500 is as follows:

```
*A:PE-2# configure
    router
        autonomous-system 64500
        bgp
            min-route-advertisement 1
            rapid-withdrawal
            split-horizon
            group "EBGP"
                local-as 64500
                neighbor 192.168.23.2
                    family label-ipv4
                    peer-as 64501
                    advertise-inactive
                exit
            exit
            group "IBGP"
                neighbor 192.0.2.1
                    family label-ipv4
                    next-hop-self
                    peer-as 64500
                    advertise-inactive
                exit
            exit
        exit
```

The BGP configuration of ASBR PE-4 in AS 64502 is as follows:

```
*A:PE-4# configure
    router
         autonomous-system 64502
         bgp
            min-route-advertisement 1
            rapid-withdrawal
            split-horizon
            group "EBGP"
                local-as 64502
                neighbor 192.168.34.1
                    family label-ipv4
                    peer-as 64501
                    advertise-inactive
                exit
            exit
            group "IBGP"
                neighbor 192.0.2.5
                    family label-ipv4
                    next-hop-self
                    peer-as 64502
                    advertise-inactive
                exit
            exit
        exit
```

PE-1 and PE-5 are the PEs to which the CEs are connected in AS 64500 and AS 64502. PE-1 and PE-5 advertise their system prefixes as labeled BGP routes to their BGP peers within the AS.

The BGP configuration of PE-1 is as follows:

```
*A:PE-1# configure
    router
         autonomous-system 64500
         bgp
            min-route-advertisement 1
            rapid-withdrawal
            split-horizon
            group "IBGP"
                export "export-PE-1"
                neighbor 192.0.2.2
```

```
                    family label-ipv4
                    next-hop-self
                    peer-as 64500
                exit
            exit
        exit
```

The BGP configuration of PE-5 in AS 64502 is as follows:

```
*A:PE-5# configure
    router
         autonomous-system 64502
         bgp
            min-route-advertisement 1
            rapid-withdrawal
            split-horizon
            group "IBGP"
                export "export-PEsys"
                neighbor 192.0.2.4
                    family label-ipv4
                    next-hop-self
                    peer-as 64502
                exit
            exit
        exit
```

Policy Configuration

The export policies on PE-1 and PE-5 advertise the system addresses to the remote AS.

The export policy on PE-1 has a prefix list that only contains prefix 192.0.2.1/32 as follows:

```
*A:PE-1# configure
    router
        policy-options
            begin
            prefix-list "PE-1"
                prefix 192.0.2.1/32 exact
            exit
            policy-statement "export-PE-1"
                entry 10
                    from
                        prefix-list "PE-1"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit
```

A similar export policy can be configured for prefix 192.0.2.5/32 on PE-5. However, the export policy on PE-5 is slightly different: the policy has a prefix list that can be applied for prefixes on multiple PEs, but in this case, only prefix 192.0.2.5/32 will be exported:

```
*A:PE-5# configure
    router
        policy-options
            begin
            prefix-list "PEsys"
                prefix 192.0.2.0/29 longer
            exit
            policy-statement "export-PEsys"
                entry 10
                    from
                        protocol direct
                        prefix-list "PEsys"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit
```

The same policy could have been applied on PE-1.

Service Configuration

Once BGP is configured, the configuration requires the service to be defined (Epipe 1). The focus here is a VLL service, however, it is also possible to have a similar configuration with VPLS services.

The following shows the service level configuration on PE-1:

```
*A:PE-1# configure
     service
        sdp 15 mpls create
            far-end 192.0.2.5
            bgp-tunnel
            no shutdown
        exit
        epipe 1 customer 1 create
            description "Tunnel-PE-1-PE-5"
            sap 1/1/3:1 create
            exit
            spoke-sdp 15:1 create
            exit
            no shutdown
        exit
```

The following CLI shows the service level configuration on PE-5:

```
*A:PE-5# configure
    service
        sdp 51 mpls create
            far-end 192.0.2.1
            bgp-tunnel
            no shutdown
        exit
        epipe 1 customer 1 create
            description "Tunnel-PE-5-PE-1"
            sap 1/1/3:1 create
            exit
            spoke-sdp 51:1 create
            exit
            no shutdown
        exit
```

Show Commands and Troubleshooting

On PE-5, BGP tunnels exist to the remote AS system addresses that are using LDP as a transport mechanism and the configuration of end-to-end SDPs over which TLDP service labels are exchanged.

The following shows information about SDP 15 on PE-1:

```
*A:PE-1# show service sdp

============================================================================
Services: Service Destination Points
============================================================================
SdpId  AdmMTU  OprMTU  Far End          Adm  Opr         Del     LSP   Sig
----------------------------------------------------------------------------
15     0       1552    192.0.2.5        Up   Up          MPLS    B     TLDP
----------------------------------------------------------------------------
Number of SDPs : 1
----------------------------------------------------------------------------
Legend: R = RSVP, L = LDP, B = BGP, M = MPLS-TP, n/a = Not Applicable
        I = SR-ISIS, O = SR-OSPF, T = SR-TE, F = FPE
============================================================================
*A:PE-1#
```

The VLL Epipe service is up, as follows:

```
*A:PE-1# show service service-using

===============================================================================
Services
===============================================================================
ServiceId    Type      Adm  Opr  CustomerId Service Name
-------------------------------------------------------------------------------
1            Epipe     Up   Up   1
2147483648   IES       Up   Down 1          _tmnx_InternalIesService
```

```
2147483649   intVpls   Up   Down 1          _tmnx_InternalVplsService
-------------------------------------------------------------------------------
Matching Services : 3
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

Two LDP sessions have been established from PE-1: a link LDP session with neighbor PE-2 in AS 64500 and a targeted LDP session with PE-5 in AS 64502, as follows:

```
*A:PE-1# show router ldp session ipv4

==============================================================================
LDP IPv4 Sessions
==============================================================================
Peer LDP Id         Adj Type  State         Msg Sent  Msg Recv  Up Time
------------------------------------------------------------------------------
192.0.2.2:0         Link      Established   95        98        0d 00:04:02
192.0.2.5:0         Targeted  Established   9         10        0d 00:00:27
------------------------------------------------------------------------------
No. of IPv4 Sessions: 2
=============================================================================
*A:PE-1#
```

The routing table on PE-1 shows that the system IP address of PE-5 is reachable using a BGP tunnel:

```
*A:PE-1# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Local   Local     00h51m09s  0
       system                                                      0
192.0.2.2/32                                  Remote  ISIS      00h51m02s  15
       192.168.12.2                                                10
192.0.2.5/32                                  Remote  BGP_LABEL 00h49m07s  170
       192.0.2.2 (tunneled)                                        10
192.168.12.0/30                               Local   Local     00h51m09s  0
       int-PE-1-PE-2                                               0
-------------------------------------------------------------------------------
No. of Routes: 4
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

The following tunnel-table on PE-1 shows the details of the LDP, SDP, and BGP tunnels.

```
*A:PE-1# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.2/32      ldp       MPLS  65537     9        192.168.12.2   10
192.0.2.5/32      sdp       MPLS  15        5        192.0.2.5      0
192.0.2.5/32      bgp       MPLS  262145    12       192.0.2.2      1000
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-1#
```

The service details for Epipe 1 on PE-1 are as follows:

```
*A:PE-1# show service id 1 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 1                   Vpn Id            : 0
Service Type      : Epipe
Name              : (Not Specified)
Description       : Tunnel-PE-1-PE-5
Customer Id       : 1                   Creation Origin   : manual
Last Status Change: 04/26/2017 07:53:04
Last Mgmt Change  : 04/26/2017 06:48:04
Test Service      : No
Admin State       : Up                  Oper State        : Up
MTU               : 1514
Vc Switching      : False
SAP Count         : 1                   SDP Bind Count    : 1
Per Svc Hashing   : Disabled
Vxlan Src Tep Ip  : N/A
Force QTag Fwd    : Disabled

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/3:1                              q-tag        1518    1518    Up   Up
sdp:15:1 S(192.0.2.5)                    Spok         0       1552    Up   Up
===============================================================================
*A:PE-1#
```

ICMP is used to verify the IP connectivity from PE-1 to the system IP address of PE-5:

```
*A:PE-1# ping 192.0.2.5
PING 192.0.2.5 56 data bytes
64 bytes from 192.0.2.5: icmp_seq=1 ttl=64 time=1.91ms.
64 bytes from 192.0.2.5: icmp_seq=2 ttl=64 time=2.06ms.
64 bytes from 192.0.2.5: icmp_seq=3 ttl=64 time=2.02ms.
```

```
64 bytes from 192.0.2.5: icmp_seq=4 ttl=64 time=2.01ms.
64 bytes from 192.0.2.5: icmp_seq=5 ttl=64 time=2.02ms.

---- 192.0.2.5 PING Statistics ----
5 packets transmitted, 5 packets received, 0.00% packet loss
round-trip min = 1.91ms, avg = 2.01ms, max = 2.06ms, stddev = 0.050ms
*A:PE-1#
```

The same commands on PE-5 result in the following output:

```
*A:PE-5# show service sdp

============================================================================
Services: Service Destination Points
============================================================================
SdpId  AdmMTU  OprMTU  Far End          Adm  Opr         Del     LSP   Sig
----------------------------------------------------------------------------
51     0       1552    192.0.2.1        Up   Up          MPLS    B     TLDP
----------------------------------------------------------------------------
Number of SDPs : 1
----------------------------------------------------------------------------
Legend: R = RSVP, L = LDP, B = BGP, M = MPLS-TP, n/a = Not Applicable
        I = SR-ISIS, O = SR-OSPF, T = SR-TE, F = FPE
============================================================================
*A:PE-5#
```

```
*A:PE-5# show service service-using

===============================================================================
Services
===============================================================================
ServiceId    Type      Adm  Opr  CustomerId Service Name
-------------------------------------------------------------------------------
1            Epipe     Up   Up   1
2147483648   IES       Up   Down 1          _tmnx_InternalIesService
2147483649   intVpls   Up   Down 1          _tmnx_InternalVplsService
-------------------------------------------------------------------------------
Matching Services : 3
-------------------------------------------------------------------------------
===============================================================================
*A:PE-5#
```

```
*A:PE-5# show router ldp session ipv4

==============================================================================
LDP IPv4 Sessions
==============================================================================
Peer LDP Id         Adj Type  State         Msg Sent  Msg Recv  Up Time
------------------------------------------------------------------------------
192.0.2.1:0         Targeted  Established   12        13        0d 00:00:44
192.0.2.4:0         Link      Established   98        100       0d 00:04:11
------------------------------------------------------------------------------
No. of IPv4 Sessions: 2
==============================================================================
*A:PE-5#
```

```
*A:PE-5# show router route-table
```

```
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  BGP_LABEL 00h01m16s  170
       192.0.2.4 (tunneled)                                         0
192.0.2.4/32                                  Remote  ISIS      00h05m32s  15
       192.168.45.1                                                 10
192.0.2.5/32                                  Local   Local     00h05m33s  0
       system                                                       0
192.168.45.0/30                               Local   Local     00h06m33s  0
       int-PE-5-PE-4                                                0
-------------------------------------------------------------------------------
No. of Routes: 4
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-5#

*A:PE-5# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.1/32      sdp       MPLS  51        5        192.0.2.1      0
192.0.2.1/32      bgp       MPLS  262145    12       192.0.2.4      1000
192.0.2.4/32      ldp       MPLS  65537     9        192.168.45.1   10
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-5#

*A:PE-5# show service id 1 base
===============================================================================
Service Basic Information
===============================================================================
Service Id        : 1                   Vpn Id            : 0
Service Type      : Epipe
Name              : (Not Specified)
Description       : Tunnel-PE-5-PE-1
Customer Id       : 1                   Creation Origin   : manual
Last Status Change: 04/26/2017 07:52:31
Last Mgmt Change  : 04/26/2017 06:47:41
Test Service      : No
Admin State       : Up                  Oper State        : Up
MTU               : 1514
Vc Switching      : False
SAP Count         : 1                   SDP Bind Count    : 1
Per Svc Hashing   : Disabled
Vxlan Src Tep Ip  : N/A
```

```
Force QTag Fwd    : Disabled

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/3:1                              q-tag        1518    1518    Up   Up
sdp:51:1 S(192.0.2.1)                    Spok         0       1552    Up   Up
===============================================================================
*A:PE-5#


*A:PE-5# ping 192.0.2.1
PING 192.0.2.1 56 data bytes
64 bytes from 192.0.2.1: icmp_seq=1 ttl=64 time=1.83ms.
64 bytes from 192.0.2.1: icmp_seq=2 ttl=64 time=2.06ms.
64 bytes from 192.0.2.1: icmp_seq=3 ttl=64 time=2.01ms.
64 bytes from 192.0.2.1: icmp_seq=4 ttl=64 time=2.08ms.
64 bytes from 192.0.2.1: icmp_seq=5 ttl=64 time=2.15ms.

---- 192.0.2.1 PING Statistics ----
5 packets transmitted, 5 packets received, 0.00% packet loss
round-trip min = 1.83ms, avg = 2.03ms, max = 2.15ms, stddev = 0.107ms
*A:PE-5#
```

On PE-5, the BGP route to the system IP address of PE-1 can been seen with PE-4 as the next hop:

```
*A:PE-5# show router bgp routes label-ipv4
===============================================================================
 BGP Router ID:192.0.2.5        AS:64502       Local AS:64502
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  192.0.2.1/32                                       100         None
      192.0.2.4                                          None        262140
      64501 64500
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-5#
```

On PE-5, the FIB on slot 1 shows that the system IP address of PE-1 is reachable using BGP over an LDP transport to PE-4:

```
*A:PE-5# show router fib 1
```

```
===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
    NextHop
-------------------------------------------------------------------------------
192.0.2.1/32                                                BGP_LABEL
    192.0.2.4 (Transport:LDP)
192.0.2.4/32                                                ISIS
  192.168.45.1 (int-PE-5-PE-4)
192.0.2.5/32                                                LOCAL
  192.0.2.5 (system)
192.168.45.0/30                                             LOCAL
  192.168.45.0 (int-PE-5-PE-4)
-------------------------------------------------------------------------------
Total Entries : 4
-------------------------------------------------------------------------------
===============================================================================
*A:PE-5#
```

The show commands on router PE-3 in AS 64501 are as follows:

```
*A:PE-3# show router bgp summary all

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.168.23.1
Def. Instance  64500       36    0 00h16m20s 1/1/1 (Lbl-IPv4)
                           36    0
192.168.34.2
Def. Instance  64502       34    0 00h15m16s 1/1/1 (Lbl-IPv4)
                           34    0

-------------------------------------------------------------------------------
*A:PE-3#


*A:PE-3# show router bgp routes label-ipv4
===============================================================================
 BGP Router ID:192.0.2.3        AS:64501       Local AS:64501
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                          LocalPref   MED
      Nexthop (Router)                                 Path-Id     Label
      As-Path
--------------------------------------------------------------------------------
u*>i  192.0.2.1/32                                        None        None
      192.168.23.1                                        None        262141
      64500
u*>i  192.0.2.5/32                                        None        None
      192.168.34.2                                        None        262140
      64502
--------------------------------------------------------------------------------
Routes : 8
================================================================================
*A:PE-3#
```

The BGP labels are swapped at PE-3, as follows:

```
*A:PE-3# show router bgp inter-as-label

================================================================================
BGP Inter-AS labels
Flags: B - entry has backup, P - entry is promoted
================================================================================
NextHop                       Received       Advertised     Label
                              Label          Label          Origin
--------------------------------------------------------------------------------
192.168.23.1                  262141         262143         External
192.168.34.2                  262140         262142         External
--------------------------------------------------------------------------------
Total Labels allocated:   2
================================================================================
*A:PE-3#
```

The routing table on PE-3 includes BGP labeled routes to PE-1 and PE-5, as follows:

```
*A:PE-3# show router route-table

================================================================================
Route Table (Router: Base)
================================================================================
Dest Prefix[Flags]                             Type    Proto     Age        Pref
      Next Hop[Interface Name]                                     Metric
--------------------------------------------------------------------------------
192.0.2.1/32                                   Remote  BGP_LABEL 00h50m04s  170
       192.168.23.1                                                  0
192.0.2.3/32                                   Local   Local     00h51m40s  0
       system                                                        0
192.0.2.5/32                                   Remote  BGP_LABEL 00h49m50s  170
       192.168.34.2                                                  0
192.168.23.0/30                                Local   Local     00h51m40s  0
       int-PE-3-PE-2                                                 0
192.168.34.0/30                                Local   Local     00h51m40s  0
       int-PE-3-PE-4                                                 0
--------------------------------------------------------------------------------
No. of Routes: 5
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
```

```
===============================================================================
*A:PE-3#
```

The commands on PE-2 are as follows:

```
*A:PE-2# show router bgp summary all

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.1
Def. Instance  64500       36    0 00h16m24s 1/0/1 (Lbl-IPv4)
                           36    0
192.168.23.2
Def. Instance  64501       36    0 00h16m19s 1/1/1 (Lbl-IPv4)
                           36    0

-------------------------------------------------------------------------------
*A:PE-2#
```

The BGP labels are swapped by PE-2 as follows:

```
*A:PE-2# show router bgp inter-as-label

===============================================================================
BGP Inter-AS labels
Flags: B - entry has backup, P - entry is promoted
===============================================================================
NextHop                       Received       Advertised     Label
                              Label          Label          Origin
-------------------------------------------------------------------------------
192.0.2.1                     262141         262141         Internal
192.168.23.2                  262142         262140         External
-------------------------------------------------------------------------------
Total Labels allocated:   2
===============================================================================
*A:PE-2#


*A:PE-2# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  ISIS      00h51m43s  15
       192.168.12.1                                                 10
192.0.2.2/32                                  Local   Local     00h51m44s  0
       system                                                       0
```

```
192.0.2.5/32                                  Remote   BGP_LABEL 00h49m48s  170
       192.168.23.2                                                  0
192.168.12.0/30                               Local    Local     00h51m44s  0
       int-PE-2-PE-1                                                 0
192.168.23.0/30                               Local    Local     00h51m44s  0
       int-PE-2-PE-3                                                 0
-------------------------------------------------------------------------------
No. of Routes: 5
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-2#
```

The show commands on PE-4 are the following:

```
*A:PE-4# show router bgp summary all

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.5
Def. Instance  64502       34    0 00h15m10s 1/0/1 (Lbl-IPv4)
                           34    0
192.168.34.1
Def. Instance  64501       34    0 00h15m17s 1/1/1 (Lbl-IPv4)
                           34    0

-------------------------------------------------------------------------------
*A:PE-4#
```

```
*A:PE-4# show router bgp routes label-ipv4
===============================================================================
 BGP Router ID:192.0.2.4        AS:64502       Local AS:64502
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  192.0.2.1/32                                       None        None
      192.168.34.1                                       None        262143
      64501 64500
```

```
*i    192.0.2.5/32                                       100         None
      192.0.2.5                                          None        262141
      No As-Path
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-4#

*A:PE-4# show router bgp inter-as-label
===============================================================================
BGP Inter-AS labels
Flags: B - entry has backup, P - entry is promoted
===============================================================================
NextHop                       Received       Advertised     Label
                              Label          Label          Origin
-------------------------------------------------------------------------------
192.168.34.1                  262143         262141         External
192.0.2.5                     262141         262140         Internal
------------------------------------------------------------------------------
Total Labels allocated:   2
===============================================================================
*A:PE-4#
```

Conclusion

The BGP tunnel based SDP binding is allowed for VLL and VPLS services, including PBB-VPLS. Using RFC 3107, it is possible to implement inter-AS Model C VLLs.

The example used in this chapter illustrates the configuration of an Inter-AS VLL providing access to CE sites. Troubleshooting commands also have been shown to verify all the procedures.

Layer 2 Multicast Optimization for EVPN-VXLAN - Assisted Replication

This chapter provides information about Layer 2 Multicast Optimization for EVPN-VXLAN - Assisted Replication.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

The information in this chapter is based on SR OS Release 14.0.R4 and the CLI is based on SR OS Release 15.0.R3. Layer 2 multicast optimization for EVPN-VXLAN - Assisted Replication (AR) is supported in SR OS Release 14.0.R4, and later.

Overview

Typically, EVPN-VXLAN can use either Ingress Replication (IR) or Protocol Independent Multicast (PIM) for Broadcast, Unknown unicast, and Multicast (BUM) traffic (although SR OS does not support PIM along with EVPN-VXLAN). PIM requires keeping multicast state awareness per subnet per tenant in the core routers, which may not scale. Not all core routers support PIM.

IR inefficiency is usually tolerable in EVPN networks for broadcast and unknown unicast traffic; however, it is not tolerable for multicast traffic:

- Broadcast traffic can be reduced by the proxy-ARP and proxy-ND capabilities supported by EVPN.

- Unknown unicast traffic is greatly reduced in virtualized Data Center (DC) networks where all MAC and IP addresses are learned in the control or management planes. In such cases, unknown MAC addresses are always outside the DC. An **unknown-mac-route** can be enabled to ensure that the unknown unicast traffic is sent only to the DC gateway, which minimizes flooding within the DC.
- Multicast traffic may be an issue for the hypervisors holding the multicast sources, because the hypervisors need to replicate the multicast traffic to the remote VXLAN Tunnel Endpoints (VTEPs). The multicast replication at the hypervisors is a software process and the throughput can be heavily impacted. This is also true when VPLS services are used in the Virtual Service Router (VSR) and many replicas must be done from the VSR. Using a dedicated service node to replicate the multicast traffic on behalf of the hypervisors can help, but the replication capabilities of such service nodes are limited too.

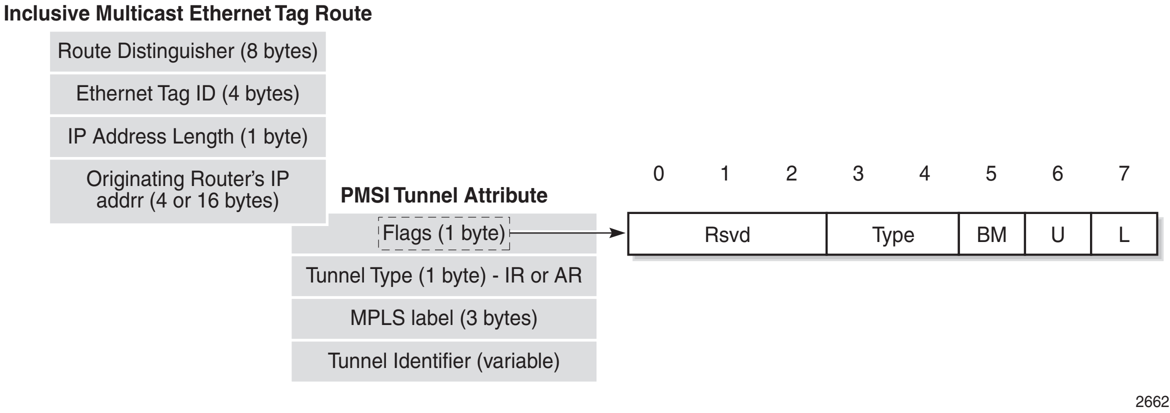
SR OS supports the Assisted Replication (AR) feature for IPv4 VXLAN tunnels (both replicator and leaf functions) in compliance with the non-selective mode described in *draft-ietf-bess-evpn-optimized-ir*. AR is a Layer 2 multicast optimization feature that helps software-based PEs and Network Virtualization Edge (NVE) devices with low-performance replication capabilities to deliver Broadcast and Multicast (BM) Layer 2 traffic to remote VTEPs in the VPLS.

SR OS nodes support the AR-Replicator (AR-R) and AR-Leaf (AR-L) functions, although not simultaneously on the same service. Nodes configured as AR-L select an AR-R within a service and send all BM packets to this AR-R. AR-Rs replicate traffic to all the VTEPs in the VPLS on behalf of the AR-Ls, so BM traffic is delivered to all VPLS participants without any packet loss caused by performance issues. Unknown unicast packets follow the same path as known unicast packets to avoid packet reordering. Therefore, no AR-R is used for unknown unicast traffic.

When multiple AR-Rs exist in a service, the AR-L performs per-service load-balancing of the BM traffic. The AR-L lists the candidate AR-Rs, ordered by IP and VXLAN Network Identifier (VNI); candidate 0 having the lowest IP and VNI. The replicator is selected using a modulo function of the service ID and the number of candidate AR-Rs. For example, assume that VPLS 1 has two candidate AR-Rs: because 1 modulo 2 equals 1, the second AR-R in the list will be selected. In case of failure, a new AR-R is selected. If there are no more AR-Rs, the system falls back to IR.

Figure 84 shows an EVPN route-type 3, an Inclusive Multicast Ethernet Tag (IMET) route containing a PMSI tunnel attribute with a flags octet. Flag L was already defined in RFC6514. *Draft-ietf-bess-evpn-optimized-ir* defines additional flags: type, BM, and U. The BM and U flags are used for Pruned Flood Lists (PFL) signaling and they are not supported.

Figure 84 **PMSI Tunnel Attribute - Flags**


Inclusive Multicast Ethernet Tag Route
Route Distinguisher (8 bytes)
Ethernet Tag ID (4 bytes)
IP Address Length (1 byte)
Originating Router's IP addrr (4 or 16 bytes)
PMSI Tunnel Attribute
Flags (1 byte)
Tunnel Type (1 byte) - IR or AR
MPLS label (3 bytes)
Tunnel Identifier (variable)
0
1
2
3
4
5
6
7
Rsvd
Type
BM
U
L
26626

The type field has two bits that define the AR role of the advertising router, as follows:

- Type 00 = Regular Network Virtualization Edge (RNVE) - indicates that AR is not supported and IR is applied instead (for backward compatibility)
- Type 01= AR-R
- Type 10 = AR-L
- Type 11 = reserved

The tunnel type in the PMSI tunnel attribute can be configured with the following options for IR and AR:

- Tunnel type 0x06 = (non-optimized) IR, sent by AR-R and AR-L if **ingress-repl-inc-mcast-advertisement** is enabled, which is the default option
- Tunnel type 0x0A = type AR, originated by AR-R

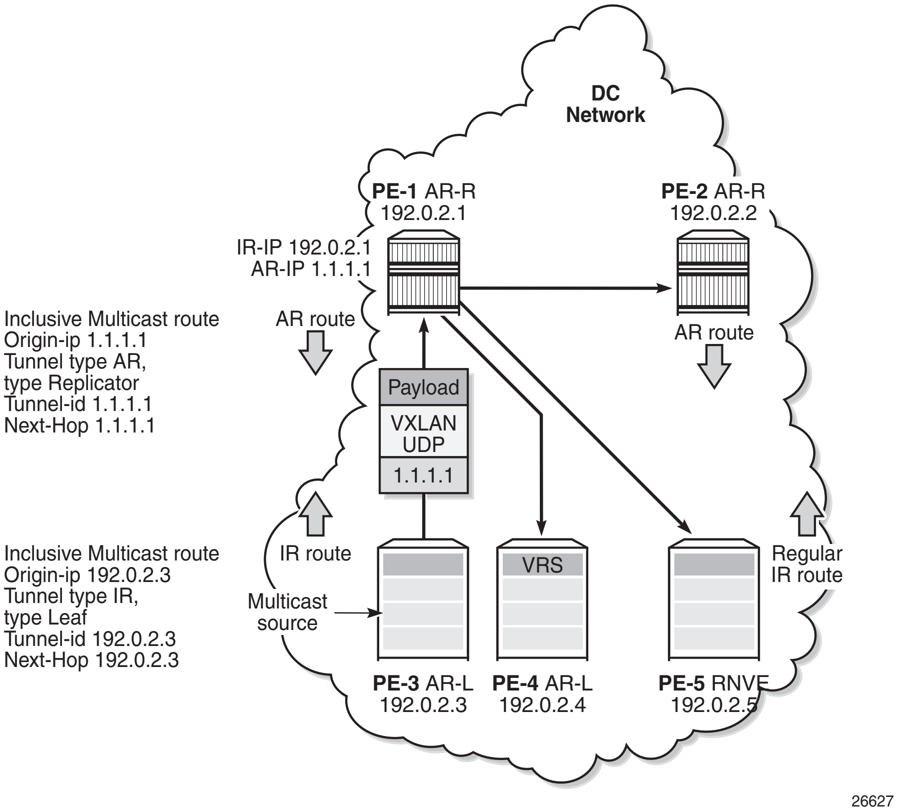
For regular IR routes, the originating router's IP address equals the system IP address. The MPLS label and tunnel identifier must be used as described in RFC7432. The tunnel identifier is set to a routable address of the PE.

For AR routes, the originating router's IP address and the tunnel identifier are both set to the AR IP address (AR-IP) configured in the **service system bgp-evpn** context. The AR-IP must be previously defined as a loopback interface address in the base router and must be different from the IR IP address (IR-IP).

Note: If the AR-IP loopback interface is down, the router will not withdraw the AR route. However, the remote AR-Ls will not be able to resolve the AR route's BGP next-hop if the AR-IP is no longer propagated in the IGP.

Figure 85 shows the example topology with the multicast source connected to a hypervisor PE-3 that acts as AR-L, which will send an IR route containing the system address of PE-3. The AR-R PE-1 sends an AR route that uses AR-IPs instead of IR-IPs; for example, PE-1 has AR-IP 1.1.1.1 and IR-IP 192.0.2.1.

Figure 85 **EVPN Assisted Replication for VXLAN**



Hypervisor PE-3 will send the BM traffic to the AR-R, which will replicate it to all the VTEPs in the VPLS, except to PE-3.

Table 7 shows the inclusive multicast route information sent by each role in an AR-capable service.

Table 7 **Inclusive Multicast Route Information Sent By Different AR Roles**

AR Role	Function	Inclusive Multicast Route Advertised
AR-R	Assists AR-Ls	IR inclusive multicast route (tunnel = 0x06 = IR, IR-IP, type = 0 = none) AR inclusive multicast route (tunnel = 0x0A = AR, AR-IP, type = 1 = AR-R)
AR-L	Sends BM only to AR-R	IR inclusive multicast route (tunnel = 0x06 = IR, IR- IP, type = 2 = AR-L)
RNVE	Non-AR support	IR inclusive multicast route (tunnel = 0x06 = IR, IR- IP, type = 0 = none)

Unicast traffic (known or unknown) is processed as normal. For BM traffic, the AR-R uses AR or IR based on the IP destination address (DA):

- If IP DA equals the AR-IP, the AR-R replicates to the VTEPs in the VXLAN service, except for the VTEP over which the BM traffic was received.
- If IP DA equals the IR-IP, normal IR forwarding is done.

Non-optimized-IR nodes will be unaware of the PMSI tunnel attribute flag definition with the additional flags for AR, so they will ignore the information in the flags field.

The *draft-ietf-bess-evpn-optimized-ir* describes the following three types of IR optimizations:

- Non-selective AR - the chosen AR-R replicates the BM traffic to all NVEs in the Ethernet VPN Instance (EVI) except for the source NVE.
- Selective AR - AR-Rs replicate BM traffic to only their AR-L set and the rest of the AR-Rs. Selective AR allows a "multi-stage" AR replication, as opposed to a "single-stage" AR replication.
- Pruned Flood Lists - AR-Ls can signal PFL flags to be pruned from the flood lists for BM or for unknown unicast traffic. PFL may be used in combination with AR.

SR OS Release 14.0.R4 only supports non-selective AR; therefore, the other two optimization modes are not described in this chapter.

Configure AR-R and AR-L

The AR-IP is configured on the AR-R, as follows:

```
*A:PE-1# configure service system vxlan assisted-replication-ip
  - assisted-replication-ip <ip-address>
  - no assisted-replication-ip

 <ip-address>         : a.b.c.d
```

The AR-IP is the IPv4 address of a loopback interface in the base router instance. When attempting to configure an AR-IP and the loopback address does not exist, the following error message is raised:

```
*A:PE-1# configure service system vxlan assisted-replication-ip 1.1.1.1
MINOR: SVCMGR #8110 Cannot change assisted-replicated address -
 loopback interface with address does not exist
```

The AR types replicator and leaf are configured in a VPLS with the following command:

```
*A:PE-1# configure service vpls 10 vxlan vni 1 assisted-replication
  - assisted-replication {replicator|leaf} [replicator-activation-time <seconds>]
  - no assisted-replication

 <replicator|leaf>    : replicator|leaf
 <seconds>            : [1..255]
```

When attempting to configure an AR-R before the AR-IP is set, the following error is raised:

```
*A:PE-1# configure service vpls 10 customer 1 create vxlan vni 1 create assisted-
replication replicator
MINOR: SVCMGR #8111 Cannot change assisted-replicated role -
 assisted replicator ip not set
```

The AR type (AR-R or AR-L) cannot be changed while being used by any BGP-EVPN service. The following error is raised in such a case:

```
*A:PE-1# configure service vpls 10 vxlan vni 1 assisted-replication replicator
MINOR: SVCMGR #8111 Cannot change assisted-replicated role - Evpn not shut
```

The assisted-replication-time can only be configured on leaf nodes. The following error is raised after an attempt to configure the assisted-replication-time on an AR-R:

```
*A:PE-1# configure service vpls 10 vxlan vni 1 assisted-
replication replicator replicator-activation-time 5
MINOR: SVCMGR #8112 Cannot change replicator activation time - valid only on leaf
```

The **replicator-activation-time** can optionally be activated, and works as follows. When the router creates an AR-R destination for the first time, the assisted-replication-timer must expire before this AR-R destination is eligible as candidate AR-R to forward BM traffic. Upon timer expiration, the router will then run the AR-R selection (service ID modulo the number of AR-Rs provides the selected AR-R in the ordered list of candidate AR-Rs). The AR-R EVPN destination will be created as "BM" and the destinations to the remaining nodes will be shown as "U".

The **replicator-activation-time** allows the AR-R some time to program the leaf VTEPs in the following cases:

- Configuration of a new AR-R
- AR-R rebooting
- AR-R going operationally down and up again

If the timer is zero (default value), the AR-R may receive packets from a VTEP that has not been programmed yet, in which case the AR-R drops the packets.

With the AR-Rs and AR-Ls configured, IMET AR routes can be exchanged. IR can be enabled or disabled independently of the AR configuration. The following command is required to enable IR inclusive multicast routes, and is enabled by default:

```
*A:PE-1# configure service vpls 10 bgp-evpn ingress-repl-inc-mcast-advertisement
```

BGP-EVPN Routes

By default, IR is enabled in BGP-EVPN. The following IMET IR route is sent from PE-5 (RNVE) to Route Reflector (RR) PE-1. The flags in the PMSI Tunnel Attribute (PTA) indicate that regular IR is used to forward BUM traffic (tunnel type: 0x06). The AR type is "None", because AR is disabled on PE-5. The IR-IP 192.0.2.5 is used as next-hop, originator IP address, and tunnel endpoint. The MPLS label corresponds to the VNI.

```
*A:PE-5# show debug
debug
    router "Base"
        bgp
            update
        exit
    exit
exit


12 2017/05/29 07:00:50.86 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 84
    Flag: 0x90 Type: 14 Len: 28 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.5
        Type: EVPN-Incl-
mcast Len: 17 RD: 192.0.2.5:1, tag: 0, orig_addr len: 32, orig_addr: 192.0.2.5
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:64500:1
        bgp-tunnel-encap:VXLAN
    Flag: 0xc0 Type: 22 Len: 9 PMSI:
        Tunnel-type Ingress Replication (6)
        Flags: (0x0)[Type: None BM: 0 U: 0 Leaf: not required]
        MPLS Label 1
        Tunnel-Endpoint 192.0.2.5
"
```

A similar IMET IR route is sent from AR-L PE-3 toward RR PE-1, as follows. The difference is that the flags indicate that PE-3 is configured as an AR-L for the VPLS. The IR-IP 192.0.2.3 is used as next-hop, originator address, and tunnel endpoint.

```
8 2017/05/29 07:00:51.19 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 84
    Flag: 0x90 Type: 14 Len: 28 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.3
        Type: EVPN-Incl-
mcast Len: 17 RD: 192.0.2.3:1, tag: 0, orig_addr len: 32, orig_addr: 192.0.2.3
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:64500:1
        bgp-tunnel-encap:VXLAN
    Flag: 0xc0 Type: 22 Len: 9 PMSI:
        Tunnel-type Ingress Replication (6)
        Flags: (0x10)[Type: AR Leaf BM: 0 U: 0 Leaf: not required]
        MPLS Label 1
        Tunnel-Endpoint 192.0.2.3
"
```

The IMET IR routes contain the system IP addresses of the nodes, not the AR-IPs.

The following AR route is advertised from AR-R PE-1. The tunnel type is AR and the flags indicate that PE-1 is configured as AR-R. The AR-IP 1.1.1.1 is the next-hop address, the originator address, and the tunnel endpoint.

```
4 2017/05/29 06:59:57.15 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.4
```

```
"Peer 1: 192.0.2.4: UPDATE
Peer 1: 192.0.2.4 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 84
    Flag: 0x90 Type: 14 Len: 28 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 1.1.1.1
        Type: EVPN-Incl-
mcast Len: 17 RD: 192.0.2.1:1, tag: 0, orig_addr len: 32, orig_addr: 1.1.1.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:64500:1
        bgp-tunnel-encap:VXLAN
    Flag: 0xc0 Type: 22 Len: 9 PMSI:
        Tunnel-type Assisted Replication (10)
        Flags: (0x8)[Type: AR Replicator BM: 0 U: 0 Leaf: not required]
        MPLS Label 1
        Tunnel-Endpoint 1.1.1.1
"
```

Besides IMET AR routes, PE-1 may also advertise IMET IR routes to the other nodes using IR-IP 192.0.2.1 (system IP address). By default, BGP-EVPN has IR enabled. For example, the following IMET IR route is advertised to PE-4:

```
3 2017/05/29 06:59:57.15 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.4
"Peer 1: 192.0.2.4: UPDATE
Peer 1: 192.0.2.4 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 84
    Flag: 0x90 Type: 14 Len: 28 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.1
        Type: EVPN-Incl-
mcast Len: 17 RD: 192.0.2.1:1, tag: 0, orig_addr len: 32, orig_addr: 192.0.2.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:64500:1
        bgp-tunnel-encap:VXLAN
    Flag: 0xc0 Type: 22 Len: 9 PMSI:
        Tunnel-type Ingress Replication (6)
        Flags: (0x0)[Type: None BM: 0 U: 0 Leaf: not required]
        MPLS Label 1
        Tunnel-Endpoint 192.0.2.1
"
```

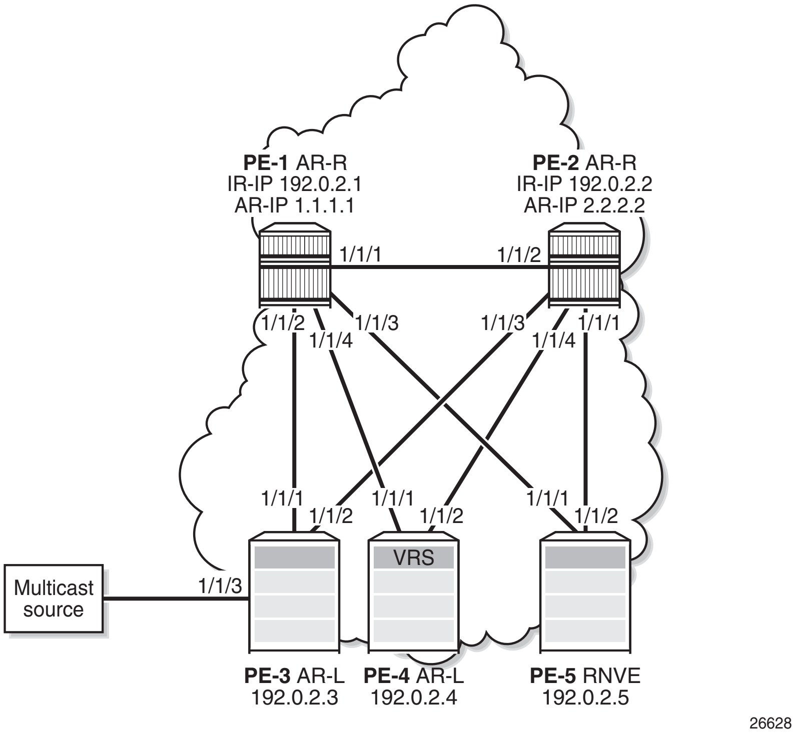
The following IMET routes have been received by PE-4:

```
*A:PE-4# show router bgp routes evpn inclusive-mcast
===============================================================================
 BGP Router ID:192.0.2.4        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN Inclusive-Mcast Routes
===============================================================================
Flag  Route Dist.         OrigAddr
      Tag                 NextHop
-------------------------------------------------------------------------------
u*>i  192.0.2.1:1         192.0.2.1
      0                   192.0.2.1
u*>i  192.0.2.1:1         1.1.1.1
      0                   1.1.1.1
u*>i  192.0.2.2:1         192.0.2.2
      0                   192.0.2.2
u*>i  192.0.2.2:1         2.2.2.2
      0                   2.2.2.2
u*>i  192.0.2.3:1         192.0.2.3
      0                   192.0.2.3
u*>i  192.0.2.5:1         192.0.2.5
      0                   192.0.2.5
-------------------------------------------------------------------------------
Routes : 6
===============================================================================
*A:PE-4#
```

Configuration

Figure 86 shows the example topology with PE-1 and PE-2 as AR-R nodes, PE-3 and PE-4 as AR-L nodes, and PE-5 as RNVE node. The multicast source is connected to PE-3, which is a low-performance node. PE-1 acts as an RR for all nodes.

Figure 86 **Example Topology**



The initial configuration on the nodes includes:

- Cards, MDAs, ports
- Router interfaces between the nodes
- IS-IS as IGP (alternatively, OSPF can be used)

BGP is configured for address family EVPN with RR PE-1. The BGP configuration on PE-1 is as follows:

```
configure
    router
        autonomous-system 64500
        bgp
            vpn-apply-import
            vpn-apply-export
            rapid-withdrawal
            split-horizon
            rapid-update evpn
            group "DC"
                family evpn
                cluster 192.0.2.1
                peer-as 64500
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
                neighbor 192.0.2.5
                exit
            exit
        exit
```

The BGP configuration on the other nodes is as follows:

```
configure
    router
        autonomous-system 64500
        bgp
            vpn-apply-import
            vpn-apply-export
            rapid-withdrawal
            split-horizon
            rapid-update evpn
            group "DC"
                family evpn
                peer-as 64500
                neighbor 192.0.2.1
                exit
            exit
        exit
```

VPLS 10 is configured on all nodes. PE-1 is configured as AR-R with AR-IP 1.1.1.1, which must be configured as loopback IPv4 address in the base router and as AR-IP that can be shared between services. When attempting to configure an AR-IP with an IP address that does not exist in the base router, the following error is raised:

```
*A:PE-1# configure service system vxlan assisted-replication-ip 1.1.1.1
MINOR: SVCMGR #8110 Cannot change assisted-replicated address -
 loopback interface with address does not exist
```

First, a loopback interface is configured in the base router. The IP address needs to be routable and, in this example, an export policy exporting this IP address is configured in IS-IS. Alternatively, a static route can be configured or an additional IS-IS passive interface can be configured for the loopback interface. The IP address is then configured as AR-IP in the **service system bgp-evpn** context and VPLS 10 in configured with AR-R role. PE-1 is configured as AR-R for VPLS 10, as follows:

```
configure
    router
        interface "AR-IP"
            address 1.1.1.1/32
            loopback
        exit
        policy-options
            begin
            prefix-list "AR-IP"
```

```
                prefix 1.1.1.1/32 exact
            exit
            policy-statement "export_AR-IP"
                entry 10
                    from
                        prefix-list "AR-IP"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit
        isis
            export "export_AR-IP"
        exit
    exit
    service
        system
            vxlan
                assisted-replication 1.1.1.1
            exit
        exit
        vpls 10 customer 1 create
            vxlan vni 1 create
                assisted-replication replicator
            exit
            bgp
            exit
            bgp-evpn
                evi 1
                vxlan
                    no shutdown
                exit
            exit
            no shutdown
        exit
```

The configuration is similar on PE-2, but with AR-IP 2.2.2.2 instead of 1.1.1.1.

PE-3 and PE-4 are configured as AR-L nodes for VPLS 10. No AR-IP needs to be configured. The configuration of VPLS 10 on PE-3 is as follows:

```
configure
    service
        vpls 10 customer 1 create
            vxlan vni 1 create
                assisted-replication leaf
            exit
            bgp
            exit
            bgp-evpn
                evi 1
                vxlan
                    no shutdown
                exit
            exit
            sap 1/1/3 create
            exit
            sap 1/2/1:1 create
            exit
            no shutdown
        exit
```

Multicast traffic enters SAP 1/1/3, whereas receiving hosts can be connected to other SAPs, such as SAP 1/2/1:1. The configuration of VPLS 10 on PE-4 is similar, but no multicast source is connected. When a node is configured as AR-L, optionally the **replicator-activation-time** can be configured to define the waiting time before the leaf can begin sending multicast traffic to a new replicator or a replicator that was rebooted. The default is zero seconds, in which case the AR-L starts sending packets to the AR-R without delay. Nokia recommends configuring a **replicator-activation-time** value different from zero.

```
*A:PE-3# configure service vpls 10 vxlan vni 1 assisted-
replication                                  - assisted-
replication {replicator|leaf} [replicator-activation-time <seconds>]
  - no assisted-replication
 <replicator|leaf>    : replicator|leaf
 <seconds>            : [1..255]
```

PE-5 is configured as an RNVE node for VPLS 10, as follows:

```
configure
    service
        vpls 10 customer 1 create
            vxlan vni 1 create
            exit
            bgp
            exit
            bgp-evpn
                evi 1
                vxlan
                    no shutdown
                exit
            exit
            sap 1/2/1:1 create
            exit
            no shutdown
        exit
```

BGP-EVPN IMET routes are exchanged between the nodes. The following IMET routes are used on AR-L PE-3, with two routes from each AR-R: one IR route with BGP next-hop 192.0.2.x and one AR route with BGP next-hop x.x.x.x (with x equal to 1 or 2).

```
*A:PE-3# show router bgp routes evpn inclusive-mcast
===============================================================================
 BGP Router ID:192.0.2.3        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
```

```
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN Inclusive-Mcast Routes
===============================================================================
Flag  Route Dist.         OrigAddr
      Tag                 NextHop
-------------------------------------------------------------------------------
u*>i  192.0.2.1:1         192.0.2.1
      0                   192.0.2.1
u*>i  192.0.2.1:1         1.1.1.1
      0                   1.1.1.1
u*>i  192.0.2.2:1         192.0.2.2
      0                   192.0.2.2
u*>i  192.0.2.2:1         2.2.2.2
      0                   2.2.2.2
u*>i  192.0.2.4:1         192.0.2.4
      0                   192.0.2.4
u*>i  192.0.2.5:1         192.0.2.5
      0                   192.0.2.5
-------------------------------------------------------------------------------
Routes : 6
===============================================================================
*A:PE-3#
```

When the AR-R has no local attachment circuits, such as SAPs or SDP-bindings, it should not generate regular IR routes. This can be controlled by disabling **ingress-repl-inc-mcast-advertisement** on PE-1 and PE-2, as follows:

```
configure
    service
        vpls 10
            bgp-evpn
                vxlan shutdown
                no ingress-repl-inc-mcast-advertisement
                vxlan no shutdown
            exit
        exit
```

When IR is disabled on the AR-Rs, no IR routes are sent to the other nodes and PE-3 only sees the AR routes from PE-1 and PE-2, as follows:

```
*A:PE-3# show router bgp routes evpn inclusive-mcast
===============================================================================
 BGP Router ID:192.0.2.3        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN Inclusive-Mcast Routes
===============================================================================
Flag  Route Dist.         OrigAddr
```

```
      Tag                 NextHop
-------------------------------------------------------------------------------
u*>i  192.0.2.1:1         1.1.1.1
      0                   1.1.1.1
u*>i  192.0.2.2:1         2.2.2.2
      0                   2.2.2.2
u*>i  192.0.2.4:1         192.0.2.4
      0                   192.0.2.4
u*>i  192.0.2.5:1         192.0.2.5
      0                   192.0.2.5
-------------------------------------------------------------------------------
Routes : 4
```

The detailed information about the AR route sent by AR-R PE-1 can be shown with the following command. The AR tunnel has endpoint 1.1.1.1.

```
*A:PE-3# show router bgp routes evpn inclusive-mcast rd 192.0.2.1:1 hunt
===============================================================================
 BGP Router ID:192.0.2.3        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN Inclusive-Mcast Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : N/A
Nexthop        : 1.1.1.1
From           : 192.0.2.1
Res. Nexthop   : 192.168.13.1
Local Pref.    : 100                    Interface Name : int-PE-3-PE-1
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : target:64500:1 bgp-tunnel-encap:VXLAN
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.1
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
EVPN type      : INCL-MCAST
ESI            : N/A
Tag            : 0
Originator IP  : 1.1.1.1
Route Dist.    : 192.0.2.1:1
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h25m07s
-------------------------------------------------------------------------------
```

```
PMSI Tunnel Attributes :
Tunnel-type    : Assisted Replication
Flags          : Type: AR-Replicator(1) BM: 0 U: 0 Leaf: not required
MPLS Label     : VNI 1
Tunnel-Endpoint: 1.1.1.1
-------------------------------------------------------------------------------

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-3#
```

The following command shows the VXLAN-related information for VPLS 10 on PE-3:

```
*A:PE-3# show service id 10 vxlan
===============================================================================
Vxlan Src Vtep IP: N/A
===============================================================================
VPLS VXLAN, Ingress VXLAN Network Id: 1
Creation Origin: manual
Assisted-Replication: leaf      Replicator-Activation-Time: None
RestProtSrcMacAct: none

===============================================================================
VPLS VXLAN service Network Specifics
===============================================================================
Ing Net QoS Policy : none                            Vxlan VNI Id     : 1
Ingress FP QGrp    : (none)                          Ing FP QGrp Inst : (none)

===============================================================================
Egress VTEP, VNI
===============================================================================
VTEP Address                            Egress VNI  Num. MACs   Mcast Oper  L2
                                                                      State PBR
-------------------------------------------------------------------------------
1.1.1.1                                 1           0           BM    Up    No
2.2.2.2                                 1           0           -     Up    No
192.0.2.4                               1           0           U     Up    No
192.0.2.5                               1           0           U     Up    No
-------------------------------------------------------------------------------
Number of Egress VTEP, VNI : 4
-------------------------------------------------------------------------------
===============================================================================
*A:PE-3#
```

PE-3 is configured as AR-L and no **replicator-activation-time** is defined (default). Four egress VTEPs are listed: the system IP addresses are used for IR routes and the AR-IPs are used for AR routes. All BM traffic will be forwarded to AR-IP 1.1.1.1 on PE-1. The AR-R in use is selected by the modulo operation on the service ID (10). In this example, two AR-Rs are available, and the service ID modulo 2 equals zero: 10 mod 2 = 0. This is the lowest possible outcome, so the first AR-R in the ordered candidate list is used. The AR-Rs are ordered by IP and VNI, with candidate 0 the lowest IP and VNI.

```
*A:PE-3# show service id 10 vxlan assisted-replication replicator

===============================================================================
Vxlan AR Replicator Candidates
===============================================================================
VTEP Address           Egress VNI     In Use  In Candidate List Pending Time
-------------------------------------------------------------------------------
1.1.1.1                1              yes     yes               0
2.2.2.2                1              no      yes               0
-------------------------------------------------------------------------------
Number of entries : 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-3#
```

Within a service, no load-sharing is done between the AR-Rs. However, different AR-Rs can be used for different services.

- If PE-3 were configured as AR-L in VPLS 11, the calculation would be as follows: 11 mod 2 = 1; therefore, the second AR-R in the list would be selected.
- When three AR-Rs were available for VPLS 11, the calculation would be: 11 mod 3 = 2, so the third AR-R in the list would be used.

In case different VNIs are configured for the AR-Rs, the lowest IP address is always higher in the list, even when the VNI is higher. This can be shown when the VPLS VXLAN configuration on PE-1 is modified with VNI 99 instead of VNI 1, as follows:

```
*A:PE-1# configure service vpls 10 bgp-evpn vxlan shutdown
*A:PE-1# configure service vpls 10 no vxlan vni 1
*A:PE-1# configure service vpls 10 vxlan vni 99 create assisted-replication
replicator
*A:PE-1# configure service vpls 10 bgp-evpn vxlan no shutdown
```

The list of AR-Rs on PE-3 shows that the first entry is the VTEP with the lowest IP address (1.1.1.1), even though the VNI 99 is higher than 1:

```
*A:PE-3# show service id 10 vxlan assisted-replication replicator
===============================================================================
Vxlan AR Replicator Candidates
===============================================================================
VTEP Address           Egress VNI     In Use  In Candidate List Pending Time
-------------------------------------------------------------------------------
1.1.1.1                99             yes     yes               0
```

```
2.2.2.2                1              no      yes               0
-------------------------------------------------------------------------------
Number of entries : 2
```

Note: If the AR-IP loopback interface is down, BGP will not withdraw the AR route. When the route to the AR-IP is signaled using IGP, the route will be removed from the routing table and the AR-L will select another AR-R. However, when a static route is defined for the AR-IP, a black-hole exists when the AR-IP interface is down.

PE-5 is configured as an RNVE node that signals regular IMET IR routes and is unaware of the AR-R and AR-L roles in the EVI. RNVE nodes ignore IMET AR routes. In the example, only PE-3, PE-4, and PE-5 send IMET IR updates, so the list of VTEP addresses on PE-5 only contains PE-3 and PE-4, as follows:

```
*A:PE-5# show service id 10 vxlan
===============================================================================
Vxlan Src Vtep IP: N/A
===============================================================================
VPLS VXLAN, Ingress VXLAN Network Id: 1
Creation Origin: manual
Assisted-Replication: none
RestProtSrcMacAct: none

===============================================================================
VPLS VXLAN service Network Specifics
===============================================================================
Ing Net QoS Policy : none                            Vxlan VNI Id     : 1
Ingress FP QGrp    : (none)                          Ing FP QGrp Inst : (none)

===============================================================================
Egress VTEP, VNI
===============================================================================
VTEP Address                            Egress VNI  Num. MACs   Mcast Oper  L2
                                                                      State PBR
-------------------------------------------------------------------------------
192.0.2.3                               1           1           BUM   Up    No
192.0.2.4                               1           0           BUM   Up    No
-------------------------------------------------------------------------------
Number of Egress VTEP, VNI : 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-5#
```

The RNVE is unaware of AR-Rs; therefore, the list of AR-Rs is empty on PE-5:

```
*A:PE-5# show service id 10 vxlan assisted-replication replicator

===============================================================================
Vxlan AR Replicator Candidates
===============================================================================
VTEP Address           Egress VNI     In Use  In Candidate List Pending Time
-------------------------------------------------------------------------------
No Matching Entries
```

```
===============================================================================
*A:PE-5#
```

Verification of Multicast Traffic

The multicast source connected to PE-3 generates multicast traffic. PE-3 acts as AR-L and forwards the multicast packets to AR-R PE-1. In this example topology, multicast traffic enters port 1/1/3 on PE-3 and is forwarded to egress port 1/1/1 toward PE-1. Port statistics are cleared and traffic is generated, then the port statistics are verified.

```
*A:PE-3# show port 1/1/[1..4] statistics

===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/1                         2             209           76881         8764456
===============================================================================

===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/2                         2             272               1             136
===============================================================================

===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/3                     76879         4920256               0               0
===============================================================================
*A:PE-3#
```

Besides the multicast traffic, IGP signaling is sent and received on the network interfaces. This explains why the counters on the network interface 1/1/1 toward PE-1 show a slightly higher value than on the interface 1/1/3 toward the multicast source. No multicast traffic is forwarded to PE-2, which is an AR-R candidate, but not used. AR-L PE-3 selected PE-1 for VPLS 10.

When the AR-R PE-1 receives the multicast traffic from PE-3, it forwards the traffic to PE-4 and PE-5 within the VXLAN service. The VXLAN information for VPLS 10 on PE-1 shows that PE-2 is not in the list of egress VTEPs. The reason is that PE-2 does not have any SAPs or SDP-bindings and no IMET IR route is sent by PE-2 because **ingress-repl-inc-mcast-advertisement** is disabled.

```
*A:PE-1# show service id 10 vxlan
===============================================================================
Vxlan Src Vtep IP: N/A
===============================================================================
VPLS VXLAN, Ingress VXLAN Network Id: 1
Creation Origin: manual
Assisted-Replication: replicator
RestProtSrcMacAct: none

===============================================================================
VPLS VXLAN service Network Specifics
===============================================================================
Ing Net QoS Policy : none                            Vxlan VNI Id     : 1
Ingress FP QGrp    : (none)                          Ing FP QGrp Inst : (none)

===============================================================================
Egress VTEP, VNI
===============================================================================
VTEP Address                            Egress VNI  Num. MACs   Mcast Oper  L2
                                                                      State PBR
-------------------------------------------------------------------------------
192.0.2.3                               1           1           BUM   Up    No
192.0.2.4                               1           0           BUM   Up    No
192.0.2.5                               1           0           BUM   Up    No
-------------------------------------------------------------------------------
Number of Egress VTEP, VNI : 3
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

AR-R PE-1 receives the multicast traffic from PE-3 on port 1/1/2 and forwards it to the egress ports 1/1/3 toward PE-5 and 1/1/4 toward PE-4, as follows. No multicast traffic needs to be forwarded to egress port 1/1/1 toward PE-2. Source squelching ensures that the traffic is not sent back to the originator AR-L PE-3. PE-1 has no local SAPs or SDP-bindings.

```
*A:PE-1# show port 1/1/[1..4] statistics

===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/1                         3             342               2             162
===============================================================================

===============================================================================
Port Statistics on Slot 1
```

```
================================================================================
Port                    Ingress          Ingress         Egress           Egress
Id                      Packets           Octets         Packets          Octets
--------------------------------------------------------------------------------
1/1/2                     60681          6917634               2             209
================================================================================

================================================================================
Port Statistics on Slot 1
================================================================================
Port                    Ingress          Ingress         Egress           Egress
Id                      Packets           Octets         Packets          Octets
--------------------------------------------------------------------------------
1/1/3                         3              295           60685         6918002
================================================================================

================================================================================
Port Statistics on Slot 1
================================================================================
Port                    Ingress          Ingress         Egress           Egress
Id                      Packets           Octets         Packets          Octets
--------------------------------------------------------------------------------
1/1/4                         5              504           60685         6918002
================================================================================
*A:PE-1#
```

An egress AR-L or RNVE node will perform regular egress BUM forwarding procedures. Packets will be replicated to local SAPs or SDP-bindings, but not to VXLAN-bindings.

AR-R Failure Scenarios

When the AR-IP interface on the used AR-R is down for any kind of reason, the route to this AR-IP will be removed from the routing table on AR-L PE-3, and PE-3 will select AR-R PE-2. To simulate an AR-R failure, the AR-IP interface on PE-1 is disabled, as follows:

```
*A:PE-1# configure router interface "AR-IP" shutdown
```

After a while, the routing table on PE-3 will not contain an entry for prefix 1.1.1.1/32 anymore, as follows:

```
*A:PE-3# show router route-table 1.1.1.1/32

================================================================================
Route Table (Router: Base)
================================================================================
Dest Prefix[Flags]                             Type    Proto     Age        Pref
      Next Hop[Interface Name]                                     Metric
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
```

```
No. of Routes: 0
```

AR-R PE-1 is not eligible anymore when the AR-IP is not reachable. PE-2 is now selected as AR-R, so BM traffic is forwarded to PE-2. Log 99 on PE-3 shows the change in AR-R from PE-1 to PE-2, as follows:

```
77 2017/05/29 10:05:23.54 UTC MINOR: SVCMGR #2098 Base
"Assisted replicator in service 10 changed to VTEP 2.2.2.2, Egress VNI 1."
```

The VXLAN-related information for VPLS 10 on PE-3 does not include VTEP 1.1.1.1 anymore, as follows:

```
*A:PE-3# show service id 10 vxlan
================================================================================
Vxlan Src Vtep IP: N/A
================================================================================
VPLS VXLAN, Ingress VXLAN Network Id: 1
Creation Origin: manual
Assisted-Replication: leaf      Replicator-Activation-Time: None
RestProtSrcMacAct: none
================================================================================
VPLS VXLAN service Network Specifics
================================================================================
Ing Net QoS Policy : none                            Vxlan VNI Id     : 1
Ingress FP QGrp    : (none)                          Ing FP QGrp Inst : (none)
================================================================================
Egress VTEP, VNI
================================================================================
VTEP Address                            Egress VNI  Num. MACs   Mcast Oper  L2
                                                                      State PBR
--------------------------------------------------------------------------------
2.2.2.2                                 1           0           BM    Up    No
192.0.2.4                               1           0           U     Up    No
192.0.2.5                               1           0           U     Up    No
--------------------------------------------------------------------------------
Number of Egress VTEP, VNI : 3
```

Only PE-2 is listed as AR-R for VPLS 10 on PE-3, and PE-2 is the selected AR-R for VPLS 10, as follows:

```
*A:PE-3# show service id 10 vxlan assisted-replication replicator
================================================================================
Vxlan AR Replicator Candidates
================================================================================
VTEP Address           Egress VNI     In Use  In Candidate List Pending Time
--------------------------------------------------------------------------------
2.2.2.2                1              yes     yes               0
--------------------------------------------------------------------------------
Number of entries : 1
```

Incoming multicast traffic on port 1/1/3 on PE-3 will now be forwarded to port 1/1/2 toward PE-2, as follows:

```
*A:PE-3# clear port 1/1/[1..4] statistics
```

```
*A:PE-3# sleep 10
*A:PE-3# show port 1/1/[1..4] statistics

===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/1                         3             345               3             345
===============================================================================

===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/2                         3             345          169566        19330527
===============================================================================

===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/3                    169698        10860672               0               0
===============================================================================
*A:PE-3#
```

When the AR-IP interface on AR-R PE-2 is also disabled, no AR-R is available anymore and PE-3 will revert to IR instead.

```
*A:PE-2# configure router interface "AR-IP" shutdown
```

The following log 99 message on AR-L PE-3 indicates that there is no AR-R anymore (VTEP 0.0.0.0, Egress VNI 0).

```
78 2017/05/29 10:05:23.55 UTC MINOR: SVCMGR #2098 Base
"Assisted replicator in service 10 changed to VTEP 0.0.0.0, Egress VNI 0."
```

The VXLAN-related information for VPLS 10 on PE-3 does not include any AR-R (VTEP 1.1.1.1 or 2.2.2.2) anymore, as follows:

```
*A:PE-3# show service id 10 vxlan
===============================================================================
Vxlan Src Vtep IP: N/A
===============================================================================
VPLS VXLAN, Ingress VXLAN Network Id: 1
Creation Origin: manual
Assisted-Replication: leaf      Replicator-Activation-Time: None
RestProtSrcMacAct: none

===============================================================================
VPLS VXLAN service Network Specifics
```

```
===============================================================================
Ing Net QoS Policy : none                            Vxlan VNI Id     : 1
Ingress FP QGrp    : (none)                          Ing FP QGrp Inst : (none)

===============================================================================
Egress VTEP, VNI
===============================================================================
VTEP Address                            Egress VNI  Num. MACs   Mcast Oper  L2
                                                                      State PBR
-------------------------------------------------------------------------------
192.0.2.4                               1           0           BUM   Up    No
192.0.2.5                               1           0           BUM   Up    No
-------------------------------------------------------------------------------
Number of Egress VTEP, VNI : 2

*A:PE-3# show service id 10 vxlan assisted-replication replicator

===============================================================================
Vxlan AR Replicator Candidates
===============================================================================
VTEP Address           Egress VNI     In Use  In Candidate List Pending Time
-------------------------------------------------------------------------------
No Matching Entries
```

In this case, IR is done for all BUM traffic toward PE-4 and PE-5.

Conclusion

AR uses replicators to forward broadcast and multicast traffic on behalf of less-performing nodes that are configured as AR-Ls. AR is primarily used for L2 multicast optimization in data centers, but might also be used in any network using overlay EVPN-VXLAN tunnels.

LDP VPLS Using BGP Auto-Discovery

This chapter provides information about LDP VPLS using BGP Auto-Discovery.

Topics in this chapter include:

- Applicability
- Summary
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SR OS Release 9.0.R3. The CLI in this edition is based on SR OS Release 15.0.R2. There are no prerequisites for this configuration.

Summary

MPLS-based Virtual Private LAN Services (VPLS) may have many different provisioning models to allow the signaling of pseudowires between Provider Edge (PE) routers containing VPLS instances.

Network Management System (NMS) provisioning using Label Distribution Protocol (LDP) signaling is a well understood method of provisioning of Layer 2 VPLS services as described in RFC 4762. This relies on the provisioning of pseudowires between VPLS instances using LDP signaling with a common Virtual Circuit (VC) identifier within the label mapping message to instantiate pseudowires.

Border Gateway Protocol (BGP) Auto-Discovery (RFC 6074) is an alternative method of provisioning of Layer 2 PE routers containing VPLS service instances to those described above where PEs in a common VPLS instance are automatically discovered using BGP Auto-Discovery (BGP-AD) techniques.

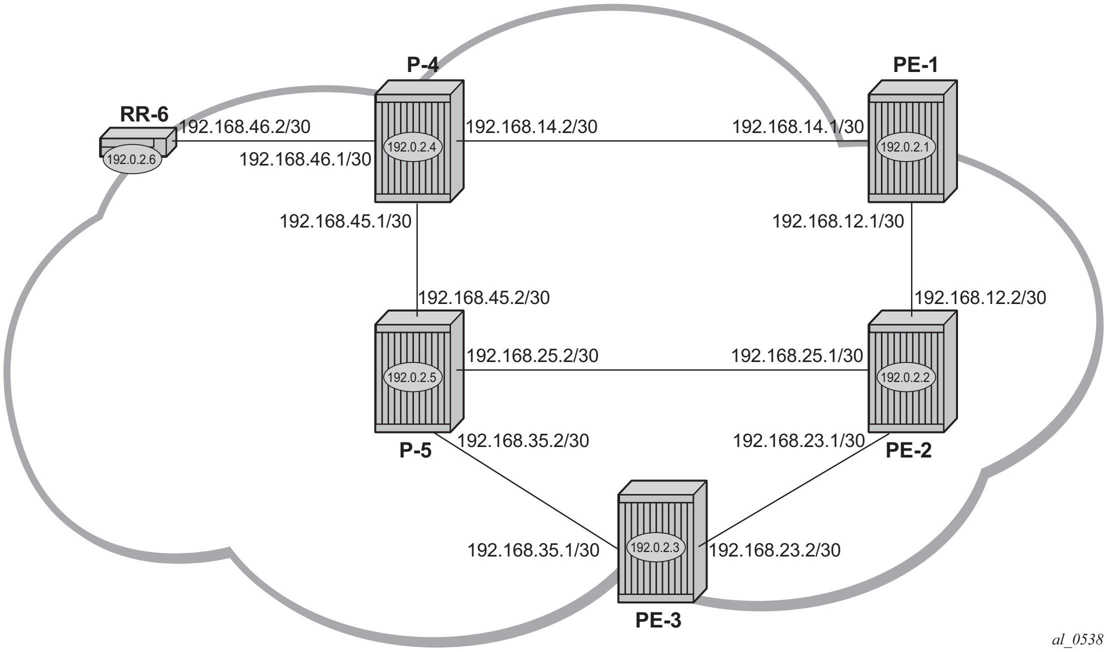
Each PE router advertises the presence of VPLS instances to other PE routers using defined parameters within a BGP update message.

LDP is used as the pseudowire signaling protocol and relies on the auto-discovery of VPLS endpoints to instantiate pseudowires instead of manually provisioning virtual circuits. Locally configured parameters, along with BGP learned parameters, are used to determine local and remote VPLS endpoints, which are used by LDP to signal service labels to peer routers.

Knowledge of BGP-auto-discovery RFC 6074 architecture and functionality, RFC 4447 Pseudo-wire set-up using label distribution protocol is assumed throughout this section, as well as knowledge of Multi-Protocol BGP (MP-BGP).

Overview

Figure 87 **Example Topology**



The example topology is displayed in Figure 87. The setup uses six 7x50 nodes located in the same autonomous system (AS). There are three PEs and RR-6 will act as a route reflector for the AS. The PE routers are all VPLS aware. The provider (P) routers are VPLS unaware and do not take part in the BGP process. A full mesh VPLS between PE-1, PE-2 and PE-3 is described.

The following configuration tasks should be completed as a pre-requisite:

- IS-IS or OSPF should be enabled on all network interfaces between each of the PE/P routers and route reflector RR-6.

- MPLS should be configured on all interfaces between PE and P routers; MPLS is not required between P-4 and RR-6.
- LDP should be configured on interfaces between PE and P routers. It is not required between P-4 and RR-6.
- RSVP protocol is disabled by default, so the RSVP protocol should be enabled.

BGP-AD

In this architecture, a VPLS service is a collection of local VPLS instances present on a number of PEs in a provider network. In this context, VPLS-aware devices are PE routers. Each VPLS instance has a unique identifier known as the VPLS identifier (VPLS-id). All PEs that have this VPLS instance present will have a common VPLS-id configured.

Each VPLS instance within a PE contains a Virtual Switching Instance (VSI). The VPLS attachment circuits and pseudowires are associated with the VSI. Each VSI within a given VPLS has a unique identifier called the VSI identifier (VSI-id) and is a concatenation of the VPLS-id plus an IP address, usually the system IP address.

The PEs communicate with each other at the control plane level by means of BGP updates containing BGP Layer 2 Network Layer Reachability Information (NLRI). Each update contains enough information for a PE to determine the presence of other local VPLS instances on peering PEs. In turn, this allows peer PE routers to set up pseudowire connectivity using LDP signaling for data flow between peers containing a local VPLS within the same VPLS instances.

Each update contains parameters usually associated with Multi-Protocol BGP updates:

- NLRI encoded as route-target (usually the VPLS-id) and PE system address.
- Next-Hop — The system IP address of the sending PE router.
- Extended communities — Contains the route target extended community and the VPLS-id as community values.

Each VPLS instance is configured with import and export route target extended communities to create the required pseudowire topology by controlling the distribution of each NLRI.

The purpose of this section is to describe the provisioning of a VPLS instance across three PE routers. A full mesh of pseudowires interconnects the VSI of each PE within the VPLS instance. A single attachment circuit is also configured on each VSI.

Configuration

The first step is to configure an MP-iBGP session using the L2VPN address family between each of the PEs and the route reflector.

The configuration for PE-1 is as follows:

```
configure
     router
        autonomous-system 65536
        bgp
            group "internal"
                family l2-vpn
                peer-as 65536
                neighbor 192.0.2.6
                exit
            exit
            no shutdown
        exit
    exit
```

The configuration for the other PE nodes is identical. The IP addresses can be derived from Figure 87.

The configuration for route reflector RR-6 is as follows:

```
configure
     router
     autonomous-system 65536
        bgp
            cluster 1.1.1.1
            group "rr-internal"
                family l2-vpn
                peer-as 65536
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
            exit
            no shutdown
        exit
    exit
```

On PE-1, verify that the BGP session with RR-6 is established with address family l2-vpn capability negotiated:

```
*A:PE-1# show router bgp neighbor 192.0.2.6

===============================================================================
BGP Neighbor
===============================================================================
-------------------------------------------------------------------------------
Peer                 : 192.0.2.6
Description          : (Not Specified)
Group                : internal
-------------------------------------------------------------------------------
Peer AS              : 65536            Peer Port            : 51039
Peer Address         : 192.0.2.6
Local AS             : 65536            Local Port           : 179
Local Address        : 192.0.2.1
Peer Type            : Internal         Dynamic Peer         : No
State                : Established      Last State           : Established
Last Event           : recvKeepAlive
Last Error           : Cease (Connection Collision Resolution)
Local Family         : L2-VPN
Remote Family        : L2-VPN
---snip---
Local Capability     : RtRefresh MPBGP 4byte ASN
Remote Capability    : RtRefresh MPBGP 4byte ASN
---snip---
-------------------------------------------------------------------------------
Neighbors shown : 1
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-1#
```

On RR-6, show that BGP sessions with each PE are established, and have correctly negotiated the l2-vpn address family capability.

```
*A:RR-6# show router bgp summary all

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.1
Def. Instance  65536        3    0 00h00m22s 0/0/0 (L2VPN)
                            3    0
192.0.2.2
Def. Instance  65536        3    0 00h00m22s 0/0/0 (L2VPN)
                            3    0
192.0.2.3
Def. Instance  65536        3    0 00h00m22s 0/0/0 (L2VPN)
                            3    0

-------------------------------------------------------------------------------
*A:RR-6#
```

A full mesh of RSVP Label Switched Paths (LSPs) is configured between the PE routers. For reference, the MPLS interface configuration and LSPs for PE-1 to PE-2 and PE-3 is:

```
*A:PE-1# configure
    router
        mpls
            interface "int-PE-1-P-4"
                no shutdown
            exit
            interface "int-PE-1-PE-2"
                no shutdown
            exit
            path "loose"
                no shutdown
            exit
            lsp "LSP-PE-1-PE-2"
                to 192.0.2.2
                primary "loose"
                exit
                no shutdown
            exit
            lsp "LSP-PE-1-PE-3"
                to 192.0.2.3
                primary "loose"
                exit
                no shutdown
            exit
            no shutdown
```

VPLS PE Configuration

Pseudowire-Templates

Pseudowire templates are used by BGP to dynamically instantiate Service Distribution Point (SDP) bindings. For a given service, pseudowire templates signal the egress service de-multiplexer labels used by remote PEs to reach the local PE.

The template determines the signaling parameters of the pseudowire, control word presence, plus other usage characteristics such as Split Horizon Groups (SHG), MAC-pinning, filters, and so on.

The MPLS transport tunnel between PE routers can be signaled using either LDP or RSVP.

LDP based pseudowires can be automatically instantiated; RSVP based SDPs have to be pre-provisioned.

Pseudowire Templates for Auto-SDP Creation using LDP

In order to use an LDP transport tunnel for data flow between PEs, it is necessary for link layer LDP to be configured between all PEs/Ps so that a transport label for each PE's system interface address is available. Using this mechanism SDPs can be auto-instantiated with SDP IDs starting at 17407. Any subsequent SDPs created use SDP-ids decrementing from this value.

A pseudowire template is required which may contain a split-horizon group. Each SDP created with this template is contained within the configured split horizon group so that traffic cannot be forwarded between them.

```
configure
    service
        pw-template 1 create
            split-horizon-group "vpls-shg"
            exit
        exit
```

A pseudowire template can also be created that does not contain a split horizon group. The split horizon group can then be specified when the pw-template is included within the service.

```
configure
    service
        pw-template 2 create
        exit
```

Pseudowire Templates for Provisioned SDPs using RSVP

To use an RSVP tunnel as transport between PEs, it is necessary to bind the RSVP LSPs to the SDPs between each PE.

SDP 12 from PE-1 to PE-2 is configured on PE-1, as follows:

```
*A:PE-1# configure
    service
        sdp 12 mpls create
            far-end 192.0.2.2
            lsp "LSP-PE-1-PE-2"
            no shutdown
```

To create an SDP within a service that uses the RSVP transport tunnel, a pseudowire template is required that has the **use-provisioned-sdp** parameter.

```
*A:PE-1# configure
    service
        pw-template 3 use-provisioned-sdp create
```

```
                exit
            exit
```

Alternatively, the **prefer-provisioned-sdp** parameter can be used, see chapter LDP VPLS Using BGP Auto-Discovery - Prefer Provisioned SDP.

VPLS BGP-AD using Auto-Provisioned SDPs

Figure 88 **VPLS Instance with Auto-Provisioned SDPs**

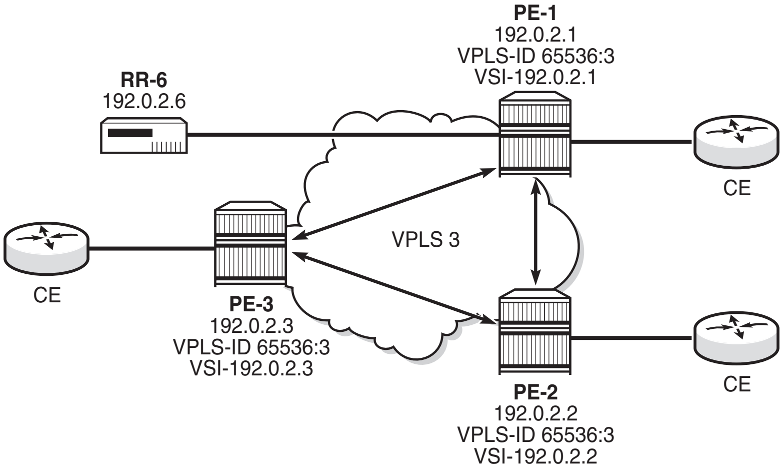


Figure 88 shows a schematic of a VPLS instance where the SDPs are auto-provisioned. SDPs are instantiated by a PE router using LDP signaling upon receipt of BGP Auto-discovery (BGP-AD) updates from peer PE routers.

PE-1 Configuration:

The following output shows the configuration required for a VPLS service using a pseudowire template configured for auto-provisioning of SDPs.

VPLS 3 is configured on PE-1, as follows:

```
configure
    service
        vpls 3 customer 1 create
            bgp
                route-distinguisher 65536:3
                route-target export target:65536:3 import target:65536:3
                pw-template-binding 2 split-horizon-group "vpls-shg"
                                      import-rt "target:65536:3"
                exit
            exit
            bgp-ad
```

```
                vpls-id 65536:3
                vsi-id
                    prefix 192.0.2.1
                exit
                no shutdown
            exit
            sap 1/1/4:3.0 create
            exit
            no shutdown
        exit
```

Within the **bgp** context, the pseudowire template is referenced which can be linked to a split-horizon-group and an import route-target, if required.

Within the **bgp-ad** context, the signaling parameters are configured. These are two parameters used by each PE to determine the presence of a VPLS instance on a PE router. In turn, these are translated into endpoint identifiers for LDP signaling of pseudowires. As previously discussed, these parameters are:

- VPLS-id - a unique identifier of the VPLS instance. Each PE that is a member of a VPLS must share the same VPLS-id. This is inserted as an extended community value in the format AS:n. In this case, the VPLS-id for VPLS 3 is 65536:3. This is a mandatory parameter and if it is not configured, it is not possible to enable BGP-AD using no shutdown.
- Virtual Switching Instance (VSI) prefix — This identifies a specific instance of the VPLS. This must be unique within the VPLS instance, and is encoded using the 4 byte dotted decimal notation. Generally, the system address is used as the VSI prefix. If this parameter is not configured, then the system address is used automatically.

The VPLS-id and VSI prefix for VPLS 3 on each PE is shown in Figure 88.

The VPLS-id and VSI prefix are concatenated to form a unique VSI-id. In this case, PE-1 has a VSI-id of 65536:3:192.0.2.1. This uniquely identifies the VPLS instance on each individual PE and is advertised as an L2 VPN BGP update.

A BGP-AD update is transmitted to all other PEs via the Route Reflector as follows:

```
*A:PE-1# show router bgp routes l2-vpn rd 65536:3 hunt
===============================================================================
 BGP Router ID:192.0.2.1        AS:65536       Local AS:65536
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP L2VPN Routes
===============================================================================
-------------------------------------------------------------------------------
```

```
RIB In Entries
-------------------------------------------------------------------------------
---snip---
-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
Route Type     : AutoDiscovery
Route Dist.    : 65536:3
Prefix         : 192.0.2.1
Nexthop        : 192.0.2.1
To             : 192.0.2.6
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : target:65536:3 l2-vpn/vrf-imp:65536:3
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.6
Origin         : IGP
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
-------------------------------------------------------------------------------
Routes : 4
===============================================================================
A:PE-1#
```

The preceding BGP update is transmitted by PE-1 and has route type auto discovery.

In this L2 VPN update, the VPLS-id is encoded as the L2VPN extended community 65536:3.

The VSI is seen as the prefix 192.0.2.1. The combination of the VPLS-id and the VSI forms the VSI-id and uniquely identifies the VPLS instance within this PE router.

The nexthop is also encoded as the local system IP address 192.0.2.1, which allows remote PEs to identify a suitable transport tunnel to PE-1 and for the targeted-LDP peer for instantiating the SDP.

As can be seen within the update, the VPLS-id 65536:3 is also used to determine the route target extended community and the route distinguisher.

PE-2 Configuration

On PE-2, VPLS 3 is created using pseudowire template 1, with VPLS-id 65536:3 and VSI-id prefix 192.0.2.2 (system IP address), as follows”

```
configure
    service
        vpls 3 customer 1 create
```

```
                bgp
                    route-distinguisher 65536:3
                    route-target export target:65536:3 import target:65536:3
                    pw-template-binding 2 split-horizon-group "vpls-shg"
                                          import-rt "target:65536:3"
                    exit
                exit
                bgp-ad
                    vpls-id 65536:3
                    vsi-id
                        prefix 192.0.2.2
                    exit
                    no shutdown
                exit
                sap 1/1/4:3.0 create
                exit
                no shutdown
            exit
```

PE-3 Configuration

On PE-3, VPLS 3 is created using pseudowire template 2, with VPLS-id 65536:3—identical to the VPLS-id of PE-1 and PE-2—and VSI-id 192.0.2.3 (system IP address), as follows:

```
On PE-3:
configure
    service
        vpls 3 customer 1 create
            bgp
                route-distinguisher 65536:3
                route-target export target:65536:3 import target:65536:3
                pw-template-binding 2 split-horizon-group "vpls-shg"
                                      import-rt "target:65536:3"
                exit
            exit
            bgp-ad
                vpls-id 65536:3
                vsi-id
                    prefix 192.0.2.3
                exit
                no shutdown
            exit
            sap 1/1/4:3.0 create
            exit
            no shutdown
        exit
```

PE-1 Service Operation Verification

The following output shows that the service is operationally up on PE-1:

```
*A:PE-1# show service id 3 base

===============================================================================
Service Basic Information
```

```
===============================================================================
Service Id        : 3                   Vpn Id            : 0
Service Type      : VPLS
---snip---
Admin State       : Up                  Oper State        : Up
MTU               : 1514                Def. Mesh VC Id   : 3
SAP Count         : 1                   SDP Bind Count    : 2
---snip---
-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/4:3.0                            qinq         1522    1522    Up   Up
sdp:17406:4294967294 SB(192.0.2.2)       BgpAd        0       1556    Up   Up
sdp:17407:4294967295 SB(192.0.2.3)       BgpAd        0       1556    Up   Up
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-1#
```

As seen from the output, the service is operationally up, with the SAPs and SDPs also up. The **SB** flag indicates that the SDP is of type spoke BGP.

BGP is used to discover the VPLS endpoints and exchange network reachability information. LDP is used to signal the pseudowires between the PEs.

LDP signaling occurs when each PE has discovered the endpoints of the VPLS instance. This compares with the use of the provisioned virtual circuit IDs used in an NMS provisioned VPLS instances as per RFC 4762.

Verification of the ability of PE-1 to reach the other PE routers with VSIs within the VPLS instance can be seen from the Layer 2 routing table as follows:

```
*A:PE-1# show service l2-route-table bgp-ad

========================================================================
Services: L2 Route Information - Summary
========================================================================
Svc Id     L2-Routes (RD-Prefix)                Next Hop        Origin
               Sdp Bind Id                          PW Temp Id
------------------------------------------------------------------------
3          *65536:3-192.0.2.2                   192.0.2.2       BGP-L2
               17406:4294967294                     2
3          *65536:3-192.0.2.3                   192.0.2.3       BGP-L2
               17407:4294967295                     2
------------------------------------------------------------------------
No. of L2 Route Entries: 2
========================================================================
*A:PE-1#
```

This output shows the presence of the signaled pseudowire SDPs. SDPs from PE-1 to PE-2 and PE-3 are signaled using LDP Forwarding Equivalence Class (FEC) Element 129.

Each PE router uses targeted LDP to signal the local and remote endpoints. If there is an endpoint match, then SDPs are instantiated. This compares with the use of LDP for NMS provisioned SDPs, which uses virtual circuit IDs to signal pseudowires using LDP FEC Element 128.

In order to signal the SDPs, the following parameters are required:

1. Attachment Group Identifier (AGI): this is used to carry the VPLS-id of the local PE router VPLS instance. The VPLS-id must be identical for all PEs in the same VPLS instance.
2. Source Attachment Individual Identifier (SAII) and Target Attachment Individual Identifier (TAII): These use AII type 1 (RFC 4446) and are used to carry the NLRI (VSI-id minus the RD) of the remote PE router VPLS instance.

The AGI for each PE must be identical. SAII and TAII must be different.

The following shows the service LDP bindings for VPLS 3 on PE-1:

```
*A:PE-1# show router ldp bindings services service-id 3

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        S - Status Signaled Up,  D - Status Signaled Down, e - Label ELC
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
Service Type:
        E - Epipe Service, V - VPLS Service, M - Mirror Service
        A - Apipe Service, F - Fpipe Service, I - IES Service, R - VPRN service
        P - Ipipe Service, C - Cpipe Service
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
===============================================================================
LDP Service FEC 128 Bindings
===============================================================================
Type                                          VCId       SDPId     IngLbl  LMTU
Peer                                          SvcId                EgrLbl  RMTU
-------------------------------------------------------------------------------
No Matching Entries Found
===============================================================================
===============================================================================
LDP Service FEC 129 Bindings
===============================================================================
SAII                                            AGII        IngLbl    LMTU
TAII                                            Type        EgrLbl    RMTU
Peer                                            SvcId       SDPId
-------------------------------------------------------------------------------
192.0.2.1                                       null        262135U   1500
192.0.2.2                                       V-Eth       262133S   1500
192.0.2.2:0                                     3           17406

192.0.2.1                                       null        262136U   1500
```

```
192.0.2.3                                       V-Eth       262136S   1500
192.0.2.3:0                                     3           17407

-------------------------------------------------------------------------------
No. of FEC 129s: 2
===============================================================================
*A:PE-1#
```

This shows the two T-LDP bindings for PE-1 toward PE-2 and PE-3 for VPLS 3. The label bindings from this LDP output is identical to the SDP bindings output that follows. The following command can be used to list the SDP IDs and the SDP label bindings:

```
*A:PE-1# show service id 3 sdp

===============================================================================
Services: Service Destination Points
===============================================================================
SdpId            Type     Far End addr    Adm     Opr       I.Lbl     E.Lbl
-------------------------------------------------------------------------------
17406:4294967294 BgpAd    192.0.2.2       Up      Up        262135    262133
17407:4294967295 BgpAd    192.0.2.3       Up      Up        262136    262136
-------------------------------------------------------------------------------
Number of SDPs : 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

The SDP ID for the auto-provisioned SDP toward PE-2 is 17406, the SDP ID toward PE-3 is 17407.

The actual AGI, SAII, and TAII values are seen in the following detailed SDP output.

- AGI — 65536:3
- SAII — Local system IP address 192.0.2.1
- TAII — Remote system IP address 192.0.2.2 or 192.0.2.3

```
*A:PE-1# show service id 3 sdp 17407:4294967295 detail

===============================================================================
Service Destination Point (Sdp Id : 17407:4294967295) Details
===============================================================================
-------------------------------------------------------------------------------
 Sdp Id 17407:4294967295  -(192.0.2.3)
-------------------------------------------------------------------------------
Description     : (Not Specified)
SDP Id             : 17407:4294967295         Type              : BgpAd
PW-Template Id     : 2
AGI                : 65536:3                  SDP Bind Source   : bgp-l2vpn
Local AII          : 192.0.2.1
Remote AII         : 192.0.2.3
Split Horiz Grp    : vpls-shg
Etree Root Leaf Tag: Disabled                 Etree Leaf AC     : Disabled
VC Type            : Ether                    VC Tag            : n/a
```

```
Admin Path MTU     : 0                        Oper Path MTU     : 1556
Delivery           : MPLS
Far End            : 192.0.2.3
---snip---
```

PE-2 Service Operation Verification

For completeness, verify the service is operationally up on PE-2.

```
*A:PE-2# show service id 3 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 3                     Vpn Id            : 0
Service Type      : VPLS
---snip---
Admin State       : Up                    Oper State        : Up
MTU               : 1514                  Def. Mesh VC Id   : 3
SAP Count         : 1                     SDP Bind Count    : 2
---snip---
-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/4:3.0                            qinq         1522    1522    Up   Up
sdp:17406:4294967294 SB(192.0.2.1)       BgpAd        0       1556    Up   Up
sdp:17407:4294967295 SB(192.0.2.3)       BgpAd        0       1556    Up   Up
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-2#
```

```
*A:PE-2# show service l2-route-table bgp-ad

========================================================================
Services: L2 Route Information - Summary
========================================================================
Svc Id     L2-Routes (RD-Prefix)                 Next Hop        Origin
               Sdp Bind Id                           PW Temp Id
------------------------------------------------------------------------
3          *65536:3-192.0.2.1                    192.0.2.1       BGP-L2
               17406:4294967294                      2
3          *65536:3-192.0.2.3                    192.0.2.3       BGP-L2
               17407:4294967295                      2
------------------------------------------------------------------------
No. of L2 Route Entries: 2
========================================================================
*A:PE-2#
```

```
*A:PE-2# show router ldp bindings services service-id 3

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        S - Status Signaled Up,  D - Status Signaled Down, e - Label ELC
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
Service Type:
        E - Epipe Service, V - VPLS Service, M - Mirror Service
        A - Apipe Service, F - Fpipe Service, I - IES Service, R - VPRN service
        P - Ipipe Service, C - Cpipe Service
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
===============================================================================
LDP Service FEC 128 Bindings
===============================================================================
Type                                          VCId       SDPId      IngLbl  LMTU
Peer                                          SvcId                 EgrLbl  RMTU
-------------------------------------------------------------------------------
No Matching Entries Found
===============================================================================
===============================================================================
LDP Service FEC 129 Bindings
===============================================================================
SAII                                            AGII        IngLbl    LMTU
TAII                                            Type        EgrLbl    RMTU
Peer                                            SvcId       SDPId
-------------------------------------------------------------------------------
192.0.2.2                                       null        262133U   1500
192.0.2.1                                       V-Eth       262135S   1500
192.0.2.1:0                                     3           17406

192.0.2.2                                       null        262134U   1500
192.0.2.3                                       V-Eth       262135S   1500
192.0.2.3:0                                     3           17407

-------------------------------------------------------------------------------
No. of FEC 129s: 2
===============================================================================
*A:PE-2#

*A:PE-2# show service id 3 sdp

===============================================================================
Services: Service Destination Points
===============================================================================
SdpId            Type     Far End addr    Adm     Opr       I.Lbl     E.Lbl
-------------------------------------------------------------------------------
17406:4294967294 BgpAd    192.0.2.1       Up      Up        262133    262135
17407:4294967295 BgpAd    192.0.2.3       Up      Up        262134    262135
-------------------------------------------------------------------------------
Number of SDPs : 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-2#
```

PE-3 Service Operation Verification

Verify service is operationally up on PE-3.

```
*A:PE-3# show service id 3 base
```

```
===============================================================================
Service Basic Information
===============================================================================
Service Id        : 3                   Vpn Id            : 0
Service Type      : VPLS
---snip---
Admin State       : Up                  Oper State        : Up
MTU               : 1514                Def. Mesh VC Id   : 3
SAP Count         : 1                   SDP Bind Count    : 2
---snip---
-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                              Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/4:3.0                           qinq         1522    1522    Up   Up
sdp:17406:4294967294 SB(192.0.2.2)      BgpAd        0       1556    Up   Up
sdp:17407:4294967295 SB(192.0.2.1)      BgpAd        0       1556    Up   Up
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-3#

*A:PE-3# show service l2-route-table bgp-ad

========================================================================
Services: L2 Route Information - Summary
========================================================================
Svc Id     L2-Routes (RD-Prefix)                Next Hop        Origin
               Sdp Bind Id                          PW Temp Id
------------------------------------------------------------------------
3          *65536:3-192.0.2.1                   192.0.2.1       BGP-L2
               17407:4294967295                     2
3          *65536:3-192.0.2.2                   192.0.2.2       BGP-L2
               17406:4294967294                     2
------------------------------------------------------------------------
No. of L2 Route Entries: 2
========================================================================
*A:PE-3#

*A:PE-3# show router ldp bindings services service-id 3

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.3)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        S - Status Signaled Up,  D - Status Signaled Down, e - Label ELC
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
Service Type:
        E - Epipe Service, V - VPLS Service, M - Mirror Service
        A - Apipe Service, F - Fpipe Service, I - IES Service, R - VPRN service
        P - Ipipe Service, C - Cpipe Service
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
===============================================================================
LDP Service FEC 128 Bindings
```

```
===============================================================================
Type                                          VCId       SDPId     IngLbl  LMTU
Peer                                          SvcId                EgrLbl  RMTU
-------------------------------------------------------------------------------
No Matching Entries Found
===============================================================================
===============================================================================
LDP Service FEC 129 Bindings
===============================================================================
SAII                                            AGII        IngLbl    LMTU
TAII                                            Type        EgrLbl    RMTU
Peer                                            SvcId       SDPId
-------------------------------------------------------------------------------
192.0.2.3                                       null        262136U   1500
192.0.2.1                                       V-Eth       262136S   1500
192.0.2.1:0                                     3           17407

192.0.2.3                                       null        262135U   1500
192.0.2.2                                       V-Eth       262134S   1500
192.0.2.2:0                                     3           17406

-------------------------------------------------------------------------------
No. of FEC 129s: 2
===============================================================================
*A:PE-3#

*A:PE-3# show service id 3 sdp

===============================================================================
Services: Service Destination Points
===============================================================================
SdpId            Type     Far End addr    Adm     Opr       I.Lbl     E.Lbl
-------------------------------------------------------------------------------
17406:4294967294 BgpAd    192.0.2.2       Up      Up        262135    262134
17407:4294967295 BgpAd    192.0.2.1       Up      Up        262136    262136
-------------------------------------------------------------------------------
Number of SDPs : 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-3#
```

BGP AD using Pre-Provisioned SDPs

It is possible to configure BGP-AD instances that use RSVP transport tunnels. In this case, the LSPs and SDPs must be manually created.

Figure 89 **VPLS Instance using Pre-Provisioned SDPs**

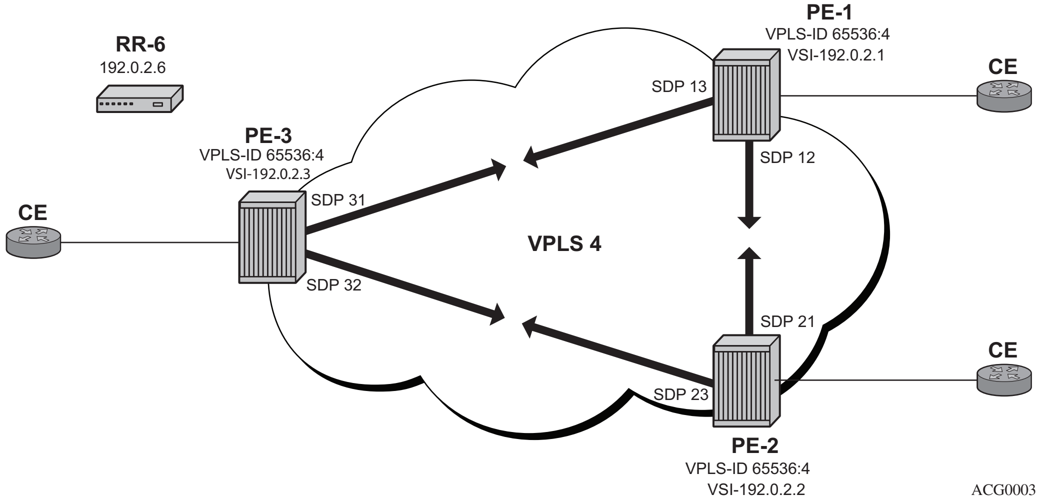


Figure 89 shows a VPLS instance configured across three Provider Edge routers as before.

The SDP configurations for the three PEs are as follows:

SDPs on PE-1

```
configure
    service
        sdp 12 mpls create
            far-end 192.0.2.2
            lsp "LSP-PE-1-PE-2"
            no shutdown
        exit
        sdp 13 mpls create
            far-end 192.0.2.3
            lsp "LSP-PE-1-PE-3"
            no shutdown
        exit
    exit
```

SDPs on PE-2

```
configure
    service
       sdp 21 mpls create
            far-end 192.0.2.1
            lsp "LSP-PE-2-PE-1"
            no shutdown
        exit
        sdp 23 mpls create
            far-end 192.0.2.3
            lsp "LSP-PE-2-PE-3"
            no shutdown
        exit
    exit
```

SDPs on PE-3

```
configure
    service
        sdp 31 mpls create
            far-end 192.0.2.1
            lsp "LSP-PE-3-PE-1"
            no shutdown
        exit
        sdp 32 mpls create
            far-end 192.0.2.2
            lsp "LSP-PE-3-PE-2"
            no shutdown
        exit
    exit
```

The pw-template that is to be used within each VPLS instance must be provisioned on all PEs and must use the keyword **use-provisioned-sdp**. The pw-template is configured on all PEs with the following command:

```
configure
    service
        pw-template 3 use-provisioned-sdp create
            exit
        exit
```

The following output shows the configuration required for a VPLS service using a pseudowire template configured for pre-provisioned RSVP SDPs.

On PE-1:

```
configure
    service
        vpls 4 customer 1 create
            bgp
                route-distinguisher 65536:4
                route-target export target:65536:4 import target:65536:4
                pw-template-binding 3 split-horizon-group "vpls-shg"
                                      import-rt "target:65536:4"
                exit
            exit
            bgp-ad
                vpls-id 65536:4
                vsi-id
                    prefix 192.0.2.1
                exit
                no shutdown
            exit
            sap 1/1/4:4.0 create
            exit
            no shutdown
        exit
```

Similarly, on PE-2 the configuration is as follows:

```
configure
    service
        vpls 4 customer 1 create
            bgp
                route-distinguisher 65536:4
                route-target export target:65536:4 import target:65536:4
                pw-template-binding 3 split-horizon-group "vpls-shg"
                                      import-rt "target:65536:4"
                exit
            exit
            bgp-ad
                vpls-id 65536:4
                vsi-id
                    prefix 192.0.2.2
                exit
                no shutdown
            exit
            sap 1/1/4:4.0 create
            exit
            no shutdown
        exit
```

On PE-3, VPLS 4 is configured as follows:

```
configure
    service
        vpls 4 customer 1 create
            bgp
                route-distinguisher 65536:4
                route-target export target:65536:4 import target:65536:4
                pw-template-binding 3 split-horizon-group "vpls-shg"
                                      import-rt "target:65536:4"
                exit
            exit
            bgp-ad
                vpls-id 65536:4
                vsi-id
                    prefix 192.0.2.3
                exit
                no shutdown
            exit
            sap 1/1/4:4.0 create
            exit
            no shutdown
        exit
```

The following output shows that the service is operationally up on PE-1.

```
*A:PE-1# show service id 4 base

===============================================================================
Service Basic Information
```

```
===============================================================================
Service Id        : 4                   Vpn Id            : 0
Service Type      : VPLS
---snip---
Admin State       : Up                  Oper State        : Up
MTU               : 1514                Def. Mesh VC Id   : 4
SAP Count         : 1                   SDP Bind Count    : 2
---snip---
-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/4:4.0                            qinq         1522    1522    Up   Up
sdp:12:4294967292 S(192.0.2.2)           BgpAd        0       1556    Up   Up
sdp:13:4294967293 S(192.0.2.3)           BgpAd        0       1556    Up   Up
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-1#
```

The SDP identifiers are the pre-provisioned SDPs: SDP 12 and 13.

The following command shows that the service is operationally up on PE-2.

```
*A:PE-2# show service id 4 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 4                   Vpn Id            : 0
Service Type      : VPLS
---snip---
Admin State       : Up                  Oper State        : Up
MTU               : 1514                Def. Mesh VC Id   : 4
SAP Count         : 1                   SDP Bind Count    : 2
---snip---
-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/4:4.0                            qinq         1522    1522    Up   Up
sdp:21:4294967292 S(192.0.2.1)           BgpAd        0       1556    Up   Up
sdp:23:4294967293 S(192.0.2.3)           BgpAd        0       1556    Up   Up
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-2#
```

The following command shows that the service is operationally up on PE-3.

```
*A:PE-3# show service id 4 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 4                   Vpn Id            : 0
Service Type      : VPLS
```

```
---snip---
Admin State       : Up                  Oper State        : Up
MTU               : 1514                Def. Mesh VC Id   : 4
SAP Count         : 1                   SDP Bind Count    : 2
---snip---
-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/4:4.0                            qinq         1522    1522    Up   Up
sdp:31:4294967292 S(192.0.2.1)           BgpAd        0       1556    Up   Up
sdp:32:4294967293 S(192.0.2.2)           BgpAd        0       1556    Up   Up
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-3#
```

Conclusion

BGP-auto-discovery coupled with LDP pseudowire signaling allows the delivery of L2 VPN services to customers where BGP is commonly used. This example shows the configuration of BGP-Auto discovery together with the associated show outputs which can be used for verification and troubleshooting.

LDP VPLS Using BGP Auto-Discovery - Prefer Provisioned SDP

This chapter provides information about LDP VPLS Using BGP Auto-Discovery - Prefer Provisioned SDP.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SR OS Release 14.0.R6, but the CLI in the current edition is based on SR OS Release 15.0.R2. BGP Auto-Discovery (BGP-AD) based on RFC 6074 is supported in SR OS Release 6.0, and later. The **prefer-provisioned-sdp** option is supported in SR OS Release 14.0.R1, and later.

Overview

As described in chapter LDP VPLS Using BGP Auto-Discovery, BGP-AD based on RFC 6074 can auto-create SDP bindings, but an operator can force the system to use a provisioned SDP by specifying the **use-provisioned-sdp** option. This chapter compares the **use-provisioned-sdp** option with the **prefer-provisioned-sdp** option. The chapter describes a migration scenario for a VPLS service with a pseudowire (PW) template binding, restricted to using provisioned SDPs toward a PW template binding preferring to use provisioned SDPs, but auto-creating SDPs in case there is no suitable manually created SDP available.

PW Templates

PW templates can be configured with the following command:

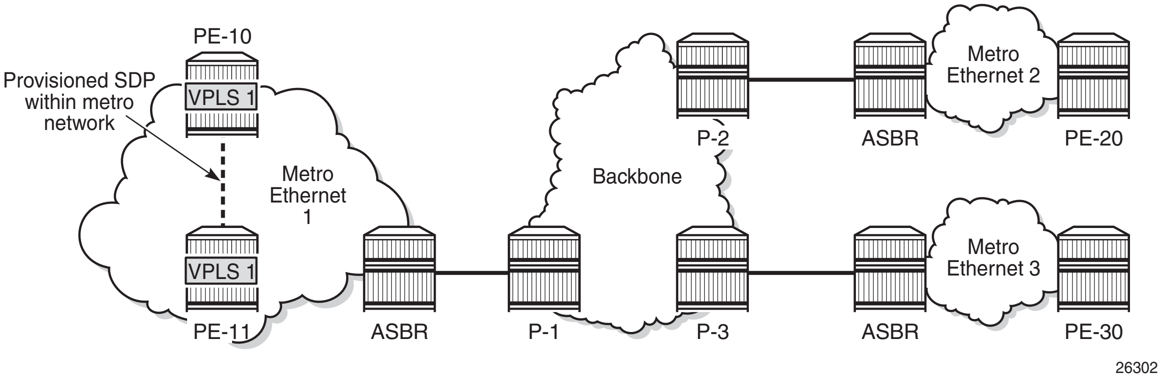
```
*A:PE-1# configure service pw-template
```

```
 - pw-template <policy-id> [create] prefer-provisioned-sdp
 - no pw-template <policy-id>
 - pw-template <policy-id> [use-provisioned-sdp] [create]

<policy-id>           : [1..2147483647]
<use-provisioned-s*> : keyword
<create>              : keyword - mandatory while creating an entry.
<prefer-provisione*> : keyword
```

- When the **use-provisioned-sdp** keyword is added at creation time, the tunnel manager is forced to look for a provisioned and active SDP to the far-end PE. The far-end PE is auto-discovered from the BGP next hop. If multiple SDPs are active to this far-end PE, the tunnel manager chooses the SDP template with the best metric. If there is a tie, the SDP ID is used as a tie-breaker and the highest SDP ID wins. However, if no provisioned SDP exists, the SDP binding will not be instantiated.
- When the **prefer-provisioned-sdp** keyword is added at creation time, the behavior is the same as when a provisioned SDP exists. When the tunnel manager finds an existing matching SDP, it will use it even if it is operationally down. Only when no provisioned SDP exists, will the SDP binding be auto-created.
- When a PW template is created without the **use-provisioned-sdp** or **prefer-provisioned-sdp** keyword, the SDP bindings will be auto-created.

Figure 90 shows the following use case: the metro Ethernet networks were initially built with provisioned SDPs. Intra-metro services are provisioned using provisioned SDPs; for example, customer X has a VPLS service defined in the metro Ethernet networks, using BGP-AD with a PW template to use the provisioned SDPs in the metro Ethernet networks.

Figure 90 **LDP VPLS Using BGP-AD with use-provisioned-sdp Option**

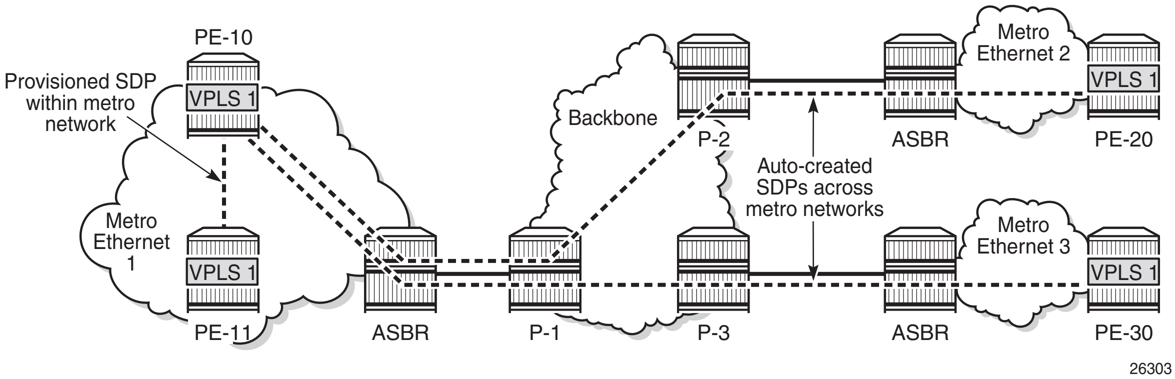


The service provider initially started with PE-10 and PE-11 in metro Ethernet 1, but now wants to add PE-20 and PE-30 as new sites to the VPLS service. Therefore, the BGP-AD routes should propagate beyond the boundaries of the metro Ethernet network. The backbone network may be in a different AS, but in this example, all networks are in the same AS. VPLS 1 of customer X can have sites added to the service on PEs in different metro Ethernet networks. A new PW template is created with the **prefer-provisioned-sdp** option and applied to the VPLS service.

- When a new site within the metro Ethernet network is added, an SDP is already provisioned to this site and this SDP is used for the SDP binding in the VPLS.
- When a new site in a different metro Ethernet network is added, no SDP is available to the site in the remote metro Ethernet network and the SDP binding is auto-created.

Figure 91 shows the SDP bindings in VPLS 1 between PE-10 and the other PEs. For simplicity, the SDP bindings between the other PEs are not shown.

Figure 91 **LDP VPLS Using BGP-AD with prefer-provisioned-sdp Option**

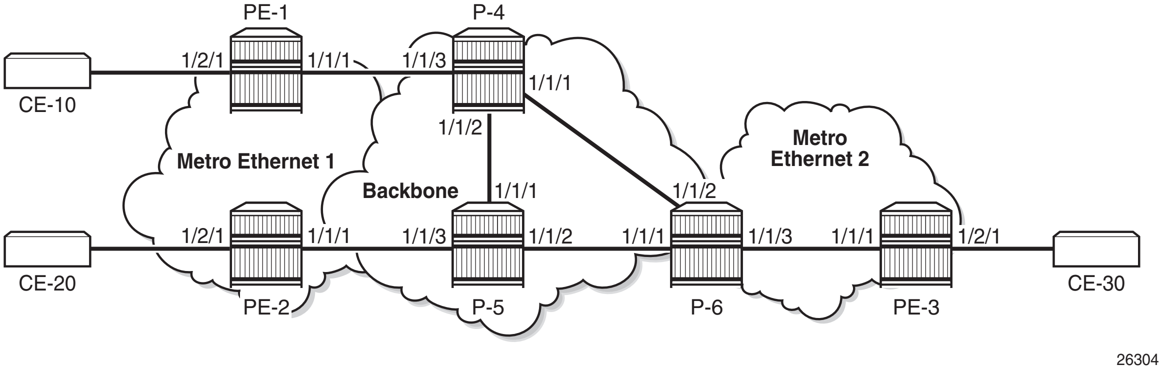


The **prefer-provisioned-sdp** and **use-provisioned-sdp** options can only be defined at creation time, implying that existing PW templates cannot be changed from prefer-provisioned-sdp to use-provisioned-sdp and vice versa. To support migration from one PW template to another with minimal service impact, two PW templates can be applied in parallel, as shown in the Configuration section.

Configuration

Figure 92 shows the example topology. For simplicity, all nodes are in the same AS.

Figure 92 **Example Topology**



The initial configuration includes the following:

- Cards, MDAs, ports
- Router interfaces
- IS-IS as IGP (or OSPF) on all interfaces
- MPLS and RSVP on all interfaces, except "int-P-4-P-6" and "int-P-5-P-6".
- LDP on all interfaces

BGP is configured on all PE routers for address family l2-vpn, as follows:

```
configure
    router
        autonomous-system 64496
        bgp
            group "internal"
                family l2-vpn
                peer-as 64496
                neighbor 192.0.2.6
                exit
            exit
```

The BGP configuration on the route reflector (RR) P-6 is as follows:

```
configure
    router
        autonomous-system 64496
        bgp
            group "rr-internal"
                cluster 1.1.1.1
                family l2-vpn
                peer-as 64496
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
            exit
```

On PE-1 and PE-2 in metro Ethernet network 1, an RSVP LSP is created that is used in a manually created SDP. The LSP configuration on PE-1 is as follows:

```
configure
    router
        mpls
            path "loose"
                no shutdown
            exit
            lsp "LSP-PE-1-PE-2"
                to 192.0.2.2
                primary "loose"
                exit
                no shutdown
            exit
            no shutdown
```

On PE-1, SDP 12 is configured as follows:

```
configure
    service
        sdp 12 mpls create
            description "SDP12 to 192.0.2.2"
            far-end 192.0.2.2
            lsp "LSP-PE-1-PE-2"
            no shutdown
        exit
```

The configuration on PE-2 is similar.

LDP VPLS Using AD without prefer-provisioned-sdp Option

Initially, the following two PW templates are created on all PEs: PW template 1 has the **use-provisioned-sdp** option and PW template 2 is created without any option; therefore, SDP bindings will be auto-created.

```
configure
    service
        pw-template 1 use-provisioned-sdp create
        exit
        pw-template 2 create
        exit
```

The following lists the PW templates configured on PE-1:

```
*A:PE-1# show service pw-template
```

```
===============================================================================
PW Template information
===============================================================================
PW Template Id      SDP                           Last Update
-------------------------------------------------------------------------------
1                   Use-provisioned               05/02/2017 13:47:04
2                   Auto-create                   05/02/2017 13:33:24
===============================================================================
*A:PE-1#
```

On all PEs, two VPLS services are created: VPLS 1 with BGP-AD PW template 1 and VPLS 2 with PW template 2, as follows:

```
configure
    service
        vpls 1 customer 1 create
            bgp
                route-distinguisher 64496:1
                route-target export target:64496:1 import target:64496:1
                pw-template-binding 1
                exit
            exit
            bgp-ad
                vpls-id 64496:1
                no shutdown
            exit
            sap 1/2/1:1 create
            exit
            no shutdown
        exit
        vpls 2 customer 1 create
            bgp
                route-distinguisher 64496:2
                route-target export target:64496:2 import target:64496:2
                pw-template-binding 2 import-rt "target:64496:2"
                exit
            exit
            bgp-ad
                vpls-id 64496:2
                no shutdown
            exit
            sap 1/2/1:2 create
            exit
            no shutdown
        exit
```

On PE-1, the following SDP bindings have been created:

```
*A:PE-1# show service sdp-using

===============================================================================
SDP Using
===============================================================================
SvcId      SdpId              Type   Far End              Opr   I.Label E.Label
                                                          State
-------------------------------------------------------------------------------
```

```
1          12:4294967295      BgpAd  192.0.2.2           Up    262136  262136
2          17406:4294967293   BgpAd  192.0.2.3           Up    262134  262136
2          17407:4294967294   BgpAd  192.0.2.2           Up    262135  262135
-------------------------------------------------------------------------------
Number of SDPs : 3
```

The first SDP binding is created by BGP-AD in VPLS 1 and uses the configured SDP 12 with far-end PE-2; the other two SDP bindings have been auto-created by BGP-AD in VPLS 2 and have far-end PE-2 and PE-3.

The list of SDP bindings on PE-2 looks similar:

```
*A:PE-2# show service sdp-using

===============================================================================
SDP Using
===============================================================================
SvcId      SdpId              Type   Far End             Opr   I.Label E.Label
                                                         State
-------------------------------------------------------------------------------
1          21:4294967295      BgpAd  192.0.2.1           Up    262136  262136
2          17406:4294967293   BgpAd  192.0.2.3           Up    262134  262137
2          17407:4294967294   BgpAd  192.0.2.1           Up    262135  262135
-------------------------------------------------------------------------------
Number of SDPs : 3
```

On PE-3, there are only two SDP bindings, both in VPLS 2:

```
*A:PE-3# show service sdp-using

===============================================================================
SDP Using
===============================================================================
SvcId      SdpId              Type   Far End             Opr   I.Label E.Label
                                                         State
-------------------------------------------------------------------------------
2          17406:4294967294   BgpAd  192.0.2.2           Up    262137  262134
2          17407:4294967295   BgpAd  192.0.2.1           Up    262136  262134
-------------------------------------------------------------------------------
Number of SDPs : 2
```

Log 99 on PE-3 shows that the system failed to create a dynamic BGP-L2VPN SDP binding because no provisioned SDP was found, as follows:

```
59 2017/01/10 13:46:16.59 UTC MAJOR: SVCMGR #2322 Base
"The system failed to create a dynamic bgp-l2vpn SDP Bind  in service 1 with SDP pw-
template policy 1 for the following reason: suitable manual SDP not found."
```

Figure 93 shows the SDPs used in VPLS 1. PE-1 and PE-2 both used the provisioned SDP. PE-3 has no SDP bindings in VPLS 1.

Figure 93 **SDP Bindings in VPLS 1 with use-provisioned-sdp Option**

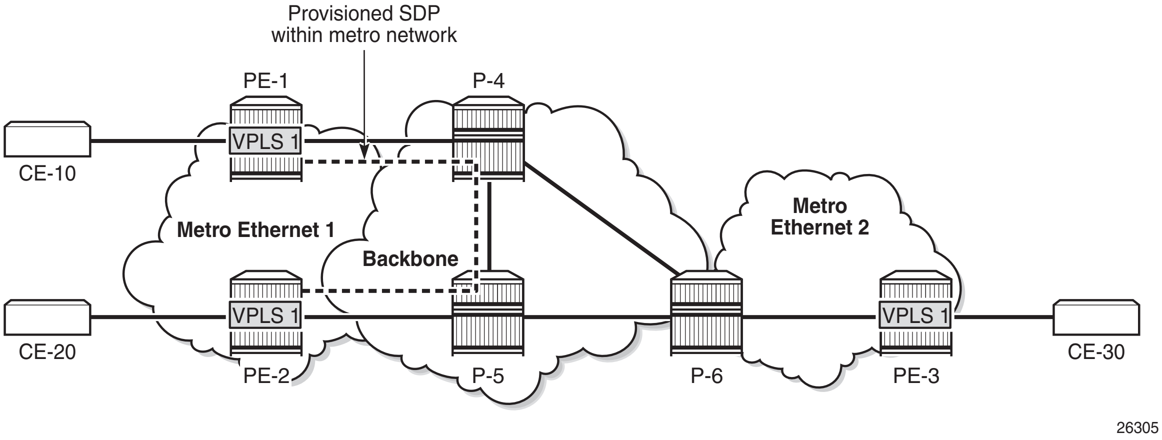
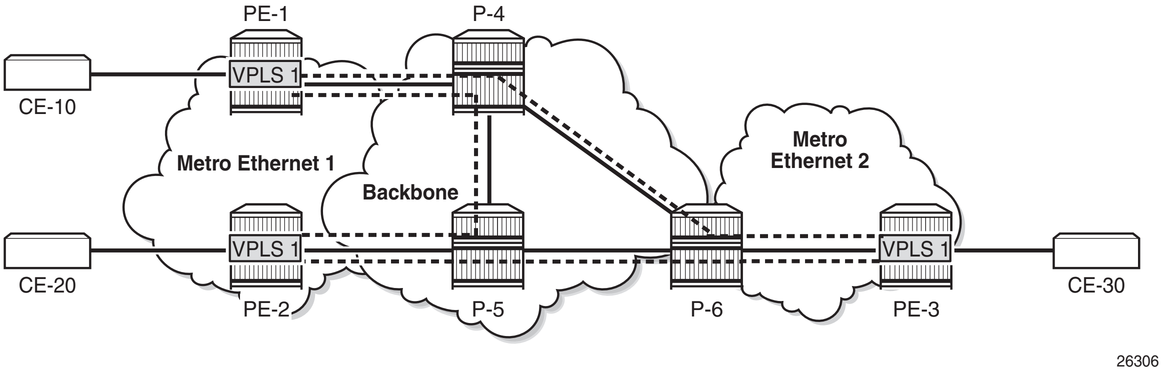


Figure 94 shows the auto-created SDP bindings in VPLS 2. Each PE has two auto-created SDP bindings to each other PE.

Figure 94 **Auto-Created SDP Bindings in VPLS 2**



Migrate VPLS 1 to prefer-provisioned-sdp Option

VPLS 1 uses PW template 1 with the **use-provisioned-sdp** option. This option is defined at creation time and cannot be modified afterward, as follows:

```
*A:PE-1# configure service pw-template 1 prefer-provisioned-sdp
MINOR: CLI The prefer-provisioned-sdp option cannot be modified after creation.
```

The following steps are needed to migrate to another PW template with the **prefer-provisioned-sdp** option without service outage:

Step 1. Create new PW template with **prefer-provisioned-sdp** option.

Step 2. Add new PW template binding to VPLS and verify which PW template is used.

Step 3. Modify old PW template binding to make it not usable.

Step 4. Launch tools command to re-evaluate old PW template in the VPLS.

Step 5. When the old PW template is not used anymore, remove PW template binding from the VPLS configuration.

A new PW template with the prefer-provisioned-sdp option is configured on all PEs, as follows:

```
configure service pw-template 10 prefer-provisioned-sdp create
```

An additional PW template binding is configured in VPLS 1 on all PEs, as follows:

```
configure service vpls 1 bgp pw-template-binding 10
```

The configuration of VPLS 1 includes two PW template bindings, as follows:

```
*A:PE-1>config>service>vpls# info
----------------------------------------------
            bgp
                route-distinguisher 64496:1
                route-target export target:64496:1 import target:64496:1
                pw-template-binding 1
                exit
                pw-template-binding 10
                exit
            exit
            bgp-ad
                vpls-id 64496:1
                no shutdown
            exit
            stp
                shutdown
            exit
            sap 1/2/1:1 create
                no shutdown
            exit
            no shutdown
```

The following shows that no additional SDP bindings have been created. The only SDP binding in VPLS 1 on PE-1 uses the provisioned SDP 12.

```
*A:PE-1# show service id 1 sdp

===============================================================================
Services: Service Destination Points
```

```
===============================================================================
SdpId             Type     Far End addr    Adm     Opr       I.Lbl     E.Lbl
-------------------------------------------------------------------------------
12:4294967295     BgpAd    192.0.2.2       Up      Up        262136    262136
-------------------------------------------------------------------------------
Number of SDPs : 1
-------------------------------------------------------------------------------
===============================================================================
```

The following shows that PW template 1 was used for the creation of the SDP binding:

```
*A:PE-1# show service id 1 sdp detail | match "SDP Id|PW-Template Id" expression
SDP Id              : 12:4294967295            Type              : BgpAd
PW-Template Id      : 1
```

The PW template 10 has a higher ID than PW template 1 and is not used. Re-evaluating the PW template binding for PW template 1 in VPLS 1 will make no difference if both PW templates are usable. However, PW template 1 can be made unusable by adding a dummy **import-rt** not matching any route in the VPLS, as follows:

```
configure service vpls 1 bgp pw-template-binding 1 import-rt "target:111:111"
```

As a result, PW template 10 with the **prefer-provisioned-sdp** option is used for the automatic creation of SDP bindings where no provisioned SDP is available, as follows:

```
*A:PE-1# show service id 1 sdp

===============================================================================
Services: Service Destination Points
===============================================================================
SdpId             Type     Far End addr    Adm     Opr       I.Lbl     E.Lbl
-------------------------------------------------------------------------------
12:4294967291     BgpAd    192.0.2.2       Up      Up        262136    262136
17406:4294967292 BgpAd    192.0.2.3       Up      Up        262133    262135
-------------------------------------------------------------------------------
Number of SDPs : 2
```

For the first SDP binding, PW template 1 is used, and for the second SDP binding, PW template 10 is used, as follows:

```
*A:PE-1# show service id 1 sdp detail | match "SDP Id|PW-Template Id" expression
SDP Id              : 12:4294967295            Type              : BgpAd
PW-Template Id      : 1
SDP Id              : 17406:4294967292         Type              : BgpAd
PW-Template Id      : 10
```

The following command forces the system to re-evaluate PW template 1 in VPLS 1:

```
*A:PE-1# tools perform service id 1 eval-pw-template 1 allow-service-impact
```

```
eval-pw-template succeeded for Svc 1 12:4294967295 Policy 1
```

As a result, only PW template 10 is used for the creation of SDP bindings in VPLS 1, as follows:

```
*A:PE-1# show service id 1 sdp detail | match "SDP Id|PW-Template Id" expression
SDP Id              : 12:4294967291            Type              : BgpAd
PW-Template Id      : 10
SDP Id              : 17406:4294967292         Type              : BgpAd
PW-Template Id      : 10
```

PW template 1 is not used anymore and can be removed from the VPLS configuration, as follows:

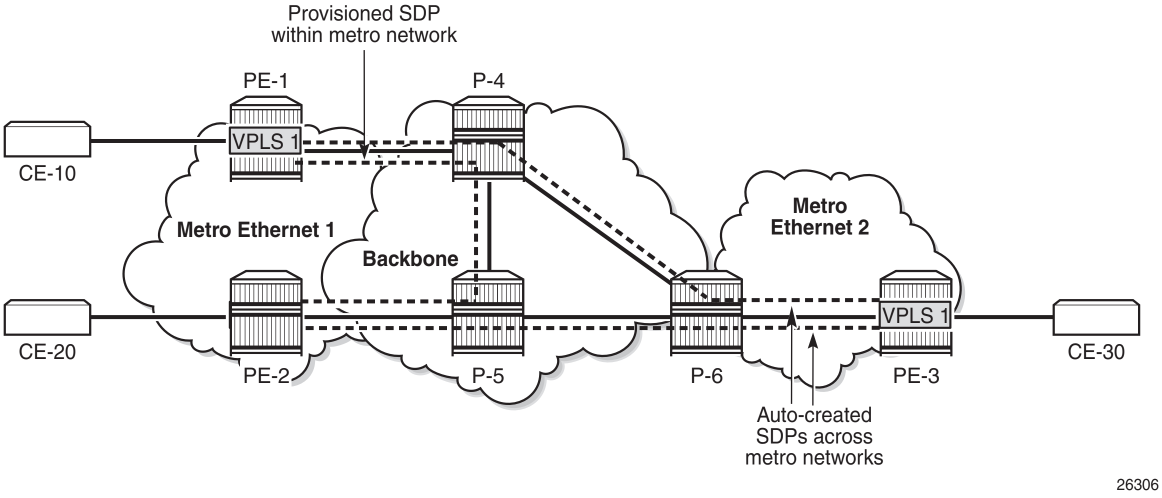
```
configure service vpls 1 bgp no pw-template-binding 1
```

The configuration of VPLS 1 on PE-1 contains only a PW template binding for PW template 10, as follows:

```
*A:PE-1>config>service>vpls# info
----------------------------------------------
            bgp
                route-distinguisher 64496:1
                route-target export target:64496:1 import target:64496:1
                pw-template-binding 10
                exit
            exit
            bgp-ad
                vpls-id 64496:1
                no shutdown
            exit
            stp
                shutdown
            exit
            sap 1/2/1:1 create
                no shutdown
            exit
            no shutdown
```

Figure 95 shows the SDP bindings in VPLS 1 with the **prefer-provisioned-sdp** option. Within metro Ethernet network 1, the provisioned SDP is used, and between metro Ethernet networks, auto-created SDP bindings are used.

Figure 95 **SDP Bindings in VPLS 1 with prefer-provisioned-sdp Option**



Conclusion

LDP VPLS using BGP-AD allows the creation of SDP bindings that are either auto-created or that use provisioned SDPs. When the **prefer-provisioned-sdp** option is used, the tunnel manager will look for a provisioned and active SDP to the far end and use it, if available, even if it is down. When no provisioned SDP is available, the system will auto-create an SDP binding.

Multi-Chassis Endpoint for VPLS Active/Standby Pseudowire

This chapter provides information about multi-chassis endpoint for VPLS active/standby pseudowire.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SROS release 7.0.R6, but the CLI in this edition is based on release 15.0.R2.

Overview

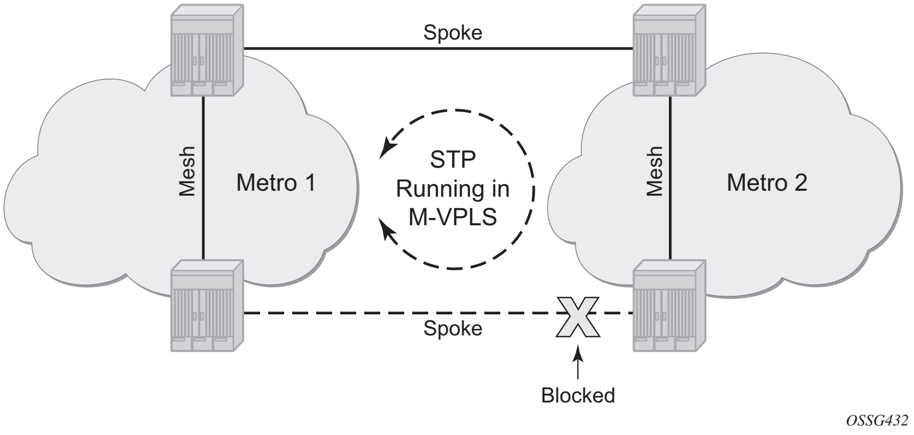
When implementing a large VPLS, one of the limiting factors is the number of T-LDP sessions required for the full mesh of SDPs. Mesh SDPs are required between all PEs participating in the VPLS with a full mesh of T-LDP sessions.

This solution is not scalable, because the number of sessions grows more rapidly than the number of participating PEs. Several options exist to reduce the number of T-LDP sessions required in a large VPLS.

The first option is hierarchical VPLS (H-VPLS) with spoke SDPs. By using spoke SDPs between two clouds of fully meshed PEs, any-to-any T-LDP sessions for all participating PEs are not required.

However, if spoke SDP redundancy is required, STP must be used to avoid a loop in the VPLS. Management VPLS can be used to reduce the number of STP instances and separate customer and STP traffic (Figure 96).

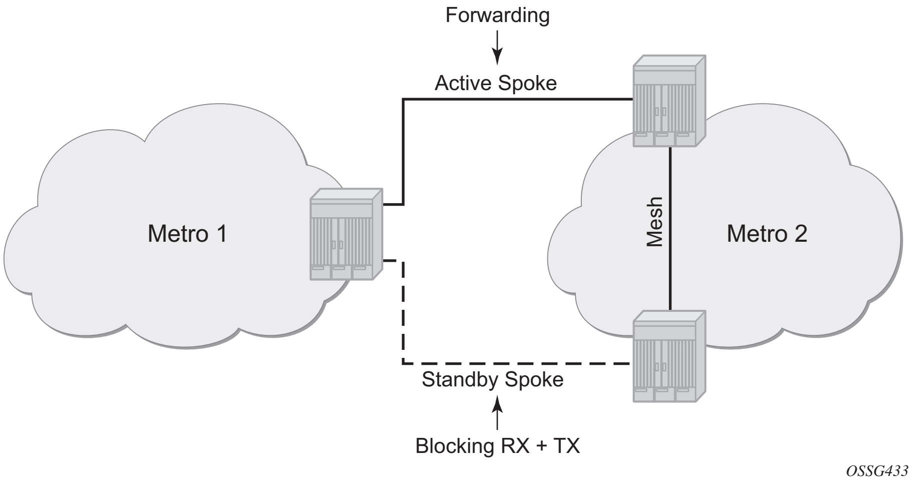
Figure 96 **H-VPLS with STP**



VPLS pseudowire redundancy provides H-VPLS redundant spoke connectivity. The active spoke is in forwarding state, while the standby spoke is in blocking state. Therefore, STP is not needed anymore to break the loop, as illustrated in Figure 97.

However, the PE implementing the active and standby spokes represents a single point of failure in the network.

Figure 97 **VPLS Pseudowire Redundancy**

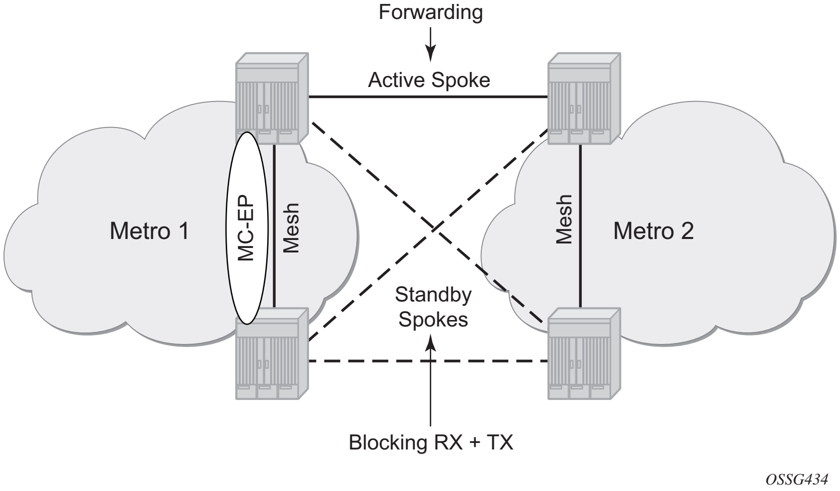


Multi-chassis endpoint (MC-EP) for VPLS active/standby pseudowire expands on the VPLS pseudowire redundancy and allows the removal of the single point of failure.

Only one spoke SDP is in forwarding state; all standby spoke SDPs are in blocking state. Mesh and square resiliency are supported.

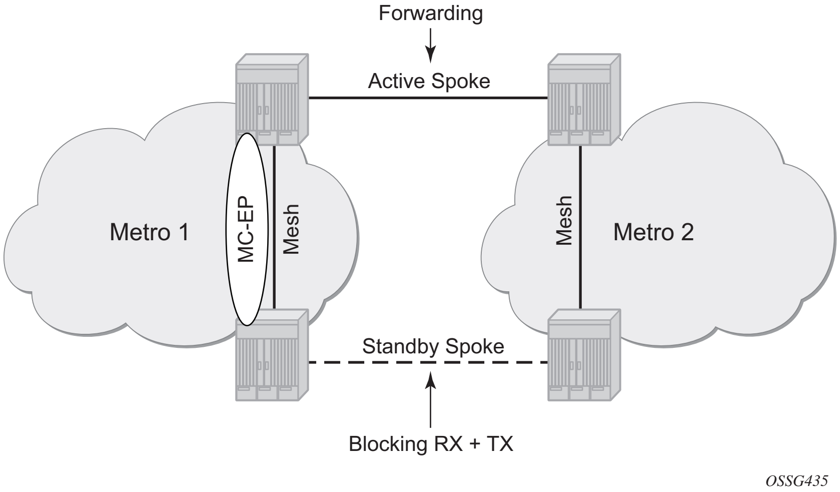
Mesh resiliency can protect against simultaneous node failure in the core and in the MC-EP (double failure), but requires more SDPs (and therefore more T-LDP sessions). Mesh resiliency is illustrated in Figure 98.

Figure 98 **Multi-Chassis Endpoint with Mesh Resiliency**



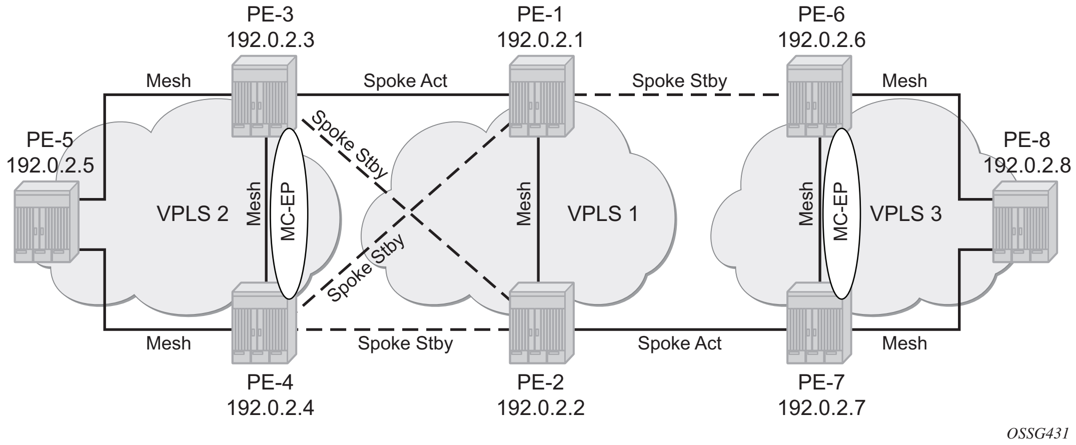
Square resiliency provides single failure node protection, and requires less SDPs (and thus less T-LDP sessions). Square resiliency is illustrated in Figure 99.

Figure 99 **Multi-Chassis Endpoint with Square Resiliency**



Example Topology

Figure 100 **Example Topology**



The network topology is displayed in Figure 100.

The setup consists of:

- Two core nodes (PE-1 and PE-2), and three nodes for each metro area (PE-3, PE-4, PE-5 and PE-6, PE-7, PE-8, respectively).
- VPLS 1 is the core VPLS, used to interconnect the two metro areas represented by VPLS 2 and VPLS 3.
- VPLS 2 will be connected to the core VPLS in mesh resiliency.
- VPLS 3 will be connected to the core VPLS in square resiliency.

Three separate VPLS identifiers are used for clarity. However, the same identifier could be used for each. For interoperation, only the same VC-ID is required to be used on both ends of the spoke SDPs.

The following configuration tasks should be done first:

- IS-IS or OSPF throughout the network.
- RSVP or LDP-signaled LSPs over the paths used for mesh/spoke SDPs.

Configuration

SDP Configuration

On each PE, SDPs are created to match the topology described in Figure 100.

The convention for the SDP naming is: XY where X is the originating node and Y the target node.

An example of the SDP configuration in PE-3 (using LDP):

```
A:PE-3# configure
    service
        sdp 31 mpls create
            far-end 192.0.2.1
            ldp
            no shutdown
        exit
        sdp 32 mpls create
            far-end 192.0.2.2
            ldp
            no shutdown
        exit
        sdp 34 mpls create
            far-end 192.0.2.4
            ldp
            no shutdown
        exit
        sdp 35 mpls create
            far-end 192.0.2.5
            ldp
            no shutdown
        exit
```

Verification of the SDPs on PE-3:

```
*A:PE-3# show service sdp

=============================================================================
Services: Service Destination Points
=============================================================================
SdpId  AdmMTU  OprMTU  Far End          Adm  Opr         Del     LSP   Sig
-----------------------------------------------------------------------------
31     0       1556    192.0.2.1        Up   Up          MPLS    L     TLDP
32     0       1556    192.0.2.2        Up   Up          MPLS    L     TLDP
34     0       1556    192.0.2.4        Up   Up          MPLS    L     TLDP
35     0       1556    192.0.2.5        Up   Up          MPLS    L     TLDP
-----------------------------------------------------------------------------
Number of SDPs : 4
-----------------------------------------------------------------------------
Legend: R = RSVP, L = LDP, B = BGP, M = MPLS-TP, n/a = Not Applicable
        I = SR-ISIS, O = SR-OSPF, T = SR-TE, F = FPE
```

```
===============================================================================
*A:PE-3#
```

Full Mesh VPLS Configuration

Next, three fully meshed VPLS services are configured.

- VPLS 1 is the core VPLS, on PE-1 and PE-2
- VPLS 2 is the metro 1 VPLS, on PE-3, PE-4 and PE-5
- VPLS 3 is the metro 2 VPLS, on PE-6, PE-7 and PE-8

On PE-1 (similar configuration on PE-2):

```
configure
    service
        vpls 1 customer 1 create
            description "core VPLS"
            mesh-sdp 12:1 create
            exit
            no shutdown
        exit
```

On PE-3 (similar configuration on PE-4 and PE-5):

```
configure
    service
        vpls 2 customer 1 create
            description "Metro 1 VPLS"
            mesh-sdp 34:2 create
            exit
            mesh-sdp 35:2 create
            exit
            no shutdown
        exit
```

On PE-6 (similar configuration on PE-7 and PE-8):

```
configure
    service
        vpls 3 customer 1 create
            description "Metro 2 VPLS"
            mesh-sdp 67:3 create
            exit
            mesh-sdp 68:3 create
            exit
            no shutdown
        exit
```

Verification of the VPLS:

- The service must be operationally up.
- All mesh SDPs must be up in the VPLS service.

On PE-6 (similar on other nodes):

```
*A:PE-6# show service id 3 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 3                   Vpn Id            : 0
Service Type      : VPLS
Name              : (Not Specified)
Description       : (Not Specified)
Customer Id       : 1                   Creation Origin   : manual
Last Status Change: 04/24/2017 08:08:29
Last Mgmt Change  : 04/24/2017 08:08:24
Etree Mode        : Disabled
Admin State       : Up                  Oper State        : Up
MTU               : 1514                Def. Mesh VC Id   : 3
SAP Count         : 0                   SDP Bind Count    : 2
Snd Flush on Fail : Disabled            Host Conn Verify  : Disabled
SHCV pol IPv4     : None
Propagate MacFlush: Disabled            Per Svc Hashing   : Disabled
Allow IP Intf Bind: Disabled
Fwd-IPv4-Mcast-To*: Disabled            Fwd-IPv6-Mcast-To*: Disabled
Def. Gateway IP   : None
Def. Gateway MAC  : None
Temp Flood Time   : Disabled            Temp Flood        : Inactive
Temp Flood Chg Cnt: 0
SPI load-balance  : Disabled
TEID load-balance : Disabled
Src Tep IP        : N/A
VSD Domain        : <none>

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sdp:67:3 M(192.0.2.7)                    Mesh         0       1556    Up   Up
sdp:68:3 M(192.0.2.8)                    Mesh         0       1556    Up   Up
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-6#
```

Multi-Chassis Configuration

Multi-chassis will be configured on the MC peers PE-3, PE-4 and PE-6, PE-7. The peer system address is configured, and **mc-endpoint** will be enabled.

On PE-3 (similar configuration on PE-4, PE-6 and PE-7):

```
configure
    redundancy
        multi-chassis
            peer 192.0.2.4 create
                mc-endpoint
                    no shutdown
                exit
                no shutdown
            exit
```

Verification of the multi-chassis synchronization:

If the multi-chassis synchronization fails, both nodes will fall back to single-chassis mode. In that case, two spoke SDPs could become active at the same time. It is important to verify the multi-chassis synchronization before enabling the redundant spoke SDPs.

```
*A:PE-3# show redundancy multi-chassis mc-endpoint peer 192.0.2.4

===============================================================================
Multi-Chassis MC-Endpoint
===============================================================================
Peer Addr       : 192.0.2.4           Peer Name            :
Admin State     : up                  Oper State           : up
Last State chg  :                     Source Addr          :
System Id       : 16:0b:ff:00:00:00   Sys Priority         : 0
Keep Alive Intvl: 10                  Hold on Nbr Fail     : 3
Passive Mode    : disabled            Psv Mode Oper        : No
Boot Timer      : 300                 BFD                  : disabled
Last update     : 04/24/2017 08:08:44 MC-EP Count          : 0
===============================================================================
*A:PE-3#
```

Mesh Resiliency Configuration

PE-3 and PE-4 will be connected to the core VPLS in mesh resiliency.

- First an endpoint is configured.
- The “no suppress-standby-signaling” is needed to block the standby spoke SDP.
- The multi-chassis endpoint peer is configured. The mc-endpoint ID must match between the two peers.

On PE-3 (similar on PE-4):

```
configure
    service
        vpls 2
            endpoint "CORE" create
            no suppress-standby-signaling
            mc-endpoint 1
                mc-ep-peer 192.0.2.4
            exit
        exit
```

After this configuration, the MP-EP Count in the preceding show command changes to 1, as follows:

```
*A:PE-3# show redundancy multi-chassis mc-endpoint peer 192.0.2.4

===============================================================================
Multi-Chassis MC-Endpoint
===============================================================================
Peer Addr       : 192.0.2.4           Peer Name            :
Admin State     : up                  Oper State           : up
Last State chg  :                     Source Addr          :
System Id       : 16:0b:ff:00:00:00   Sys Priority         : 0
Keep Alive Intvl: 10                  Hold on Nbr Fail     : 3
Passive Mode    : disabled            Psv Mode Oper        : No
Boot Timer      : 300                 BFD                  : disabled
Last update     : 04/24/2017 08:10:07 MC-EP Count          : 1
===============================================================================
*A:PE-3#
```

Two spoke SDPs are configured on each peer of the multi-chassis to the two nodes of the core VPLS (mesh resiliency). Each spoke SDP refers to the endpoint CORE.

The precedence is defined on the spoke SDPs as follows:

- Spoke SDP 31 on PE-3 will be active. It is configured as primary (= precedence 0).
- Spoke SDP 32 on PE-3 will be the first backup. It is configured with precedence 1.
- Spoke SDP 41 on PE-4 will be the second backup. It is configured with precedence 2.
- Spoke SDP 42 on PE-4 will be the third backup. It is configured with precedence 3.

On PE-3:

```
configure
    service
        vpls 2
            spoke-sdp 31:1 endpoint "CORE" create
                precedence primary
            exit
            spoke-sdp 32:1 endpoint "CORE" create
                precedence 1
            exit
```

On PE-4:

```
configure
```

```
    service
        vpls 2
            spoke-sdp 41:1 endpoint "CORE" create
                precedence 2
            exit
            spoke-sdp 42:1 endpoint "CORE" create
                precedence 3
            exit
```

Verification of the spoke SDPs:

On PE-3 and PE-4, the spoke SDPs must be up.

```
*A:PE-3# show service id 2 sdp

===============================================================================
Services: Service Destination Points
===============================================================================
SdpId            Type Far End addr    Adm     Opr       I.Lbl       E.Lbl
-------------------------------------------------------------------------------
31:1             Spok 192.0.2.1       Up      Up        262135      262131
32:1             Spok 192.0.2.2       Up      Up        262134      262131
34:2             Mesh 192.0.2.4       Up      Up        262133      262135
35:2             Mesh 192.0.2.5       Up      Up        262132      262135
-------------------------------------------------------------------------------
Number of SDPs : 4
-------------------------------------------------------------------------------
===============================================================================
*A:PE-3#
```

The endpoints on PE-3 and PE-4 can be verified. One spoke SDP is in Tx-Active mode (31 on PE-1 because it is configured as primary).

```
*A:PE-3# show service id 2 endpoint "CORE" | match "Tx Active"
Tx Active (SDP)              : 31:1
Tx Active Up Time            : 0d 01:16:04
Tx Active Change Count       : 1
Last Tx Active Change        : 04/24/2017 08:10:41
*A:PE-3#
```

There is no active spoke SDP on PE-4.

```
*A:PE-4# show service id 2 endpoint "CORE" | match "Tx Active"
Tx Active                    : none
Tx Active Up Time            : 0d 00:00:00
Tx Active Change Count       : 0
Last Tx Active Change        : 04/24/2017 07:59:47
*A:PE-4#
```

On PE-1 and PE-2, the spoke SDPs are operationally up.

```
*A:PE-1# show service id 1 sdp

===============================================================================
Services: Service Destination Points
```

```
===============================================================================
SdpId            Type Far End addr    Adm     Opr       I.Lbl       E.Lbl
-------------------------------------------------------------------------------
12:1             Mesh 192.0.2.2       Up      Up        262135      262135
13:1             Spok 192.0.2.3       Up      Up        262131      262135
14:1             Spok 192.0.2.4       Up      Up        262130      262133
-------------------------------------------------------------------------------
Number of SDPs : 3
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

However, because pseudowire signaling has been enabled, only one spoke SDP will be active, the others are set in standby.

On PE-1, spoke SDP 13:1 is active (no pseudowire bit signaled from peer PE-3) and the spoke SDP 14:1 is signaled in standby by peer PE-4.

```
*A:PE-1# show service id 1 sdp 13:1 detail | match "Peer Pw Bits"
Peer Pw Bits       : None
*A:PE-1# show service id 1 sdp 14:1 detail | match "Peer Pw Bits"
Peer Pw Bits       : pwFwdingStandby
*A:PE-1#
```

On PE-2, both spoke SDPs are signaled in standby by peers PE-3 and PE-4.

```
*A:PE-2# configure port 1/1/1 no shutdown
*A:PE-2# show service id 1 sdp 23:1 detail | match "Peer Pw Bits"
Peer Pw Bits       : pwFwdingStandby
*A:PE-2# show service id 1 sdp 24:1 detail | match "Peer Pw Bits"
Peer Pw Bits       : pwFwdingStandby
*A:PE-2#
```

There is one active and three standby spoke SDPs.

Square Resiliency Configuration

PE-6 and PE-7 will be connected to the core VPLS in square resiliency.

- First an endpoint is configured.
- The “no suppress-standby-signaling” is needed to block the standby spoke SDP.
- The multi-chassis endpoint peer is configured. The mc-endpoint ID must match between the two peers.

On PE-7 (similar on PE-6):

One spoke SDP is configured on each peer of the multi-chassis to one node of the core VPLS (square resiliency). Each spoke SDP refers to the endpoint CORE.

```
configure
    service
        vpls 3
            endpoint "CORE" create
                no suppress-standby-signaling
                mc-endpoint 1
                    mc-ep-peer 192.0.2.6
                exit
            exit
        exit
    exit
exit
```

The precedence will be defined on the spoke SDPs as follows:

- Spoke SDP 72:1 on PE-7 will be active. It is configured as primary (= precedence 0)
- Spoke SDP 61:1 on PE-6 will be the first backup with precedence 1.

On PE-7:

```
configure
    service
        vpls 3
            spoke-sdp 72:1 endpoint "CORE" create
                precedence primary
            exit
        exit
    exit
```

On PE-6:

```
configure
    service
        vpls 3
            spoke-sdp 61:1 endpoint "CORE" create
                precedence 1
            exit
        exit
    exit
```

Verification of the spoke SDPs.

```
*A:PE-7# show service id 3 sdp

===============================================================================
Services: Service Destination Points
===============================================================================
SdpId            Type Far End addr    Adm     Opr       I.Lbl       E.Lbl
-------------------------------------------------------------------------------
72:1             Spok 192.0.2.2       Up      Up        262133      262132
76:3             Mesh 192.0.2.6       Up      Up        262135      262135
78:3             Mesh 192.0.2.8       Up      Up        262128      262142
-------------------------------------------------------------------------------
Number of SDPs : 3
```

```
-------------------------------------------------------------------------------
===============================================================================
*A:PE-7#
```

On PE-6 and PE-7, the spoke SDPs must be up.

The endpoints on PE-7 and PE-6 can be verified. One spoke SDP is in Tx-Active mode (72 on PE-7 because it is configured as primary).

```
*A:PE-7# show service id 3 endpoint | match "Tx Active"
Tx Active (SDP)              : 72:1
Tx Active Up Time            : 0d 00:17:24
Tx Active Change Count       : 1
Last Tx Active Change        : 04/24/2017 08:13:18
*A:PE-7#
```

There are no active spoke SDP on PE-6.

```
*A:PE-6# show service id 3 endpoint | match "Tx Active"
Tx Active                    : none
Tx Active Up Time            : 0d 00:00:00
Tx Active Change Count       : 2
Last Tx Active Change        : 04/24/2017 08:13:18
*A:PE-6#
```

The output shows that on PE-1, spoke SDP 16 is signaled with peer in standby mode.

```
*A:PE-1# show service id 1 sdp 16:1 detail | match "Peer Pw Bits"
Peer Pw Bits       : pwFwdingStandby
*A:PE-1#
```

On PE-2, the spoke SDP 27 is signaled with peer active (no pseudowire bits).

```
*A:PE-2# show service id 1 sdp 27:1 detail | match "Peer Pw Bits"
Peer Pw Bits       : None
*A:PE-2#
```

There is one active and one standby spoke SDP.

Additional Parameters

Multi-Chassis

```
*A:PE-3# configure redundancy multi-chassis peer 192.0.2.4 mc-endpoint
  - mc-endpoint
  - no mc-endpoint
```

```
[no] bfd-enable      - Configure BFD
[no] boot-timer      - Configure boot timer interval
[no] hold-on-neighb* - Configure hold time applied on neighbor failure
[no] keep-alive-int* - Configure keep alive interval for this MC-Endpoint
[no] passive-mode    - Configure passive-mode
[no] shutdown        - Administratively enable/disable the multi-chassis
                       peer end-point
[no] system-priority - Configure system priority
```

These parameters will be explained in the following sections.

Peer Failure Detection

The default mechanism is based on the keep-alive messages exchanged between the peers.

The keep-alive-interval is the interval at which keep-alive messages are sent to the MC peer. It is set in tenths of a second from 5 to 500), with a default value of 5.

Hold-on-neighbor-failure is the number of keep-alive-intervals that the node will wait for a packet from the peer before assuming it has failed. After this interval, the node will revert to single chassis behavior. It can be set from 2 to 25 with a default value of 3.

BFD Session

BFD is another peer failure detection mechanism. It can be used to speed up the convergence in case of peer loss.

```
*A:PE-3# configure
    redundancy
        multi-chassis
            peer 192.0.2.4
                mc-endpoint
                    bfd-enable
                exit
            exit
```

It is using the centralized BFD session. BFD must be enabled on the system interface.

```
*A:PE-3# configure
    router
        interface "system"
            address 192.0.2.3/32
            bfd 100 receive 100 multiplier 3
        exit
```

Verification of the BFD session:

```
*A:PE-3# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
system                                               Up          175         53
  192.0.2.4                                           3          100        100
  mcep                                          central          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:PE-3#
```

Boot Timer

The **boot-timer** command specifies the time after a reboot that the node will try to establish a connection with the MC peer before assuming a peer failure. In case of failure, the node will revert to single chassis behavior.

System Priority

The system priority influences the selection of the MC master. The lowest priority node will become the master.

In case of equal priorities, the lowest system-id (=chassis MAC address) will become the master.

VPLS Endpoint and Spoke SDP

Ignore Standby Pseudowire Bits

```
*A:PE-1# configure service vpls 1 spoke-sdp 14:1
---snip---
[no] ignore-standby* - Ignore 'standby-bit' received from LDP peer
---snip---
```

The peer pseudowire status bits are ignored and traffic is forwarded over the spoke SDP.

It can speed up convergence for multicast traffic in case of spoke SDP failure.

Traffic sent over the standby spoke SDP will be discarded by the peer.

In this topology, if the **ignore-standby-signaling** command is enabled on PE-1, it sends MC traffic to PE-3 and PE-4 (and to PE-6). If PE-3 fails, PE-4 can start forwarding traffic in the VPLS as soon as it detects PE-3 being down. There is no signaling needed between PE-1 and PE-4.

Block-on-Mesh-Failure

```
*A:PE-3# configure service vpls 2 endpoint "CORE"
---snip---
[no] block-on-mesh-* - Block traffic on mesh-SDP failure
---snip---
```

In case a PE loses all the mesh SDPs of a VPLS, it should block the spoke SDPs to the core VPLS, and inform the MC-EP peer that can activate one of its spoke SDPs.

If block-on-mesh-failure is enabled, the PE will signal all the pseudowires of the endpoint in standby.

In this topology, if PE3 does not have any valid mesh SDP to the VPLS 2 mesh, it will set the spoke SDPs under endpoint CORE in standby.

When block-on-mesh-failure is activated under an endpoint, it is automatically set under the spoke SDPs belonging to this endpoint.

```
*A:PE-3# configure service vpls 2
*A:PE-3>config>service>vpls# info
----------------------------------------------
            description "Metro 1 VPLS"
            stp
                shutdown
            exit
            endpoint "CORE" create
                no suppress-standby-signaling
                mc-endpoint 1
                    mc-ep-peer 192.0.2.4
                exit
            exit
            spoke-sdp 31:1 endpoint "CORE" create
                stp
                    shutdown
                exit
                precedence primary
                no shutdown
            exit
            spoke-sdp 32:1 endpoint "CORE" create
                stp
                    shutdown
                exit
                precedence 1
                no shutdown
            exit
            mesh-sdp 34:2 create
                no shutdown
            exit
            mesh-sdp 35:2 create
                no shutdown
            exit
            no shutdown
----------------------------------------------
*A:PE-3>config>service>vpls# endpoint "CORE" block-on-mesh-failure
*A:PE-3>config>service>vpls# info
----------------------------------------------
            description "Metro 1 VPLS"
            stp
                shutdown
            exit
            endpoint "CORE" create
                no suppress-standby-signaling
                block-on-mesh-failure
                mc-endpoint 1
                    mc-ep-peer 192.0.2.4
                exit
            exit
            spoke-sdp 31:1 endpoint "CORE" create
                stp
                    shutdown
                exit
                block-on-mesh-failure
                precedence primary
                no shutdown
            exit
            spoke-sdp 32:1 endpoint "CORE" create
                stp
                    shutdown
                exit
                block-on-mesh-failure
                precedence 1
                no shutdown
            exit
            mesh-sdp 34:2 create
                no shutdown
            exit
            mesh-sdp 35:2 create
                no shutdown
            exit
            no shutdown
----------------------------------------------
```

Precedence

```
*A:PE-3# configure service vpls 2 spoke-sdp 31:1
---snip---
 [no] precedence      - Configure the spoke-sdp precedence
---snip---
```

The precedence is used to indicate in which order the spoke SDPs should be used. The value is from 0 to 4 (0 being primary), the lowest having higher priority. The default value is 4.

Revert-Time

```
*A:PE-3# configure service vpls 2 endpoint "CORE"
---snip---
 [no] revert-time    - Configure the time to wait before reverting to primary spoke-sdp
---snip---
```

If the precedence is equal between the spoke SDPs, there is no revertive behavior. Changing the precedence of a spoke SDP will not trigger a revert. The default is **no revert**.

MAC-Flush Parameters

When a spoke SDP goes from standby to active (due to the active spoke SDP failure), the node will send a **flush-all-but-mine** message.

After a restoration of the spoke SDP, a new **flush-all-but-mine** message will be sent.

```
*A:PE-1# configure service vpls 1 propagate-mac-flush
```

A node configured with **propagate MAC flush** will forward the flush messages received on the spoke-SDP to its other mesh/spoke SDPs.

A node configured with **send flush on failure** will send a **flush-all-from-me** message when one of its SDPs goes down.

```
A:PE-1# configure service vpls 1 send-flush-on-failure
```

Failure Scenarios

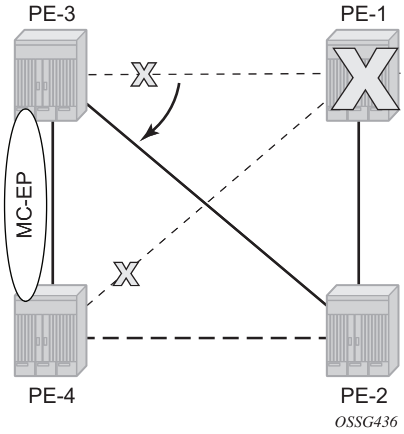
For the subsequent failure scenarios, the configuration of the nodes is as described in the Configuration.

Core Node Failure

When the core node PE-1 fails, the spoke SDPs from PE-3 and PE-4 go down.

Because the spoke SDP 31 between PE-3 and PE-4 was active, the MC master (PE-3 in this case) will select the next best spoke SDP, which will be 32 between PE-3 and PE-2 (precedence 1). See Figure 101.

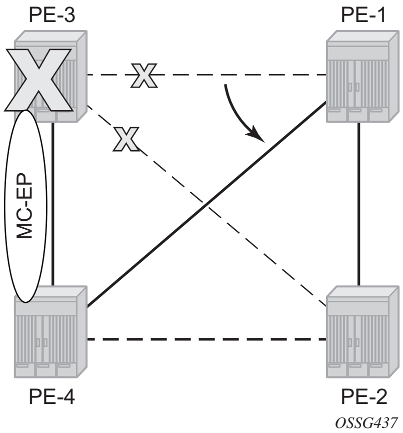
Figure 101 **Core Node Failure**



```
*A:PE-3# show service id 2 endpoint

===============================================================================
Service 2 endpoints
===============================================================================
Endpoint name                : CORE
Description                  : (Not Specified)
Creation Origin              : manual
Revert time                  : 0
Act Hold Delay               : 0
Ignore Standby Signaling     : false
Suppress Standby Signaling   : false
Block On Mesh Fail           : true
Multi-Chassis Endpoint       : 1
MC Endpoint Peer Addr        : 192.0.2.4
Psv Mode Active              : No
Tx Active (SDP)              : 32:1
Tx Active Up Time            : 0d 00:00:12
Revert Time Count Down       : N/A
Tx Active Change Count       : 1
Last Tx Active Change        : 04/24/2017 08:16:48
-------------------------------------------------------------------------------
Members
-------------------------------------------------------------------------------
Spoke-sdp: 31:1 Prec:0                                  Oper Status: Down
Spoke-sdp: 32:1 Prec:1                                  Oper Status: Up
===============================================================================
===============================================================================
*A:PE-3#

*A:PE-4# show service id 2 endpoint

===============================================================================
Service 2 endpoints
===============================================================================
Endpoint name                : CORE
Description                  : (Not Specified)
Creation Origin              : manual
Revert time                  : 0
Act Hold Delay               : 0
Ignore Standby Signaling     : false
Suppress Standby Signaling   : false
Block On Mesh Fail           : false
Multi-Chassis Endpoint       : 1
MC Endpoint Peer Addr        : 192.0.2.3
Psv Mode Active              : No
Tx Active                    : none
Tx Active Up Time            : 0d 00:00:00
Revert Time Count Down       : N/A
Tx Active Change Count       : 0
Last Tx Active Change        : 04/24/2017 07:59:47
-------------------------------------------------------------------------------
Members
-------------------------------------------------------------------------------
Spoke-sdp: 41:1 Prec:2                             Oper Status: Down
Spoke-sdp: 42:1 Prec:3                             Oper Status: Up
===============================================================================
===============================================================================
*A:PE-4#
```

Multi-Chassis Node Failure

Figure 102 **Multi-Chassis Node Failure**



When the multi-chassis node PE-3 fails, both spoke SDPs from PE-3 go down.

PE-4 reverts to single chassis mode and selects the best spoke SDP, which will be 41 between PE-4 and PE-1 (precedence 2). See Figure 102.

```
*A:PE-4# show redundancy multi-chassis mc-endpoint peer 192.0.2.3

===============================================================================
Multi-Chassis MC-Endpoint
===============================================================================
Peer Addr       : 192.0.2.3           Peer Name            :
Admin State     : up                  Oper State           : down
Last State chg  :                     Source Addr          :
System Id       : 16:0c:ff:00:00:00   Sys Priority         : 0
Keep Alive Intvl: 10                  Hold on Nbr Fail     : 3
Passive Mode    : disabled            Psv Mode Oper        : No
Boot Timer      : 300                 BFD                  : enabled
Last update     : 04/24/2017 08:13:23 MC-EP Count          : 1
===============================================================================
*A:PE-4#
```

```
*A:PE-4# show service id 2 endpoint

===============================================================================
Service 2 endpoints
===============================================================================
Endpoint name                : CORE
Description                  : (Not Specified)
Creation Origin              : manual
Revert time                  : 0
Act Hold Delay               : 0
Ignore Standby Signaling     : false
Suppress Standby Signaling   : false
Block On Mesh Fail           : false
Multi-Chassis Endpoint       : 1
MC Endpoint Peer Addr        : 192.0.2.3
Psv Mode Active              : No
Tx Active (SDP)              : 41:1
Tx Active Up Time            : 0d 00:02:40
Revert Time Count Down       : N/A
Tx Active Change Count       : 1
Last Tx Active Change        : 04/24/2017 08:17:47
-------------------------------------------------------------------------------
Members
-------------------------------------------------------------------------------
Spoke-sdp: 41:1 Prec:2                              Oper Status: Up
Spoke-sdp: 42:1 Prec:3                              Oper Status: Up
===============================================================================
===============================================================================
*A:PE-4#
```

Multi-Chassis Communication Failure

If the multi-chassis communication is interrupted, both nodes will revert to single chassis mode.

To simulate a communication failure between the two nodes, define a static route on PE-3 that will black-hole the system address of PE-4.

```
configure
    router
        static-route-entry 192.0.2.4/32
            black-hole
                no shutdown
            exit
        exit
```

Verify that the MC synchronization is down.

```
*A:PE-4# show redundancy multi-chassis mc-endpoint peer 192.0.2.3

===============================================================================
Multi-Chassis MC-Endpoint
===============================================================================
Peer Addr       : 192.0.2.3           Peer Name            :
Admin State     : up                  Oper State           : down
Last State chg  :                     Source Addr          :
System Id       : 16:0c:ff:00:00:00   Sys Priority         : 0
Keep Alive Intvl: 10                  Hold on Nbr Fail     : 3
Passive Mode    : disabled            Psv Mode Oper        : No
Boot Timer      : 300                 BFD                  : enabled
Last update     : 04/24/2017 08:13:23 MC-EP Count          : 1
===============================================================================
*A:PE-4#
```

The spoke SDPs are active on PE-3 and on PE-4.

```
*A:PE-3# show service id 2 endpoint | match "Tx Active"
Tx Active (SDP)              : 31:1
Tx Active Up Time            : 0d 00:05:58
Tx Active Change Count       : 6
Last Tx Active Change        : 04/24/2017 08:19:09

*A:PE-4# show service id 2 endpoint | match "Tx Active"
Tx Active (SDP)              : 41:1
Tx Active Up Time            : 0d 00:04:56
Tx Active Change Count       : 3
Last Tx Active Change        : 04/24/2017 08:19:05
```

This can potentially cause a loop in the system. The section Passive Mode describes how to avoid this loop.

Passive Mode

As in Multi-Chassis Communication Failure, if there is a failure in the multi-chassis communication, both nodes will assume that the peer is down and will revert to single-chassis mode. This can create loops because two spoke SDPs can become active.

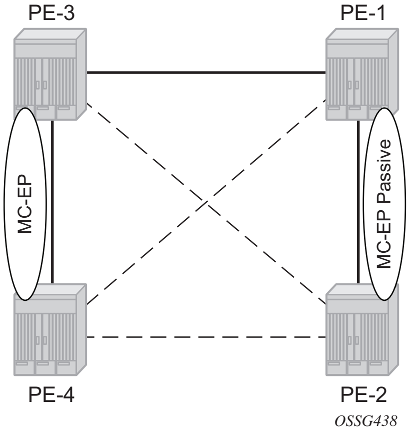
One solution is to synchronize the two core nodes, and configure them in passive mode. See Figure 103.

In passive mode, both peers will stay dormant as long as one active spoke SDP is signaled from the remote end. If more than one spoke SDP becomes active, the MC-EP algorithm will select the best SDP. All other spoke SDPs are blocked locally (in Rx and Tx directions). There is no signaling sent to the remote PEs.

If one peer is configured in passive mode, the other peer will be forced to passive mode as well.

The **no suppress-standby-signaling** and **no ignore-standby-signaling** commands are required.

Figure 103 **Multi-Chassis Passive Mode**



The following output shows the multi-chassis configuration on PE-1 (similar on PE-2).

```
*A:PE-1# configure
    redundancy
        multi-chassis
            peer 192.0.2.2 create
                mc-endpoint
                    no shutdown
                    passive-mode
                exit
                no shutdown
            exit
        exit
```

The following output shows the VPLS spoke SDPs configuration on PE-1 (similar on PE-2)

```
*A:PE-1# configure
    service
        vpls 1
            endpoint "METRO1" create
                no suppress-standby-signaling
                mc-endpoint 1
                    mc-ep-peer 192.0.2.2
                exit
            exit
            spoke-sdp 13:1 endpoint "METRO1" create
            exit
            spoke-sdp 14:1 endpoint "METRO1" create
            exit
            no shutdown
        exit
```

To simulate a communication failure between the two nodes, a static route is defined on PE-3 that will black-hole the system address of PE-4.

```
configure
    router
        static-route-entry 192.0.2.4/32
            black-hole
                no shutdown
            exit
        exit
```

The spoke SDPs are active on PE-3 and on PE-4.

```
*A:PE-3# show service id 2 endpoint | match "Tx Active"
Tx Active (SDP)             : 31:1
Tx Active Up Time           : 0d 00:00:28
Tx Active Change Count      : 8
Last Tx Active Change       : 04/24/2017 08:20:24

*A:PE-4# show service id 2 endpoint | match "Tx Active"
Tx Active (SDP)             : 41:1
Tx Active Up Time           : 0d 00:00:22
Tx Active Change Count      : 5
Last Tx Active Change       : 04/24/2017 08:20:25
```

PE-1 and PE-2 have blocked one spoke SDP which avoids a loop in the VPLS.

```
*A:PE-1# show service id 1 endpoint "METRO1" | match "Tx Active"
Tx Active (SDP)             : 13:1
Tx Active Up Time           : 0d 00:00:58
Tx Active Change Count      : 5
```

```
Last Tx Active Change        : 04/24/2017 08:20:50

*A:PE-2# show service id 1 endpoint  "METRO1" | match "Tx Active"
Tx Active                    : none
Tx Active Up Time            : 0d 00:00:00
Tx Active Change Count       : 2
Last Tx Active Change        : 04/24/2017 08:20:15
```

The passive nodes do not set the pseudowire status bits; therefore, the nodes PE-3 and PE-4 are not aware that one spoke SDP is blocked.

Conclusion

Multi-chassis endpoint for VPLS active/standby pseudowire allows the building of hierarchical VPLS without single point of failure, and without requiring STP to avoid loops.

Care must be taken to avoid loops. The multi-chassis peer communication is important and should be possible on different interfaces.

Passive mode can be a solution to avoid loops in case of multi-chassis communication failure.

Multi-Segment Pseudowire Routing

This chapter describes advanced multi-segment pseudowire routing configurations.

Topics in this chapter include:

- Applicability
- Summary
- Overview
- Configuration
- Conclusion

Applicability

Multi-Segment Pseudowire (MS-PW) routing is supported in SR OS Release 9.0.R3, and later. This chapter was initially written for SR OS Release 10.0.R4. The CLI in this edition is based on SR OS Release 15.0.R2. There are no specific prerequisites for this configuration.

Summary

SR OS supports the use of Multi-Segment Pseudowire (MS-PW) routing for Epipe services. MS-PW routing is described in draft-ietf-pwe3-dynamic-ms-pw, also known as dynamic placement of MS-PW and it is an extension of the procedures proposed in RFC 6073 (static MS-PW) to enable multi-segment pseudowires to be dynamically placed. Ultimately, MS-PW Routing provides the capability of setting up MS-PWs without provisioning the Switching PEs (S-PEs).

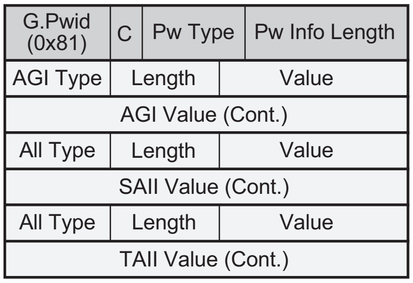
This chapter will go through the configuration process required to set up MS-PW routing and will provide two configuration examples typically deployed by service providers: MS-PW within the same Autonomous System (AS) and MS-PW across two different ASs. Different configuration options are shown and described for each example.

Overview

From a data plane perspective, MS-PW routing does not introduce any changes with respect to the existing MS-PW architecture. However, from the control plane perspective, MS-PW routing brings a new information model and set of procedures to set up a MS-PW. These are the building blocks defined by the MS-PW Routing feature:

- A new information model is introduced for dynamic MS-PW based on the FEC129, Attachment Individual Identifier (AII) Type 2. Static MS-PW uses FEC128 whereas VPLS with BGP-AD uses FEC129, but with AII Type 1 instead.
    - FEC129 is suitable for applications where the local PE with a Source Attachment Individual Identifier (SAII) must automatically learn the remote Target Attachment Individual Identifier (TAII), normally through BGP, before launching the LDP mapping message for the pseudowire setup. The following figure shows the FEC129 structure:

Figure 104 **FEC129 Structure**



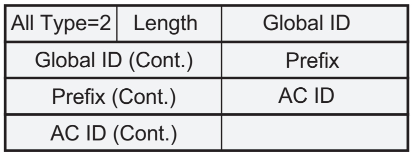
G.Pwid (0x81)	C	Pw Type	Pw Info Length
AGI Type	Length		Value
AGI Value (Cont.)			
All Type	Length		Value
SAII Value (Cont.)			
All Type	Length		Value
TAII Value (Cont.)			

ACG0004A

-
    - The Attachment Group Identifier (AGI) is not used in dynamic MS-PW signaling. In VPLS, it typically carries the instance identifier. It is zero in dynamic MS-PWs.
    - The SAII and TAII (or pseudowire end-point identifiers) are encoded in FEC129 and can have two different formats: AII Type 1 or AII Type 2.
- AII Type 1 is composed of a fixed 32-bit value unique on the local PE. This AII type is used by VPLS when BGP-AD is needed.

- AII Type 2 is composed of GID:prefix:AC-ID (global-ID:prefix: attachment-circuit-ID) and allows for summarization, thereby enhancing scalability in large networks. The GID is normally derived from the AS number, the prefix from the node system address and the AC-ID is the local pseudowire end-point identifier. The combination of the three identifiers gives us a globally unique 96-bit AII value. In general, the same global ID and prefix are assigned for all ACs belonging to the same Terminating PE (T-PE). This is not a strict requirement though.

Figure 105 **AII Type 2 Format**

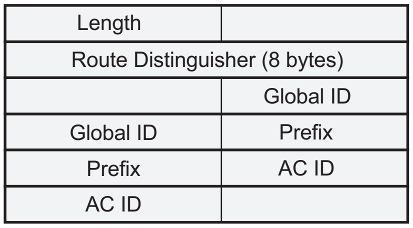


AII Type=2	Length	Global ID
Global ID (Cont.)		Prefix
Prefix (Cont.)		AC ID
AC ID (Cont.)		

ACG0004B

- A MS-PW routing table must be built in all the T-PEs and S-PEs through one of the following two mechanisms:
  - Multi-protocol BGP (MP-BGP), using a dedicated NLRI and SAFI (pseudowire routing SAFI=6, with AFI=25 L2VPN). The FEC129 AII Type 2 global values are mapped in the pseudowire routing NLRI and advertised by BGP. SR OS supports an NLRI comprising a Length, RD, Global ID, and 32-bit prefix, that is, the AC ID is not included in the advertised NLRI. The AC ID is not included as indicated in the draft-ietf-pwe3-dynamic-ms-pw because the source T-PE knows by provisioning the AC ID on the terminating T-PE to use in signaling. Therefore, there is no need to advertise a “fully qualified” 96 bit address on a per pseudowire attachment circuit basis. Only the T-PE Global ID, Prefix, and prefix length need to be advertised as part of well known BGP procedures. This also minimizes the amount of routing information that is advertised in BGP to only what is necessary to reach the far-end T-PE. The MS-PW routing NLRI is shown in Figure 106:

Figure 106 **Pseudowire Routing NLRI (the AC ID is always zero)**



Length	
Route Distinguisher (8 bytes)	
	Global ID
Global ID	Prefix
Prefix	AC ID
AC ID	

ACG0004C

  - Static routes, configurable via CLI
- Once the MS-PW routing table is populated, Targeted LDP (TLDP) will make use of it to signal the MS-PW all the way from the originating T-PE to the terminating T-PE as well as in the reverse direction. The following methods will be used:
  - At the originating T-PE, a longest-match lookup will be performed in the pseudowire routing table for the configured TAII. Based on the lookup outcome, a label mapping message will be sent to the Next Signaling Hop (NSH).

➡ **Note:** The “originating T-PE” will be the T-PE initiating the MS-PW signaling. See the Active/Passive Signaling and Auto-Configuration section for further information.

  - At the intermediate S-PEs and terminating T-PE, a longest-match lookup between the TAII Type 2 included in the TLDP signaling message and entries installed in the pseudowire routing table will be performed.
  - Alternatively to the pseudowire routing table lookup, TLDP can also use explicit routing, as per section 7.4.2 of draft-ietf-pwe3-dyn-ms-pw. If that is the case, a “path” must be configured at the T-PEs. The originating T-PE will include an ERO (Explicit Route Object) in the TLDP label mapping, containing all the S-PE hops specified in the configured path. Each S-PE along the path will remove its own entry from the ERO and will forward the label mapping message to the next hop.

SR OS supports the information model and all the previously described methods:

- Dynamic placement through MP-BGP, with the pseudowire routing NLRI
- Static routes
- Explicit paths

In addition to the above, the following features are supported on dynamic MS-PW:

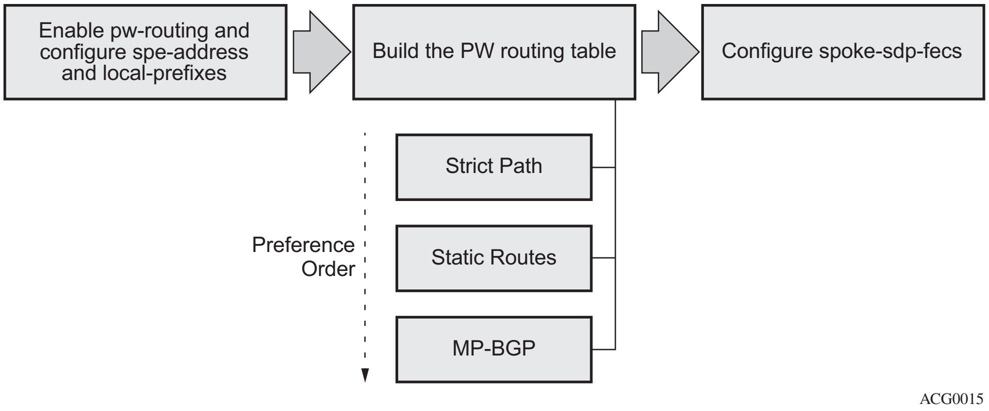
- Auto-configuration of spoke SDPs at T-PE (if enabled on a T-PE, there is no need for configuring the TAII of the remote T-PE, see Active/Passive Signaling and Auto-Configuration. The auto-configuration is typically used in hub-and-spoke scenarios. The TAII would only be configured on the spoke T-PE whereas the TAII would be automatically provisioned on the hub T-PE if the auto-config parameter is added.
- OAM using virtual circuit connectivity verification vccv-ping and vccv-trace
- Pseudowire redundancy
- Control word

- Hash label
- Standby-signaling-master and standby-signaling-slave commands
- Filters

Configuration

The following flowchart shows the configuration process to be followed when setting up MS-PW routing. Base IGP and MPLS configuration is assumed to be in place before these configuration tasks can be carried out.

Figure 107 **Configuration Flow Chart**


Enable pw-routing and configure spe-address and local-prefixes
Build the PW routing table
Configure spoke-sdp-fecs
Strict Path
Static Routes
MP-BGP
Preference Order
ACG0015

The following subsections review these three steps, including all the options in detail.

- Pseudowire Routing Enablement
- Building the Pseudowire Routing Table
- Spoke-SDP-FEC Timers

Pseudowire Routing Enablement

The first step in the configuration is to enable **pw-routing** and configure the required pseudowire routing basic parameters: the **spe-address** (in S-PEs and T-PEs) and the **local-prefix**/prefixes (only required in T-PEs). The following CLI examples show the configuration of the spe-address and local-prefixes.

```
*A:PE-1# configure
    service
        pw-routing
```

```
            spe-address 65536:192.0.2.1
            local-prefix 65536:192.0.2.11 create
                advertise-bgp route-distinguisher 65536:11 community 65535:11
            exit
            local-prefix 65536:192.0.2.12 create
                advertise-bgp route-distinguisher 65536:12 community 65535:12
            exit
        exit
```

In order to enable support for MS-PW routing on an SR OS router, a single, globally unique, S-PE ID (known as the spe-address) is first configured in the **config>service>pw-routing** context on each SR OS router to be used as a T-PE or S-PE. The S-PE address has the format global-id:prefix. It is not possible to configure any local prefixes used for pseudowire routing or to configure spoke SDPs using dynamic MS-PWs at a T-PE, unless an S-PE address has already been configured. The S-PE address is used as the address of a node when populating the switching point TLV in the LDP label mapping message and the pseudowire status notification sent for faults at an S-PE. The following CLI output shows the spe-address configuration format:

```
A:PE-1# configure service pw-routing spe-address
  - no spe-address
  - spe-address <global-id:prefix>

 <global-id:prefix>   : <global-id>:{<prefix>|<ipaddress>}
                        global-id - [1..4294967295]
                        prefix    - [1..4294967295]
                        ipaddress - a.b.c.d
```

Where:

- <global-id> is normally the 2 or 4-byte ASN identifying the network (although nothing prevents the operator from configuring any value here)
- <prefix> is normally the system address of the node (although any value in IP address or decimal format can be used)

If an S-PE is capable of dynamic MS-PW signaling, but is not assigned with an S-PE address, then on receiving a dynamic MS-PW label mapping message the S-PE will return a label release with the "LDP_RESOURCES_UNAVAILABLE" (0x38)" status code. The S-PE address cannot be changed unless the dynamic MS-PW configuration is completely removed; therefore Nokia recommends to configure the spe-address carefully and keep it for the life of the services.

The second basic pw-routing context parameter is the local-prefix:

```
A:PE-1# configure service pw-routing local-prefix
  - local-prefix <local-prefix> [create]
  - no local-prefix <local-prefix>

 <local-prefix>       : <global-id>:<ip-addr>|<raw-prefix>
                        ip-addr      - a.b.c.d
                        raw-prefix   - [1..4294967295]
                        global-id    - [1..4294967295]

 [no] advertise-bgp   - Configure BGP advertisement
```

One or more local (Layer 2) prefixes (up to a maximum of 16), which are formatted in the style of <global-id>:<ipv4-address>, are supported. A local prefix identifies a T-PE in the pseudowire routing domain. When using explicit paths or static routes, the definition of the local-prefix (or local-prefixes) without any further attribute is enough. However, when BGP is used, the **advertise-bgp** parameter along with a Route Distinguisher (RD) value and an optional BGP community is required.

```
*A:PE-1# configure service pw-routing local-prefix 65536:192.0.2.11 advertise-bgp
  - advertise-bgp route-distinguisher <rd> [community <community>]
  - no advertise-bgp route-distinguisher <rd>

 <rd>                 : <ip-addr:comm-val>|<2byte-asnumber:ext-comm-val>|
                        <4byte-asnumber:comm-val>
                        ip-addr        - a.b.c.d
                        comm-val       - [0..65535]
                        2byte-asnumber - [1..65535]
                        ext-comm-val   - [0..4294967295]
                        4byte-asnumber - [1..4294967295]
 <community>          : <asnumber:comm-val>
                        asnumber - [1..65535]
                        comm-val - [0..65535]
```

Up to four unique RDs (and communities) can be configured per each local prefix. Different RDs for the same prefix allow the operator to advertise the same prefix coming from up to four different Next Signaling Hops (NSHs). Route-Reflectors would reflect the four routes in that case, whereas only one would be reflected should the same RD be used.

```
*A:PE-1>config>service>pw-routing>local-prefix# info
----------------------------------------------
                advertise-bgp route-distinguisher 400:20
                advertise-bgp route-distinguisher 500:3
                advertise-bgp route-distinguisher 600:300
                advertise-bgp route-distinguisher 65536:11 community 65535:11


*A:PE-1>config>service>pw-routing>local-prefix# advertise-bgp route-distinguisher
700:100
MINOR: SVCMGR #6072 Maximum number of RD's has been reached
```

For each local prefix, BGP then advertises each global ID/prefix tuple and unique RD and community (if configured) using the MS-PW NLRI, based on the aggregated FEC129 AII Type 2 and the Layer 2 VPN/PW routing AFI/SAFI 25/6, to each BGP neighbor, subject to local BGP policies.

Building the Pseudowire Routing Table

Once the spe-address and the local-prefix(es) have been configured and before configuring the Epipe service itself on the T-PE nodes, we need to populate the pseudowire routing table in all the participating T-PE and S-PE nodes, so that TLDP knows what the Next Signaling Hop (NSH) is and sends LDP label mapping messages.

The pseudowire routing table will be populated with local prefixes, static routes and BGP routes, where the static routes have preference over the BGP-learned routes. The pseudowire routing table can be overridden by the explicit paths, should the operator want to configure them. Therefore, when TLDP signals an LDP Label Mapping for a TAII, it will:

- First check if there is an explicit path configured for that spoke-sdp-fec.
- Otherwise it will look up the TAII prefix into the pseudowire routing table, where static routes take precedence over BGP routes.

An aggregation scheme, similar to that used for classless IPv4 addresses, can be employed in the pseudowire routing table, where a longest match is used to find a route. Except for the default pseudowire route, which is encoded with a zero mask, masks included in the pw-routing table are:

- /64 for regular prefixes, including a global ID and prefix (as previously mentioned; the AC-ID is not included in the BGP NLRI).
- /96 for local prefixes, including the AC-ID, as well as global-id and prefix.

Each S-PE and T-PE must have a pseudowire routing table that contains a reference to the TLDP session to use to signal to a set of next hop S-PEs to reach a T-PE (or the T-PE if that is the next hop). For Epipes, this table contains aggregated AII Type 2 FECs and may be populated with routes that are learned through MP-BGP or that are statically configured.

Explicit Paths

A set of default explicit routes to a remote T-PE prefix may be configured on a T-PE under **config>services>pw-routing** using the path name command. Explicit paths are used to populate the explicit route TLV used by MS-PW TLDP signaling. Only strict (fully qualified) explicit paths are supported. It is possible to configure explicit paths independently of the configuration of BGP or static routing.

The following CLI excerpt shows an explicit path example for a MS-PW following the PE-1–PE-3–PE-5–PE-2 path (see the diagram in Figure 108). The IP addresses are the system addresses of all the S-PE and T-PE along the path (except for PE-1).

```
*A:PE-1# configure
    service
        pw-routing
            path "path-1" create
                hop 1 192.0.2.3
                hop 2 192.0.2.5
                hop 3 192.0.2.2
                no shutdown
            exit
```

Static Routes

In addition to support for BGP routing, static MS-PW routes may also be configured using the **config>services>pw-routing>static-route** command. Each static route comprises of the target T-PE Global ID and prefix, and the IP address of the TLDP session to the next hop S-PE or T-PE that should be used:

```
A:PE-1# configure service pw-routing static-route
  - no static-route <route-name>
  - static-route <route-name>

 <route-name>          : <global-id>:<prefix>:<next-hop-ip_addr>
                         global-id      - 0..4294967295
                         prefix         - a.b.c.d|0..4294967295
                         ip_addr        - a.b.c.d
```

If a static route <global-id>:<prefix> is set to 0, then this represents the default route.

```
*A:PE-1# configure service pw-routing
    ---snip---
    static-route 0:0.0.0.0:192.0.2.3
    static-route 0:0.0.0.0:192.0.2.4
```

Even though several default-routes can be configured, only one default route is added to the pseudowire routing table. The following command shows the pseudowire routing table content where only one default route (out of the two previously configured ones) is added. The default route added to the pseudowire routing table is the first valid route added to the configuration.

```
A:PE-1# show service pw-routing route-table all-routes

===============================================================================
Service PW L2 Routing Information
===============================================================================
AII-Type2/Prefix-Len                           Next-Hop       Owner  Age
 Route-Distinguisher                            Community      Best
```

```
-------------------------------------------------------------------------------
0:0.0.0.0:0/0                                          192.0.2.3      static 19h11m57s
 0:0                                                    0:0            yes
...
```

If a static route exists to a T-PE, then this is used in preference to any BGP route that may exist.

BGP Routes

As already mentioned, the dynamic advertisement of the pseudowire routes is enabled for each prefix and RD using the **advertise-bgp** command in the **config>services>pw-routing>local-prefix** context. A BGP export policy is required in order to export MS-PW routes in MP-BGP. This can be done using a default policy matching all the MS-PW routes, such as the following:

```
*A:PE-1# configure
    router
        policy-options
            begin
            policy-statement "export_ms-pw"
                entry 10
                    from
                        family ms-pw
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit

*A:PE-1# configure
    router
        autonomous-system 65536
        bgp
            enable-peer-tracking
            rapid-withdrawal
            group "region"
                family ms-pw
                export "export_ms-pw"
                peer-as 65536
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
            exit
        exit
```

MS-PW routes advertised/received can be debugged and shown on the log sessions (**debug router bgp update**). A dedicated MS-PW address family and NLRI are used to distribute the MS-PW prefixes. The following BGP update is sent by PE-1 to PE-3:

```
28 2017/04/28 07:42:59.69 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 51
    Flag: 0x90 Type: 14 Len: 26 Multiprotocol Reachable NLRI:
        Address Family MSPW
        NextHop len 4 NextHop 192.0.2.1
        [MSPW] rd: 65536:12, global-id 65536, prefix 192.0.2.12,  ac-id 0, preflen 128
    Flag: 0x40 Type: 1 Len: 1 Origin: 2
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        65535:12
"
```

MS-PW BGP routes can also be displayed in the pseudowire routing table along with the static-routes and the local-prefixes.

```
*A:PE-1# show service pw-routing route-table

===============================================================================
Service PW L2 Routing Information
===============================================================================
AII-Type2/Prefix-Len                           Next-Hop       Owner  Age
 Route-Distinguisher                            Community      Best
-------------------------------------------------------------------------------
0:0.0.0.0:0/0                                  192.0.2.3      static 01h09m00s
 0:0                                            0:0            yes
65536:192.0.2.11:0/64                          192.0.2.1      local  17h19m42s
 0:0                                            0:0            yes
65536:192.0.2.11:0/64                          192.0.2.1      local  17h19m42s
 65536:11                                       65535:11       yes
65536:192.0.2.12:0/64                          192.0.2.1      local  17h19m42s
 0:0                                            0:0            yes
65536:192.0.2.12:0/64                          192.0.2.1      local  17h19m42s
 65536:12                                       65535:12       yes
65536:192.0.2.13:0/64                          192.0.2.1      local  00h49m29s
 0:0                                            0:0            yes
65536:192.0.2.14:0/64                          192.0.2.1      local  00h49m29s
 0:0                                            0:0            yes
65536:192.0.2.21:0/64                          192.0.2.3      bgp    00h05m46s
 65536:21                                       65535:11       yes
65536:192.0.2.22:0/64                          192.0.2.4      bgp    00h05m42s
 65536:22                                       65535:12       yes
65536:192.0.2.23:0/64                          192.0.2.3      static 00h49m29s
 0:0                                            0:0            yes
65536:192.0.2.24:0/64                          192.0.2.4      static 00h49m29s
 0:0                                            0:0            yes
-------------------------------------------------------------------------------
Entries found: 11
===============================================================================
```

If there are two (or more) equal cost BGP MS-PW routes with identical <global-ID:prefix> and different RDs in the RIB, they are both tagged as best/used and both will be added to the pseudowire routing table, however, only the one with a higher RD will be shown as "Best" and as a result of that, only that one will be used by TLDP for the NSH.

The pw-routing context at PE-2 contains the following advertise-bgp entries with different RDs for local-prefix 65536:192.0.2.2:

```
*A:PE-2# configure
    service
        pw-routing
            local-prefix 65536:192.0.2.2 create
                advertise-bgp route-distinguisher 65536:21
                advertise-bgp route-distinguisher 65536:22
            exit
```

The following CLI output shows an example of two equal cost MS-PW routes. The route 65536:192.0.2.2 with RD 65536:21 and RD 65536:22 are tagged as best/used (u*>):

```
*A:PE-1# show router bgp routes ms-pw aii-type2 65536:192.0.2.2:0
===============================================================================
 BGP Router ID:192.0.2.1        AS:65536       Local AS:65536
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MSPW Routes
===============================================================================
Flag  Network                 RD
      Nexthop                 AII-Type2/Preflen
      As-Path
-------------------------------------------------------------------------------
u*>?  65536:192.0.2.2         65536:21
      192.0.2.3               65536:192.0.2.2:0/64
      No As-Path
*?    65536:192.0.2.2         65536:21
      192.0.2.4               65536:192.0.2.2:0/64
      No As-Path
u*>?  65536:192.0.2.2         65536:22
      192.0.2.4               65536:192.0.2.2:0/64
      No As-Path
*?    65536:192.0.2.2         65536:22
      192.0.2.3               65536:192.0.2.2:0/64
      No As-Path
-------------------------------------------------------------------------------
Routes : 4
===============================================================================
```

However, only the one with RD 65536:22 (higher RD) is added as “Best” to the pseudowire routing table and TLDP will use 192.0.2.4 as the NSH:

```
*A:PE-1# show service pw-routing route-table all-routes

===============================================================================
Service PW L2 Routing Information
===============================================================================
AII-Type2/Prefix-Len                           Next-Hop       Owner  Age
 Route-Distinguisher                            Community      Best
-------------------------------------------------------------------------------
65536:192.0.2.2:0/64                           192.0.2.3      bgp    00h12m22s
 65536:21                                       65535:11       no
65536:192.0.2.2:0/64                           192.0.2.4      bgp    00h12m17s
 65536:22                                       65535:12       yes
---snip---
```

How does the SR OS TLDP process select the NSH (Next-Signaling Hop) for two identical <global-ID:prefix/RD> tuples?

In case the originating T-PE or any intermediate S-PE receives two (or more) equal cost MS-PW routes with the same RD but from different Next-Hops, all the MS-PW routes will be added to the MS-PW routing table. The following output shows two MS-PW routes with the same <global-ID:prefix/RD> but different NH. Both are added to the MS-PW routing table as “Best”.

```
*A:PE-1# show service pw-routing route-table all-routes

===============================================================================
Service PW L2 Routing Information
===============================================================================
AII-Type2/Prefix-Len                           Next-Hop       Owner  Age
 Route-Distinguisher                            Community      Best
-------------------------------------------------------------------------------
---snip---
65536:192.0.2.2:0/64                           192.0.2.3      bgp    01d15h21m
 65536:21                                       65535:21       yes
65536:192.0.2.2:0/64                           192.0.2.4      bgp    01d15h21m
 65536:21                                       65535:21       yes
---snip---
```

If that is the case, TLDP will pick up the NSH out of an ECMP hashing algorithm applied to the <global-ID:prefix:AC-ID> for the SAII and the TAII of the pseudowires pointing at the same prefix. The output of that hashing algorithm will determine what the NSH will be for a spoke-SDP FEC.

When path diversity for an active and a standby pseudowire (hot standby pseudowire redundancy) is desired and the two pseudowires of the same Epipe end-point are pointing at the same remote <global-ID:prefix> coming from two different NHs, the operator has to make sure TLDP chooses a different NSH for the standby pseudowire. Only in that case, hot standby pseudowire redundancy can be achieved. As a rule of thumb, if the SAII/TAII of the active and standby pseudowires are separated by 16 or more AC-ID values, TLDP will select a different NSH for both pseudowires.

For example:

- Given the following SAII/TAII AC-ID values for the active/standby pseudowires on the originating T-PE, TLDP will select the same NSH:
  - Active pseudowire: saii-type2 — 65536:192.0.2.1:1, taii-type2 — 65536:192.0.2.2:1
  - Standby pseudowire: saii-type2 — 65536:192.0.2.1:2, taii-type2 — 65536:192.0.2.2:2
- However, the following SAII/TAII AC-ID values for the active/standby pseudowires on the originating T-PE will allow the ECMP hashing algorithm to make TLDP select different NSHs for the active and the standby pseudowires:
  - Active pseudowire: saii-type2 — 65536:192.0.2.1:1, taii-type2 — 65536:192.0.2.2:1
  - Standby pseudowire: saii-type2 — 65536:192.0.2.1:16, taii-type2 — 65536:192.0.2.2:16

Other AC-ID values greater than 16 (for the standby pseudowire) would also have achieved next hop diversity.

Configuring Dynamic Pseudowires on the T-PEs

Before any LDP signaling can take place, T-LDP sessions must be explicitly configured on T-PEs and S-PEs.

One or more spoke-SDPs may be configured for distributed Epipe VLL services. Dynamic MS-PWs use FEC129 (also known as the Generalized ID FEC) with Attachment Individual Identifier (AII) Type 2 to identify the pseudowire, as opposed to FEC128 (also known as the PW ID FEC) used for traditional single segment pseudowires and for pseudowire switching. FEC129 spoke-SDPs are configured under the spoke-sdp-fec command in the CLI. Spoke-sdp-fecs (or FEC129 spoke-SDPs) are by default fec-type 129 and aii-type 2. Spoke-sdp-fecs can be part of an endpoint and even an ICB (Inter-Chassis Backup) pseudowire.

```
*A:PE-1# configure service epipe 2 spoke-sdp-fec
```

```
  - no spoke-sdp-fec <spoke-sdp-fec-id>
  - spoke-sdp-fec <spoke-sdp-fec-id> [fec <fec-type>] [aii-type <aii-type>] [create]
  - spoke-sdp-fec <spoke-sdp-fec-id> no-endpoint
  - spoke-sdp-fec <spoke-sdp-fec-id> [fec <fec-type>] [aii-type <aii-type>] [create]
endpoint <name> [icb]

 <spoke-sdp-fec-id>   : [1..4294967295]
 <fec-type>           : [129..130]
 <aii-type>           : [1..2]
 <name>               : [32 chars max]
 <icb>                : keyword - configure spoke-sdp as inter-chassis backup
```

FEC129 AII Type 2 uses a SAII and a TAII to identify the ends of a pseudowire at the T-PE. The SAII identifies the local end, while the TAII identifies the remote end. The SAII and TAII are each structured as follows:

- Global-ID: this is a 4 byte identifier that uniquely identifies an operator or the local network. Normally this matches the ASN
- Prefix: a 4-byte prefix, which should correspond to one of the local prefixes assigned under pw-routing
- AC-ID: a 4-byte identifier for this end of the pseudowire. This should be locally unique within the scope of the global-id:prefix

In terms of the SDP tunnel being used by each spoke-sdp-fec, pw-routing chooses the MS-PW path in terms of the sequence of S-PEs to use to reach a T-PE. It does not select the SDP to use on each hop, which is instead determined at signaling time. When a label mapping is sent for a pseudowire segment, an LDP SDP will be used to reach the next-hop S-PE/T-PE if such an SDP exists. If not, and an RFC 3107 labeled BGP SDP is available, then that will be used. Otherwise, the label mapping will fail and a label release will be sent.

Note: The RSVP SDPs might be picked at the T-PE through the use of pw-template <policy-id> [use-provisioned-sdp], however there is no way to select an RSVP SDP on an S-PE.

The following CLI output shows one example of two spoke-sdp-fecs belonging to an endpoint:

```
*A:PE-1# configure
    service
        pw-template 1 create
            controlword
        exit
        epipe 2 customer 1 create
            description "ms-pw epipe with bgp - using 2 prefixes"
            endpoint "CORE" create
                description "end-point for epipe A/S PW redundancy"
                revert-time 10
                standby-signaling-master
            exit
            sap 1/1/4:2 create
            exit
            spoke-sdp-fec 21 fec 129 aii-type 2 create endpoint CORE
                precedence primary
                pw-template-bind 1
                saii-type2 65536:192.0.2.11:1
                taii-type2 65536:192.0.2.21:1
                no shutdown
            exit
            spoke-sdp-fec 22 fec 129 aii-type 2 create endpoint CORE
                pw-template-bind 1
                saii-type2 65536:192.0.2.12:1
                taii-type2 65536:192.0.2.22:1
                no shutdown
            exit
            no shutdown
        exit
```

These are all of the options available in the spoke-sdp-fec context:

```
*A:PE-1# configure service epipe 1 spoke-sdp-fec
  - no spoke-sdp-fec <spoke-sdp-fec-id>
  - spoke-sdp-fec <spoke-sdp-fec-id> [fec <fec-type>] [aii-type <aii-type>] [create]
  - spoke-sdp-fec <spoke-sdp-fec-id> no-endpoint
  - spoke-sdp-fec <spoke-sdp-fec-id> [fec <fec-type>] [aii-type <aii-type>] [create]
                                     endpoint <name> [icb]

 <spoke-sdp-fec-id>   : [1..4294967295]
 <fec-type>           : [129..130]
 <aii-type>           : [1..2]
 <name>               : [32 chars max]
 <icb>                : keyword - configure spoke-sdp as inter-chassis backup

 [no] auto-config     - Configure auto-configuration
 [no] path            - Configure path-name
 [no] precedence      - Configure precedence
 [no] pw-template-bi* - Configure Pseudo-Wire template-binding policy
 [no] retry-count     - Configure retry count
 [no] retry-timer     - Configure retry timer
 [no] saii-type2      - Configure Source Attachment Individual Identifier (SAII)
 [no] shutdown        - Administratively enable/disable the spoke SDP FEC binding
      signaling       - Configure Spoke-SDP FEC signaling
 [no] standby-signal* - Enable PW standby-signaling slave
 [no] taii-type2      - Configure Target Attachment Individual Identifier (TAII)
```

Active/Passive Signaling and Auto-Configuration

When an MS-PW is signaled, each T-PE might independently initiate signaling of the MS-PW. This could result in a different path being used in each direction of the pseudowire. To avoid this situation one of the T-PEs will start the pseudowire signaling (active role), while the other T-PE waits to receive the LDP label mapping message before sending the LDP label mapping message for the reverse direction of the pseudowire (passive role).

Enable debugging for LDP messages on PE-2:

```
*A:PE-2# debug router ldp peer 192.0.2.5 packet init detail
*A:PE-2# debug router ldp peer 192.0.2.5 packet label detail
*A:PE-2# debug router ldp peer 192.0.2.6 packet init detail
*A:PE-2# debug router ldp peer 192.0.2.6 packet label detail
```

By default, the T-PE with SAII>TAII will have the active role and will send the label mapping first. When spoke-sdp-fec 21 is first disabled, and then enabled, PE-2 sends a label mapping to PE-5 first (message 77 in following output). Afterwards, it receives a label mapping packet from PE-5 (message 78).

```
*A:PE-2# configure service epipe 2 spoke-sdp-fec 21 shutdown
*A:PE-2# configure service epipe 2 spoke-sdp-fec 21 no shutdown


*A:PE-2# show log log-id 3
===============================================================================
Event Log 3
===============================================================================

78 2017/04/28 12:07:01.09 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Recv Label Mapping packet (msgId 11220) from 192.0.2.5:0
Protocol version = 1
Label 262131 advertised for the following FECs
Service FEC GENPWE3: ENET(5)
AGI = type: 1, len: 8, val: 00:00
SAII = T: 2, L: 12, Global-id: 65536, Prefix: 192.0.2.11, AcId: 1
TAII = T: 2, L: 12, Global-id: 65536, Prefix: 192.0.2.21, AcId: 1
Group ID = 0 cBit = 1
Interface parameter Mtu = 1500
Interface parameter VCCV = 0x106
PW status bits = 0x18
Switching hop: System = 192.0.2.3, Remote System = 192.0.2.1
previous segment fec AGI = type: 1, len: 8, val: 00:00
SAII = T: 2, L: 12, Global-id: 65536, Prefix: 192.0.2.11, AcId: 1
TAII = T: 2, L: 12, Global-id: 65536, Prefix: 192.0.2.21, AcId: 1
S-PE = T: 2, L: 12, Global-id: 65536, Prefix: 192.0.2.3, AcId: 0
Switching hop: System = 192.0.2.5, Remote System = 192.0.2.3
previous segment fec AGI = type: 1, len: 8, val: 00:00
SAII = T: 2, L: 12, Global-id: 65536, Prefix: 192.0.2.11, AcId: 1
TAII = T: 2, L: 12, Global-id: 65536, Prefix: 192.0.2.21, AcId: 1
S-PE = T: 2, L: 12, Global-id: 65536, Prefix: 192.0.2.5, AcId: 0
"

77 2017/04/28 12:07:01.09 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Send Label Mapping packet (msgId 8687) to 192.0.2.5:0
Protocol version = 1
Label 262132 advertised for the following FECs
Service FEC GENPWE3: ENET(5)
AGI = type: 1, len: 8, val: 00:00
SAII = T: 2, L: 12, Global-id: 65536, Prefix: 192.0.2.21, AcId: 1
TAII = T: 2, L: 12, Global-id: 65536, Prefix: 192.0.2.11, AcId: 1
Group ID = 0 cBit = 1
Interface parameter Mtu = 1500
Interface parameter VCCV = 0x306
PW status bits = 0x0
"
```

For the other T-PE, it is the other way round. PE-1 receives a label mapping packet first before it sends a label mapping packet back.

This default behavior can be modified by the signaling command. When set to master, the T-PE will send a label mapping message regardless of the SAII and TAII. By default the parameter is set to auto (which means the T-PE will trigger label mapping if SAII>TAII).

```
*A:PE-1# configure service epipe 1 spoke-sdp-fec 21 signaling
  - signaling <signaling>

 <signaling>          : auto|master
```

```
*A:PE-1# configure
    service
        epipe 2
            spoke-sdp-fec 21
                shutdown
                signaling master
                no shutdown
                exit
```

The MS-PW routing implementation on SR OS supports single-sided auto-provisioning. This allows it to have “hub” T-PEs where the TAII is not required to be configured and as such simplifies the provisioning. In this case, the spoke T-PE PWs would be configured with specific SAII and TAII as well as signaling master, whereas the hub T-PE PWs would be configured with only the SAII and the auto-config parameter. When the auto-config attribute is set for a spoke-SDP FEC, the T-PE always passively waits for the label mapping to be received before issuing a label mapping message (because it does not know the TAII beforehand). This is a CLI example for a hub T-PE spoke-SDP FEC:

```
*A:PE-2# configure
    service
        epipe 2
            spoke-sdp-fec 21 fec 129 aii-type 2 create endpoint CORE
                auto-config
                precedence primary
                pw-template-bind 1
                saii-type2 65536:192.0.2.21:1
                no shutdown
            exit
```

Spoke-SDP-FEC Timers

MS-PW routing provides a few timers that can be configured at the global pw-routing level or at each specific spoke-sdp-fec level:

```
*A:PE-1# configure
    service
        pw-routing
            boot-timer 20
            retry-timer 40
            retry-count 50

*A:PE-1# configure
    service
        epipe 2
            spoke-sdp-fec 21
                retry-timer 10
                retry-count 10
```

Where:

- Boot-timer (the default is 10 seconds with values 0 — 600 seconds allowed): Configures a hold-off timer for MS-PW routing advertisements and signaling that is used at boot time. This timer helps to make sure all the network infrastructure is up and running before setting up the PWs.
- Retry-timer (the default is 30 seconds with values 10 — 480 seconds allowed): The exponential back-off timer that determines the interval between consecutive retries to re-establish a spoke-SDP. The configured value gives the initial retry time. The attempt fails if a label withdrawal is received. If configured at global and spoke-sdp-fec level, the latter overrides the value set by the global settings.
- Retry-count (the default 30 with values 10 — 10000): Specifies the number of attempts the system should make to re-establish the spoke-SDP after it has failed. After each successful attempt, the counter is reset to zero. When the specified number is reached, no more attempts are made and the spoke-SDP is put into the shutdown state. Use the **no shutdown** command to bring up the path after the retry limit is exceeded. It is present at the pw-routing level as well as the spoke-sdp-fec level. If configured at global and spoke-sdp-fec level, the latter overrides the value set by the global settings.
- The usual endpoint level timers are also available for MS-PW routing:
  - Revert-time <time-value|infinite> (default is 0, values 0-600 sec): configures the time to wait before reverting to the primary spoke-sdp-fec.

- Active-hold-delay (the default is 0, values 0 — 60 deci-seconds): It specifies that the node will delay sending the T-LDP status bits for VLL endpoint when the MC-LAG transitions the LAG subgroup which hosts the SAP from active to standby (MC-Ring or MC-APS are supported too) or when any object in the endpoint—SAP, ICB, or regular spoke SDP—transitions from up to down operational state. The active-hold-delay range starts from 1 (in units of deci-seconds) via CLI, and the only way to get the default value of zero is to use the **no active-hold-delay** command

Standby Signaling

Just as with a regular endpoint with regular spoke-SDPs, there can also be standby-signaling-master and standby-signaling-slave parameters for spoke-SDP FECs.

The **standby-signaling-master** command is configured in the **endpoint** context and makes sure that standby signaling (TLDP pseudowire status bits 0x20) is sent for the selected standby pseudowire.

```
*A:PE-1# configure service epipe 2 endpoint "CORE" standby-signaling-master
```

It is not allowed to add a SAP associated to an endpoint configured as standby-signaling-master to an Epipe.

```
*A:PE-1>config>service>epipe# sap 1/1/4:2 endpoint "CORE" create
MINOR: SVCMGR #6025 The endpoint has standby-signaling-master configured
```

Standby-signaling-master cannot be set if SAPs have been configured at the end-point (for MC-LAG/Ring/APS or ICB).

```
*A:PE-1>config>service>epipe>endpoint# standby-signaling-master
MINOR: SVCMGR #3805 The command is not allowed in an endpoint with sap
```

The standby-signaling-slave can be configured at endpoint or spoke-sdp-fec level (if the spoke-sdp-fec is not part of an endpoint) but never on both at the same time:

```
*A:PE1>config>service>epipe>endpoint# info
----------------------------------------------
                standby-signaling-slave


*A:PE1>config>service>epipe>spoke-sdp-fec# standby-signaling-slave
MINOR: SVCMGR #2031 Sdp-bind is in an explicit endpoint


*A:PE1>config>service>epipe# info
----------------------------------------------
            sap 1/1/3:3 create
            exit
            spoke-sdp-fec 11 fec 129 aii-type 2 create
```

```
                standby-signaling-slave
```

When this parameter is configured, the node will block the transmit forwarding direction of a spoke SDP based on the pseudowire standby bit received from a TLDP peer.

Spoke-SDP FEC Templates and Filters

PW-templates are the way to configure the control word for this type of pseudowire as well as ingress/egress filters (ipv4/mac/ipv6). Filters are only supported on the T-PEs, because there is no provisioning of a pw-template (or Epipe at all) on the S-PEs.

```
*A:PE-1# configure
    service
        pw-template 1 create
            controlword
            egress
                filter ip 1
            exit
        exit


*A:PE-1# configure
    service
        epipe 2 customer 1 create
---snip---
            spoke-sdp-fec 22 fec 129 aii-type 2 create endpoint CORE
                standby-signaling-slave
                pw-template-bind 1
                saii-type2 65536:192.0.2.12:1
                taii-type2 65536:192.0.2.22:1
                no shutdown
            exit
```

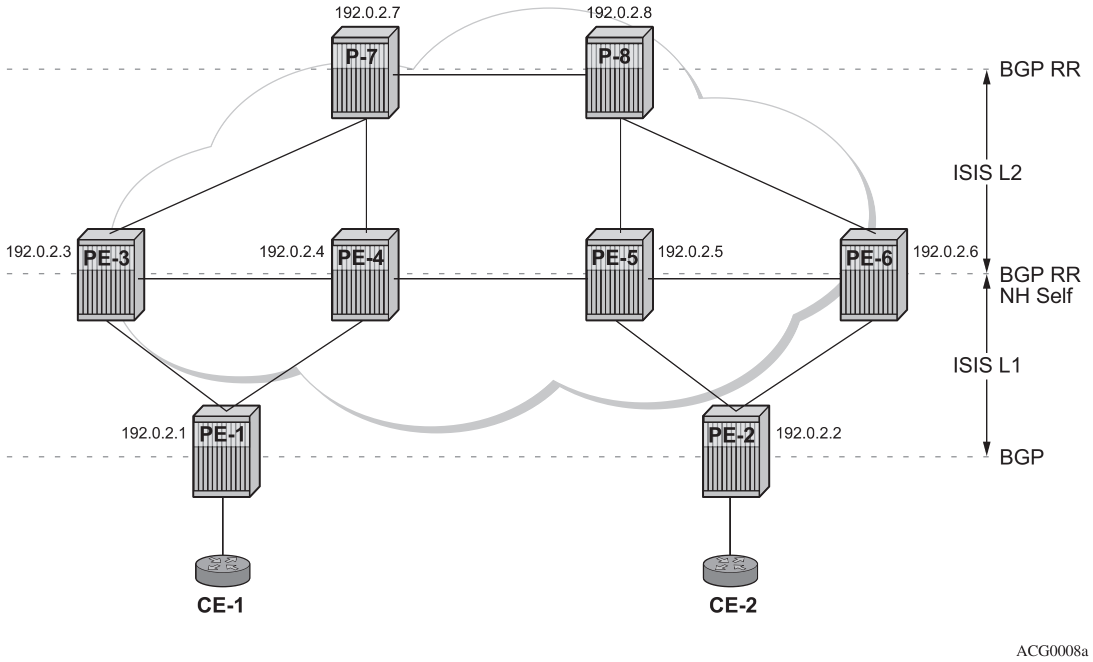
PW template changes (just like for VPLS with BGP-AD or BGP-VPLS) are not automatically propagated. A tools perform command is provided to evaluate and distribute the changes at the service level to one or all the services that use that template (if the service ID is omitted, then all the services will be updated).

```
*A:PE-1# tools perform service id 2 eval-pw-template 1 allow-service-impact
```

Intra-AS MS-PW Routing

This section provides a configuration example for an intra-AS scenario. The following example topology will be used for this section.

Figure 108 **Intra-AS MS-PW Example Topology**



Multiple MS-PW routing Epipes are to be configured between PE-1 and PE-2, with PE-3, PE-4, PE-5 and PE-6 being S-PE routers. P-7 and P-8 are pure P routers from a data plane perspective.

All the PEs are pre-configured with IS-IS as the IGP, as shown in Figure 108: PE-1 and PE-2 are level-1 routers, P-7 and P-8 are level-2 only routers and the rest of the routers are level-1/level-2. Link level LDP is also pre-configured on all the network interfaces and targeted LDP is configured between PE-1 and PE-3/PE-4, between PE-2 and PE-5/PE-6 and among PE-3, PE-4, PE-5, and PE-6. There is no targeted LDP sessions configured on P-7 and P-8.

As outlined in Figure 107, the configuration is a three-step process where the pw-routing context is configured first, then the required configuration so that routing tables get populated accordingly and finally the services themselves.

MS-PW Using BGP Routing

In this subsection, Epipe 2 will be configured between PE-1 and PE-2, where TLDP will use the BGP routes populated in the MS-PW routing table to signal the MS-PW.

The first step is the provisioning of the pw-routing context on all the T-PEs and S-PEs. The **spe-address** will be configured on all the T-PEs and S-PEs—all the routers except for P-7 and P-8—using the ASN as the global ID and the system address as the prefix. On PE-1 and PE-2, (only) the prefixes used for setting up Epipe 2 are configured. Two prefixes are configured per T-PE so that pseudowire redundancy with path diversity for the standby pseudowire can be carried out. The **spe-address** and local prefixes for the T-PEs are shown in the following CLI output. The **advertise-bgp** parameter is required because BGP is used here.

```
*A:PE-1# configure
    service
        pw-routing
            spe-address 65536:192.0.2.1
            local-prefix 65536:192.0.2.11 create
                advertise-bgp route-distinguisher 65536:11 community 65535:11
            exit
            local-prefix 65536:192.0.2.12 create
                advertise-bgp route-distinguisher 65536:12 community 65535:12
            exit
```

```
*A:PE-2# configure
    service
        pw-routing
            spe-address 65536:192.0.2.2
            local-prefix 65536:192.0.2.21 create
                advertise-bgp route-distinguisher 65536:21 community 65535:11
            exit
            local-prefix 65536:192.0.2.22 create
                advertise-bgp route-distinguisher 65536:22 community 65535:12
            exit
```

The second step is the configuration of BGP.

As shown in Figure 108, BGP is enabled in all the routers. The middle routers (PE-3, PE-4 and PE-5, PE-6) are BGP route-reflectors for PE-1 and PE-2 and they reflect MS-PW routes while changing the next-hop to their own system address. This is required so that TLDP knows where to send the label mapping message for a particular prefix. P-7 and P-8 are regular RRs reflecting routes among all the S-PEs. The BGP configuration of PE-1, PE-3, PE-4 and a P-7 is as follows. Similar commands are configured on the other PEs depending on their T-PE, S-PE or RR function.

The T-PEs have dual-homed BGP sessions to the S-PEs. Example for PE-1:

```
*A:PE-1# configure
    router
        policy-options
            begin
            policy-statement "export_ms-pw"
                entry 10
                    from
                        family ms-pw
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit

*A:PE-1# configure
    router
        autonomous-system 65536
        bgp
            enable-peer-tracking
            rapid-withdrawal
            group "region"
                family ms-pw
                export "export_ms-pw"
                peer-as 65536
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
            exit
        exit
    exit
```

The S-PEs are reflecting routes and also changing the NH and local preference based on the communities accordingly, so that pseudowire diversity can be ensured.

```
*A:PE-3# configure
    router
        policy-options
            begin
            community "65535:11" members "65535:11"
            community "65535:12" members "65535:12"
            policy-statement "export_ms-pw_ABR-to-core"
                entry 10
                    from
                        protocol bgp
                        community "65535:11"
                        family ms-pw
                    exit
                    action accept
                        local-preference 150
                        next-hop-self
                    exit
                exit
                entry 20
                    from
                        protocol bgp
                        community "65535:12"
                        family ms-pw
                    exit
                    action accept
                        local-preference 100
                        next-hop-self
                    exit
                exit
            exit
            policy-statement "export_ms-pw_ABR-to-region"
                entry 10
                    from
                        protocol bgp
                        community "65535:11"
                        family ms-pw
                    exit
                    action accept
                        local-preference 150
                        next-hop-self
                    exit
                exit
                entry 20
                    from
                        protocol bgp
                        community "65535:12"
                        family ms-pw
                    exit
                    action accept
                        local-preference 100
                        next-hop-self
                    exit
                exit
            exit
            commit
        exit
```

```
*A:PE-3# configure
    router
        autonomous-system 65536
        bgp
            rapid-withdrawal
            group "core"
                family ms-pw
                export "export_ms-pw_ABR-to-core"
                peer-as 65536
                neighbor 192.0.2.7
                exit
                neighbor 192.0.2.8
                exit
            exit
            group "region"
                family ms-pw
                cluster 3.3.3.3
                export "export_ms-pw_ABR-to-region"
                peer-as 65536
                enable-peer-tracking
                neighbor 192.0.2.1
                exit
            exit
```

The second S-PE to which PE-1 is connected has the following BGP configuration:

```
*A:PE-4# configure
    router
        policy-options
```

```
                begin
                community "65535:11" members "65535:11"
                community "65535:12" members "65535:12"
                policy-statement "export_ms-pw_ABR-to-core"
                    entry 10
                        from
                            protocol bgp
                            community "65535:12"
                            family ms-pw
                        exit
                        action accept
                            local-preference 150
                            next-hop-self
                        exit
                    exit
                    entry 20
                        from
                            protocol bgp
                            community "65535:11"
                            family ms-pw
                        exit
                        action accept
                            local-preference 100
                            next-hop-self
                        exit
                    exit
                exit
                policy-statement "export_ms-pw_ABR-to-region"
                    entry 10
                        from
                            protocol bgp
                            community "65535:12"
                            family ms-pw
                        exit
                        action accept
                            local-preference 150
                            next-hop-self
                        exit
                    exit
                    entry 20
                        from
                            protocol bgp
                            community "65535:11"
                            family ms-pw
                        exit
                        action accept
                            local-preference 100
                            next-hop-self
                        exit
                    exit
                exit
                commit
            exit
```

```
*A:PE-4# configure
    router
        autonomous-system 65536
        bgp
            rapid-withdrawal
            group "core"
                family ms-pw
                export "export_ms-pw_ABR-to-core"
                peer-as 65536
                neighbor 192.0.2.7
                exit
                neighbor 192.0.2.8
                exit
            exit
            group "region"
                family ms-pw
                cluster 4.4.4.4
                export "export_ms-pw_ABR-to-region"
                peer-as 65536
                enable-peer-tracking
                neighbor 192.0.2.1
                exit
            exit
        exit
    exit
```

The following is the BGP configuration for P-7 and P-8. These are pure RRs.

```
*A:P-7# configure
    router
        autonomous-system 65536
        bgp
            enable-peer-tracking
            rapid-withdrawal
            group "core"
                family ms-pw
                cluster 1.1.1.1
                peer-as 65536
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
                neighbor 192.0.2.5
                exit
                neighbor 192.0.2.6
                exit
            exit
        exit
```

After BGP is properly configured and the BGP update exchange takes place, the RIBs are properly populated and the required prefixes uploaded into the MS-PW routing table. An example for PE-1's RIB and pseudowire routing table is shown.

```
*A:PE-1# show router bgp routes ms-pw
===============================================================================
 BGP Router ID:192.0.2.1       AS:65536        Local AS:65536
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
```

```
===============================================================================
BGP MSPW Routes
===============================================================================
Flag  Network                 RD
      Nexthop                 AII-Type2/Preflen
      As-Path
-------------------------------------------------------------------------------
---snip---
u*>?  65536:192.0.2.21        65536:21
      192.0.2.3               65536:192.0.2.21:0/64
      No As-Path
*?    65536:192.0.2.21        65536:21
      192.0.2.4               65536:192.0.2.21:0/64
      No As-Path
u*>?  65536:192.0.2.22        65536:22
      192.0.2.4               65536:192.0.2.22:0/64
      No As-Path
*?    65536:192.0.2.22        65536:22
      192.0.2.3               65536:192.0.2.22:0/64
      No As-Path
---snip---
-------------------------------------------------------------------------------
```

```
*A:PE-1# show service pw-routing route-table

===============================================================================
Service PW L2 Routing Information
===============================================================================
AII-Type2/Prefix-Len                           Next-Hop       Owner  Age
 Route-Distinguisher                            Community      Best
-------------------------------------------------------------------------------
---snip---
65536:192.0.2.11:0/64                          192.0.2.1      local  01h03m37s
 65536:11                                       65535:11       yes
---snip---
65536:192.0.2.12:0/64                          192.0.2.1      local  01h03m37s
 65536:12                                       65535:12       yes
65536:192.0.2.21:0/64                          192.0.2.3      bgp    00h16m42s
 65536:21                                       65535:11       yes
65536:192.0.2.22:0/64                          192.0.2.4      bgp    00h17m31s
 65536:22                                       65535:12       yes
-------------------------------------------------------------------------------
```

The two prefixes advertised by PE-2 are properly learned by PE-1 through two different next hops. Now, use each one with a different pseudowire and make sure that the active and standby pseudowires follow different paths in the network.

Once the routes are installed in the MS-PW routing table, the services are configured on PE-1 and PE-2, as follows:

```
*A:PE-1# configure
    service
        pw-template 1 create
            controlword
        exit
        epipe 2 customer 1 create
```

```
            description "ms-pw epipe with bgp - using 2 prefixes"
            endpoint "CORE" create
                description "end-point for epipe A/S PW redundancy"
                revert-time 10
            exit
            sap 1/1/4:2 create
            exit
            spoke-sdp-fec 21 fec 129 aii-type 2 create endpoint CORE
                precedence primary
                pw-template-bind 1
                saii-type2 65536:192.0.2.11:1
                taii-type2 65536:192.0.2.21:1
                no shutdown
            exit
            spoke-sdp-fec 22 fec 129 aii-type 2 create endpoint CORE
                pw-template-bind 1
                saii-type2 65536:192.0.2.12:1
                taii-type2 65536:192.0.2.22:1
                no shutdown
            exit
            no shutdown
        exit
```

```
*A:PE-2# configure
    service
        pw-template 1 create
            controlword
        exit
        epipe 2 customer 1 create
            description "ms-pw epipe with bgp - using 2 prefixes"
            endpoint "CORE" create
                description "end-point for epipe A/S PW redundancy"
                revert-time 10
            exit
            sap 1/1/4:2 create
            exit
            spoke-sdp-fec 21 fec 129 aii-type 2 create endpoint CORE
                precedence primary
                pw-template-bind 1
                saii-type2 65536:192.0.2.21:1
                taii-type2 65536:192.0.2.11:1
                no shutdown
            exit
            spoke-sdp-fec 22 fec 129 aii-type 2 create endpoint CORE
                pw-template-bind 1
                saii-type2 65536:192.0.2.22:1
                taii-type2 65536:192.0.2.12:1
                no shutdown
            exit
            no shutdown
        exit
```

The following command can be executed to verify that the service and spoke-sdp-fecs are up:

```
*A:PE-1# show service id 2 base
```

```
===============================================================================
Service Basic Information
===============================================================================
Service Id        : 2                   Vpn Id            : 0
Service Type      : Epipe
Name              : (Not Specified)
Description       : ms-pw epipe with bgp - using 2 prefixes
Customer Id       : 1                   Creation Origin   : manual
Last Status Change: 04/28/2017 12:27:05
Last Mgmt Change  : 04/28/2017 12:27:05
Test Service      : No
Admin State       : Up                  Oper State        : Up
MTU               : 1514
Vc Switching      : False
SAP Count         : 1                   SDP Bind Count    : 2
Per Svc Hashing   : Disabled
Vxlan Src Tep Ip  : N/A
Force QTag Fwd    : Disabled

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/4:2                              q-tag        1518    1518    Up   Up
sdp:17405:4294967292 SB(192.0.2.4)      MS-PW        0       1556    Up   Up
sdp:17406:4294967293 SB(192.0.2.3)      MS-PW        0       1556    Up   Up
===============================================================================
*A:PE-1#
```

The SDP-binding identifiers and SDP identifiers are automatically generated by the system.

Use **vccv-trace** to check that the spoke-SDP FECs for the active and standby pseudowires follow different and disjoint paths:

```
*A:PE-1# oam vccv-trace spoke-sdp-fec 21
VCCV-TRACE  with 120 bytes of MPLS payload
1  192.0.2.3  rtt=1.75ms rc=8(DSRtrMatchLabel)
2  192.0.2.5  rtt=5.38ms rc=8(DSRtrMatchLabel)
3  192.0.2.2  rtt=8.34ms rc=3(EgressRtr)
```

```
*A:PE-1# oam vccv-trace spoke-sdp-fec 22
VCCV-TRACE  with 120 bytes of MPLS payload
1  192.0.2.4  rtt=1.80ms rc=8(DSRtrMatchLabel)
2  192.0.2.6  rtt=5.48ms rc=8(DSRtrMatchLabel)
3  192.0.2.2  rtt=7.83ms rc=3(EgressRtr)
```

MS-PW using Static Routing

In this subsection, Epipe 3 will be configured between PE-1 and PE-2, where TLDP will use static-routes in the MS-PW routing table to signal the MS-PW.

On PE-1 and PE-2 (only), the prefixes used for setting up Epipe 3 are configured. These prefixes could be the same as used for Epipe 2, however different prefixes are used in this example. The **no advertise-bgp** parameter is required now. The static routes for each remote prefix are also configured. Because we will also have pseudowire redundancy for Epipe 3, two prefixes with static-routes pointing at different next-hops will be used:

```
*A:PE-1# configure
    service
        pw-routing
            spe-address 65536:192.0.2.1
            local-prefix 65536:192.0.2.13 create
            exit
            local-prefix 65536:192.0.2.14 create
            exit
            static-route 65536:192.0.2.23:192.0.2.3
            static-route 65536:192.0.2.24:192.0.2.4
```

```
*A:PE-2# configure
    service
        pw-routing
            spe-address 65536:192.0.2.2
            local-prefix 65536:192.0.2.23 create
            exit
            local-prefix 65536:192.0.2.24 create
            exit
            static-route 65536:192.0.2.13:192.0.2.5
            static-route 65536:192.0.2.14:192.0.2.6
```

Static-routes are also required at all S-PEs along the path (keeping the path diversity for the prefixes as well) and for both directions:

```
*A:PE-3# configure
    service
        pw-routing
            spe-address 65536:192.0.2.3
            static-route 65536:192.0.2.13:192.0.2.1
            static-route 65536:192.0.2.23:192.0.2.5
```

```
*A:PE-4# configure
    service
        pw-routing
            spe-address 65536:192.0.2.4
            static-route 65536:192.0.2.14:192.0.2.1
            static-route 65536:192.0.2.24:192.0.2.6
```

Finally, once the MS-PW routing tables are properly populated, the services can be configured and brought up:

```
*A:PE-1# configure
    service
        pw-template 1 create
            controlword
        exit
```

```
            epipe 3 customer 1 create
                description "ms-pw epipe with static routes"
                endpoint "CORE" create
                    description "end-point for epipe A/S PW redundancy"
                    revert-time 10
                    standby-signaling-master
                exit
                sap 1/1/4:3 create
                exit
                spoke-sdp-fec 31 fec 129 aii-type 2 create endpoint CORE
                    precedence primary
                    pw-template-bind 1
                    saii-type2 65536:192.0.2.13:31
                    taii-type2 65536:192.0.2.23:31
                    no shutdown
                exit
                spoke-sdp-fec 32 fec 129 aii-type 2 create endpoint CORE
                    pw-template-bind 1
                    saii-type2 65536:192.0.2.14:32
                    taii-type2 65536:192.0.2.24:32
                    no shutdown
                exit
                no shutdown
            exit

    *A:PE-2# configure
        service
            pw-template 1 create
                controlword
            exit
            epipe 3 customer 1 create
                description "ms-pw epipe with static routes"
                endpoint "CORE" create
                    description "end-point for epipe A/S PW redundancy"
                    revert-time 10
                    standby-signaling-master
                exit
                sap 1/1/4:3 create
                exit
                spoke-sdp-fec 31 fec 129 aii-type 2 create endpoint CORE
                    precedence primary
                    pw-template-bind 1
                    saii-type2 65536:192.0.2.23:31
                    taii-type2 65536:192.0.2.13:31
                    no shutdown
                exit
                spoke-sdp-fec 32 fec 129 aii-type 2 create endpoint CORE
                    pw-template-bind 1
                    saii-type2 65536:192.0.2.24:32
                    taii-type2 65536:192.0.2.14:32
                    no shutdown
                exit
                no shutdown
            exit
```

Check the status and path of the spoke-sdp-fecs with the proper show commands and oam vccv-trace/ping commands (see previous subsection MS-PW Using BGP Routing).

MS-PW using Explicit Paths

In this subsection, Epipe 4 will be configured between PE-1 and PE-2, where TLDP will use explicit paths to signal the MS-PW, overriding the information given by the MS-PW routing table. Although this mode requires the specific configuration of the hops, one by one, the configuration is only done on the T-PEs, as opposed to the static-routes where all the S-PEs must be configured with static routes (a mixed of static-routes and BGP routes can coexist). The local-prefixes shown for Epipe 3 will be re-used here for Epipe 4.

Now path-1 and path-2 will be configured hop by hop, using diverse paths. All the S-PE nodes as well as the terminating T-PE must be included in the path.

```
*A:PE-1# configure
    service
        pw-routing
            spe-address 65536:192.0.2.1
            local-prefix 65536:192.0.2.13 create
            exit
            local-prefix 65536:192.0.2.14 create
            exit
            path "path-1" create
                hop 1 192.0.2.3
                hop 2 192.0.2.5
                hop 3 192.0.2.2
                no shutdown
            exit
            path "path-2" create
                hop 1 192.0.2.4
                hop 2 192.0.2.6
                hop 3 192.0.2.2
                no shutdown
            exit
        exit


*A:PE-2# configure
    service
        pw-routing
            spe-address 65536:192.0.2.2
            local-prefix 65536:192.0.2.23 create
            exit
            local-prefix 65536:192.0.2.24 create
            exit
            path "path-1" create
                hop 1 192.0.2.5
                hop 2 192.0.2.3
                hop 3 192.0.2.1
                no shutdown
            exit
            path "path-2" create
                hop 1 192.0.2.6
                hop 2 192.0.2.4
                hop 3 192.0.2.1
                no shutdown
            exit
        exit
```

Those paths must be specified when configuring the Epipe:

```
*A:PE-1# configure
    service
        epipe 4 customer 1 create
            description "ms-pw epipe with explicit paths"
            endpoint "CORE" create
                description "end-point for epipe A/S PW redundancy"
                revert-time 10
                standby-signaling-master
            exit
            sap 1/1/4:4 create
            exit
            spoke-sdp-fec 41 fec 129 aii-type 2 create endpoint CORE
                precedence primary
                saii-type2 65536:192.0.2.13:41
                taii-type2 65536:192.0.2.23:41
                path "path-1"
                no shutdown
            exit
            spoke-sdp-fec 42 fec 129 aii-type 2 create endpoint CORE
                saii-type2 65536:192.0.2.14:42
                taii-type2 65536:192.0.2.24:42
                path "path-2"
                no shutdown
            exit
            no shutdown
        exit
```

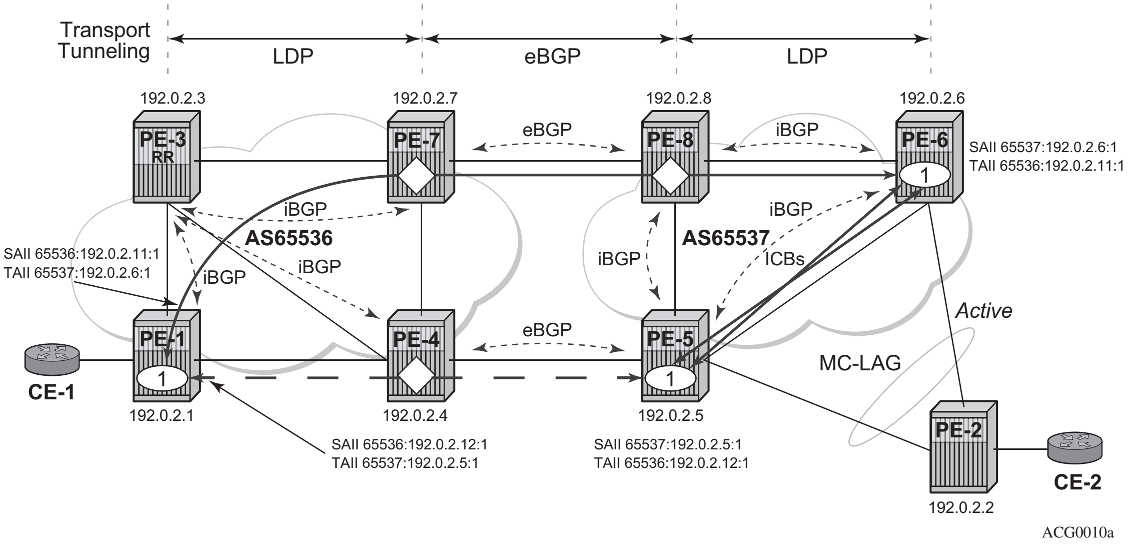
```
*A:PE-2# configure
    service
        epipe 4 customer 1 create
            description "ms-pw epipe with explicit paths"
            endpoint "CORE" create
                description "end-point for epipe A/S PW redundancy"
                revert-time 10
            exit
            sap 1/1/4:4 create
            exit
            spoke-sdp-fec 41 fec 129 aii-type 2 create endpoint CORE
                precedence primary
                saii-type2 65536:192.0.2.23:41
                taii-type2 65536:192.0.2.13:41
                path "path-1"
                no shutdown
            exit
            spoke-sdp-fec 42 fec 129 aii-type 2 create endpoint CORE
                saii-type2 65536:192.0.2.24:42
                taii-type2 65536:192.0.2.14:42
                path "path-2"
                no shutdown
            exit
            no shutdown
        exit
```

Verify the status and path of the spoke-sdp-fecs with the proper **show** commands and **oam vccv-trace/ping** commands (see subsection MS-PW Using BGP Routing).

Inter-AS MS-PW Routing

This configuration example for an inter-AS scenario uses BGP tunnels between ASBRs and BGP as the MS-PW routing mechanism. Figure 109 shows the example topology used in this section.

Figure 109 **Inter-AS MS-PW Example Topology**



In this example, only one Epipe is configured (Epipe 1, using MS-PW BGP routing). The T-PEs are PE-1, PE-5, and PE-6; the S-PEs are PE-7, PE-8, and PE-4.

A/S pseudowire redundancy together with MC-LAG at one end will be used, as shown in Figure 109. Inter-Chassis Backup (ICB) spoke SDPs between PE-5 and PE-6 are required in order to forward the in-flight packets while MC-LAG and A/S pseudowire are converging, in case of network failures. Those ICBs will also be signaled following the MS-PW routing procedures.

The example topology in Figure 109 is pre-configured with the following settings:

- There are two ASs (65536 and 65537) which are connected by two ASBR pairs (PE-7/PE-4 and PE-8/PE-5) running eBGP between them. These eBGP sessions will be used to exchange ipv4-labels (to set up the transport BGP-LBL tunnel, according to the RFC 3107) and MS-PW NLRIs.
- Within AS65536, PE-3 is used as an RR to reflect the MS-PW routes. In AS65537 there is a full mesh of iBGP sessions to distribute the MS-PW routes.
- IS-IS is used within each AS.
- LDP is used as a transport MPLS signaling protocol within each AS and a BGP tunnel will be used between the ASBRs (MS-PW routing supports LDP or BGP tunnels as transport).
- A redundant MC-LAG access to PE-6 and PE-5 is configured.

The next section will go through the configuration required to set up a redundant Epipe between CE-1 and CE-2, by combining A/S pseudowire in the network and MC-LAG at the access.

MS-PW using BGP Routing

Epipe 1 will be configured at the end of this section, including the active and redundant pseudowires from PE-1 to PE-5/PE-6, as well as the required ICBs and SAPs at the access.

As discussed, the first step is the provisioning of the pw-routing context. Again, the **spe-address** must be provisioned in all T-PEs and S-PEs whereas prefixes are mandatory only on the T-PEs involved in the service. The following shows the prefixes configured on PE-1, PE-5, and PE-6. Two prefixes are needed in PE-1 in order to make sure that active and standby pseudowires follow disjoint paths.

```
*A:PE-1# configure
    service
        pw-routing
            spe-address 65536:192.0.2.1
            local-prefix 65536:192.0.2.11 create
                advertise-bgp route-distinguisher 65536:11 community 65535:11
            exit
            local-prefix 65536:192.0.2.12 create
                advertise-bgp route-distinguisher 65536:12 community 65535:12
            exit
```

```
*A:PE-5# configure
    service
        pw-routing
            spe-address 65537:192.0.2.5
            local-prefix 65537:192.0.2.5 create
                advertise-bgp route-distinguisher 65537:5 community 65535:5
            exit
```

```
*A:PE-6# configure
    service
        pw-routing
            spe-address 65537:192.0.2.6
            local-prefix 65537:192.0.2.6 create
                advertise-bgp route-distinguisher 65537:6 community 65535:6
            exit
```

Once the spe-addresses and prefixes have been provisioned, BGP must be configured accordingly. A simple BGP export-policy is used to export all the local MS-PW prefixes. The configuration on PE-1 is as follows:

```
*A:PE-1# configure
    router
        policy-options
            begin
            policy-statement "export_ms-pw"
                entry 10
                    from
                        family ms-pw
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit
```

```
*A:PE-1# configure
    router
        autonomous-system 65536
        bgp
            min-route-advertisement 1
            rapid-withdrawal
            group "intra-AS"
                family ms-pw
                export "export_ms-pw"
                peer-as 65536
                neighbor 192.0.2.3
                exit
            exit
            no shutdown
        exit
    exit
```

The configuration on PE-6 is as follows:

```
*A:PE-6# configure
    router
        policy-options
            begin
            policy-statement "export_ms-pw"
                entry 10
                    from
                        family ms-pw
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit
```

```
*A:PE-6# configure
    router
        autonomous-system 65537
        bgp
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            group "intra-AS"
                family ms-pw
                export "export_ms-pw"
                peer-as 65537
                neighbor 192.0.2.5
                exit
                neighbor 192.0.2.8
                exit
            exit
            no shutdown
        exit
    exit
```

At the ASBR, the BGP policies are more complex because the following tasks must be accomplished:

- ASBR IPv4 system addresses must be exported to the peer ASBR to establish the RFC 3107 BGP tunnel between ASBRs.
- BGP export policies must be used so that MS-PW NLRI exchange can be controlled and attributes like MED (towards the remote AS) and/or local-preference (towards the local AS) can be modified.
- Finally, BGP import policies must also be used to modify the MS-PW route NH (next-hops) because the TLDP next signaling hop must match a peer TLDP system address.

The prefixes 65536:192.0.2.11 and 65537:192.0.2.6 must be preferred in the PE-7/PE-8 pair whereas the prefixes 65536:192.0.2.12 and 65537:192.0.2.5 must be preferred in the PE-4/PE-5 pair, so that the pseudowires are established as depicted in Figure 109. The preference can be propagated by using the BGP MED (use the Local Preference (LP) within the AS (LP is not relevant to eBGP)). The following CLI excerpt shows an example of how to modify MED and LP, as well as changing the NH with an import policy. The configuration on ASBR PE-4 is as follows:

```
*A:PE-4# configure
    router
        policy-options
            begin
            prefix-list "system"
    prefix 192.0.2.4/32 exact
exit
community "65535:5" members "65535:5"
community "65535:6" members "65535:6"
community "65535:11" members "65535:11"
community "65535:12" members "65535:12"
policy-statement "ASBR to ASBR"
    entry 10
        from
            protocol bgp
            community "65535:12"
            family ms-pw
        exit
        action accept
            origin igp
            metric set 50
        exit
    exit
    entry 20
        from
            protocol bgp
            community "65535:11"
            family ms-pw
        exit
        action accept
            origin igp
            metric set 100
        exit
    exit
exit
policy-statement "ASBR to region"
    entry 10
        from
            protocol bgp
            community "65535:5"
            family ms-pw
        exit
        action accept
            origin igp
            local-preference 150
            next-hop-self
        exit
    exit
    entry 20
        from
            protocol bgp
            community "65535:6"
            family ms-pw
        exit
        action accept
            origin igp
            next-hop-self
        exit
    exit
exit
policy-statement "export_ipv4_system"
    entry 10
        from
            prefix-list "system"
```

```
                    exit
                    action accept
                        origin igp
                    exit
                exit
            exit
            policy-statement "import ms-pw NH change"
                entry 10
                    from
                        protocol bgp
                        family ms-pw
                    exit
                    action accept
                        next-hop 192.0.2.5
                    exit
                exit
            exit
            commit
        exit

*A:PE-4# configure
    router
        autonomous-system 65536
        bgp
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            group "inter-AS"
                family ms-pw label-ipv4
                import "import ms-pw NH change"
                export "export_ipv4_system" "ASBR to ASBR"
                local-as 65536
                peer-as 65537
                neighbor 192.168.45.2
                exit
            exit
            group "intra-AS"
                family ms-pw
                export "ASBR to region"
                peer-as 65536
                neighbor 192.0.2.3
                exit
            exit
            no shutdown
        exit
    exit
```

The configuration on ASBR PE-7 is as follows:

```
*A:PE-7# configure
    router
        policy-options
            begin
            prefix-list "system"
                prefix 192.0.2.7/32 exact
            exit
            community "65535:5" members "65535:5"
            community "65535:6" members "65535:6"
            community "65535:11" members "65535:11"
            community "65535:12" members "65535:12"
            policy-statement "ASBR to ASBR"
                entry 10
                    from
                        protocol bgp
                        community "65535:11"
                        family ms-pw
                    exit
                    action accept
                        origin igp
                        metric set 50
                    exit
                exit
                entry 20
                    from
                        protocol bgp
                        community "65535:12"
                        family ms-pw
                    exit
                    action accept
                        origin igp
                        metric set 100
                    exit
                exit
            exit
            policy-statement "ASBR to region"
                entry 10
                    from
                        protocol bgp
                        community "65535:6"
                        family ms-pw
                    exit
                    action accept
                        origin igp
                        local-preference 150
                        next-hop-self
                    exit
                exit
                entry 20
                    from
                        protocol bgp
                        community "65535:5"
                        family ms-pw
                    exit
                    action accept
                        origin igp
                        next-hop-self
                    exit
                exit
            exit
            policy-statement "export_ipv4_system"
                entry 10
                    from
                        prefix-list "system"
                    exit
                    action accept
                        origin igp
                    exit
```

```
                exit
            exit
            policy-statement "import ms-pw NH change"
                entry 10
                    from
                        protocol bgp
                        family ms-pw
                    exit
                    action accept
                        next-hop 192.0.2.8
                    exit
                exit
            exit
            commit
        exit
```

```
*A:PE-7# configure
    router
       autonomous-system 65536
       bgp
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            group "inter-AS"
                family ms-pw label-ipv4
                import "import ms-pw NH change"
                export "export_ipv4_system" "ASBR to ASBR"
                local-as 65536
                peer-as 65537
                neighbor 192.168.78.2
                exit
            exit
            group "intra-AS"
                family ms-pw
                export "ASBR to region"
                peer-as 65536
                neighbor 192.0.2.3
                exit
            exit
            no shutdown
        exit
    exit
```

PE-5 and PE-8 have similar configurations to the ones shown. However, PE-5 is a T-PE as well as an ASBR, therefore a local MS-PW prefix must be exported as opposed to only remote prefixes (that is, some export entries for the local MS-PW routes will not contain **protocol bgp** in the matching criteria).

After BGP is properly configured and the updates get exchanged, the RIBs are populated and the prefixes uploaded onto the MS-PW routing table as shown for PE-1 in the following output:

```
*A:PE-1# show router bgp routes ms-pw
===============================================================================
 BGP Router ID:192.0.2.1       AS:65536       Local AS:65536
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MSPW Routes
===============================================================================
Flag  Network                 RD
      Nexthop                 AII-Type2/Preflen
      As-Path
-------------------------------------------------------------------------------
---snip---
u*>i  65537:192.0.2.5         65537:5
      192.0.2.4               65537:192.0.2.5:0/64
      65537
u*>i  65537:192.0.2.6         65537:6
      192.0.2.7               65537:192.0.2.6:0/64
      65537
-------------------------------------------------------------------------------
Routes : 4

*A:PE-1# show service pw-routing route-table

===============================================================================
Service PW L2 Routing Information
===============================================================================
AII-Type2/Prefix-Len                           Next-Hop       Owner  Age
 Route-Distinguisher                            Community      Best
-------------------------------------------------------------------------------
65536:192.0.2.11:0/64                          192.0.2.1      local  30d22h25m
 0:0                                            0:0            yes
65536:192.0.2.11:0/64                          192.0.2.1      local  30d22h24m
 65536:11                                       65535:11       yes
65536:192.0.2.12:0/64                          192.0.2.1      local  30d22h19m
 0:0                                            0:0            yes
65536:192.0.2.12:0/64                          192.0.2.1      local  30d22h19m
 65536:12                                       65535:12       yes
65537:192.0.2.5:0/64                           192.0.2.4      bgp    02h43m25s
 65537:5                                        65535:5        yes
65537:192.0.2.6:0/64                           192.0.2.7      bgp    02h45m49s
 65537:6                                        65535:6        yes
-------------------------------------------------------------------------------
Entries found: 6
===============================================================================
*A:PE-1#
```

For PE-6:

```
*A:PE-6# show router bgp routes ms-pw
===============================================================================
 BGP Router ID:192.0.2.6        AS:65537       Local AS:65537
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
```

```
===============================================================================
BGP MSPW Routes
===============================================================================
Flag  Network                 RD
      Nexthop                 AII-Type2/Preflen
      As-Path
-------------------------------------------------------------------------------
u*>i  65536:192.0.2.11        65536:11
      192.0.2.8               65536:192.0.2.11:0/64
      65536
u*>i  65536:192.0.2.12        65536:12
      192.0.2.5               65536:192.0.2.12:0/64
      65536
u*>i  65537:192.0.2.5         65537:5
      192.0.2.5               65537:192.0.2.5:0/64
      No As-Path
-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:PE-6#
```

```
*A:PE-6# show service pw-routing route-table

===============================================================================
Service PW L2 Routing Information
===============================================================================
AII-Type2/Prefix-Len                           Next-Hop       Owner  Age
 Route-Distinguisher                            Community      Best
-------------------------------------------------------------------------------
65536:192.0.2.11:0/64                          192.0.2.8      bgp    02h52m08s
 65536:11                                       65535:11       yes
65536:192.0.2.12:0/64                          192.0.2.5      bgp    02h38m51s
 65536:12                                       65535:12       yes
65537:192.0.2.5:0/64                           192.0.2.5      bgp    02h38m51s
 65537:5                                        65535:5        yes
65537:192.0.2.6:0/64                           192.0.2.6      local  28d02h16m
 0:0                                            0:0            yes
65537:192.0.2.6:0/64                           192.0.2.6      local  28d02h16m
 65537:6                                        65535:6        yes
-------------------------------------------------------------------------------
Entries found: 5
===============================================================================
*A:PE-6#
```

As can be seen in the preceding show commands on PE-6, the two PE-1 prefixes are learned on (PE-5 and) PE-6 through different and disjoint paths. On PE-1, the PE-5 and PE-6 prefixes are learned through two different and disjoint paths.

After configuring the PW routing context and configuring BGP, the last step is the service configuration on the three T-PEs, as follows. TLDP sessions must have been previously and explicitly configured between the T-PEs and S-PEs (between PE-1 and PE-4/7, between PE-4 and PE-5, PE-7 and PE-8, and between PE-6 and PE-5/8).

```
*A:PE-1# configure
    router
```

```
        ldp
            targeted-session
                peer 192.0.2.4
                exit
                peer 192.0.2.7
                exit
            exit

*A:PE-1# configure
    service
        pw-template 1 create
            controlword
        exit
        epipe 1 customer 1 create
            description "ms-pw epipe with bgp, inter-AS, MC-LAG redundancy"
            endpoint "CORE" create
                description "end-point for epipe A/S PW redundancy"
            exit
            sap 1/1/4:1 create
            exit
            spoke-sdp-fec 11 fec 129 aii-type 2 create endpoint CORE
                precedence primary
                pw-template-bind 1
                saii-type2 65536:192.0.2.11:1
                taii-type2 65537:192.0.2.6:1
                no shutdown
            exit
            spoke-sdp-fec 12 fec 129 aii-type 2 create endpoint CORE
                pw-template-bind 1
                saii-type2 65536:192.0.2.12:1
                taii-type2 65537:192.0.2.5:1
                no shutdown
            exit
            no shutdown
        exit

*A:PE-5# configure
    service
        pw-template 1 create
            controlword
        exit
        epipe 1 customer 1 create
            description "ms-pw epipe with bgp, inter-AS, MC-LAG redundancy"
            endpoint "CORE" create
                description "end-point for epipe A/S PW redundancy"
            exit
            endpoint "ACCESS" create
            exit
            sap lag-1:1 endpoint "ACCESS" create
            exit
            spoke-sdp-fec 11 fec 129 aii-type 2 create endpoint CORE
                pw-template-bind 1
                saii-type2 65537:192.0.2.5:1
                taii-type2 65536:192.0.2.12:1
                no shutdown
            exit
            spoke-sdp-fec 12 fec 129 aii-type 2 create endpoint CORE icb
                pw-template-bind 1
                saii-type2 65537:192.0.2.5:2
                taii-type2 65537:192.0.2.6:2
                no shutdown
            exit
            spoke-sdp-fec 13 fec 129 aii-type 2 create endpoint ACCESS icb
                pw-template-bind 1
                saii-type2 65537:192.0.2.5:3
                taii-type2 65537:192.0.2.6:3
                no shutdown
            exit
            no shutdown
        exit

*A:PE-6# configure
    service
        pw-template 1 create
            controlword
        exit
        epipe 1 customer 1 create
            description "ms-pw epipe with bgp, inter-AS, MC-LAG redundancy"
            endpoint "CORE" create
                description "end-point for epipe A/S PW redundancy"
            exit
            endpoint "ACCESS" create
            exit
            sap lag-1:1 endpoint "ACCESS" create
            exit
            spoke-sdp-fec 11 fec 129 aii-type 2 create endpoint CORE
                pw-template-bind 1
                saii-type2 65537:192.0.2.6:1
                taii-type2 65536:192.0.2.11:1
                no shutdown
            exit
            spoke-sdp-fec 12 fec 129 aii-type 2 create endpoint CORE icb
                pw-template-bind 1
                saii-type2 65537:192.0.2.6:3
                taii-type2 65537:192.0.2.5:3
                no shutdown
            exit
            spoke-sdp-fec 13 fec 129 aii-type 2 create endpoint ACCESS icb
                pw-template-bind 1
                saii-type2 65537:192.0.2.6:2
                taii-type2 65537:192.0.2.5:2
                no shutdown
            exit
            no shutdown
        exit
```

The following show commands can be executed to check the status of the Epipe 1 and the pseudowire status signaling received:

```
*A:PE-1# show service id 1 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 1                   Vpn Id            : 0
Service Type      : Epipe
Name              : (Not Specified)
Description       : ms-pw epipe with bgp, inter-AS, MC-LAG redundancy
Customer Id       : 1                    Creation Origin   : manual
Last Status Change: 04/28/2017 09:52:53
Last Mgmt Change  : 04/28/2017 09:56:07
Test Service      : No
Admin State       : Up                   Oper State        : Up
MTU               : 1514
Vc Switching      : False
SAP Count         : 1                    SDP Bind Count    : 2
Per Svc Hashing   : Disabled
Force QTag Fwd    : Disabled
-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/4:1                              q-tag        1518    1518    Up   Up
sdp:17406:4294967294 SB(192.0.2.7)       MS-PW        0       9190    Up   Up
sdp:17407:4294967295 SB(192.0.2.4)       MS-PW        0       9190    Up   Up
===============================================================================
```

```
*A:PE-1# show service id 1 endpoint

===============================================================================
Service 1 endpoints
===============================================================================
Endpoint name                : CORE
Description                  : end-point for epipe A/S PW redundancy
Creation Origin              : manual
Revert time                  : 0
Act Hold Delay               : 0
Standby Signaling Master     : false
Standby Signaling Slave      : false
Tx Active (SDP-FEC)          : 11
Tx Active Up Time            : 0d 00:04:59
Revert Time Count Down       : N/A
Tx Active Change Count       : 2
Last Tx Active Change        : 04/28/2017 09:56:07
-------------------------------------------------------------------------------
Members
-------------------------------------------------------------------------------
Sdp-fec: 11 Prec:0                              Oper Status: Up
Sdp-fec: 12 Prec:4                              Oper Status: Up
===============================================================================
```

PE-5 will have the MC-LAG standby interface and as such the SAP will be operationally down and will drive the standby signaling to the remote T-PEs:

```
*A:PE-5# show service id 1 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 1                    Vpn Id            : 0
Service Type      : Epipe
```

```
Name              : (Not Specified)
Description       : ms-pw epipe with bgp, inter-AS, MC-LAG redundancy
Customer Id       : 1                   Creation Origin   : manual
Last Status Change: 04/28/2017 09:52:58
Last Mgmt Change  : 04/28/2017 09:52:58
Test Service      : No
Admin State       : Up                  Oper State        : Up
MTU               : 1514
Vc Switching      : False
SAP Count         : 1                   SDP Bind Count    : 3
Per Svc Hashing   : Disabled
Force QTag Fwd    : Disabled

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:lag-1:1                              q-tag        1518    1518    Up   Down
sdp:17406:4294967293 SB(192.0.2.6)       MS-PW        0       9190    Up   Up
sdp:17406:4294967294 SB(192.0.2.6)       MS-PW        0       9190    Up   Up
sdp:17407:4294967295 SB(192.0.2.4)       MS-PW        0       9190    Up   Up
===============================================================================

*A:PE-5# show service id 1 all | match Flag
Flags              : None
Flags              : None
Flags              : None
Flags              : PortOperDown StandByForMcProtocol
```

The following commands are useful on the S-PEs in order to find the PWs automatically created as well as the SDPs automatically used for those PWs.

```
*A:PE-7# show service sdp-using

===============================================================================
SDP Using
===============================================================================
SvcId      SdpId              Type   Far End              Opr   I.Label E.Label
                                                          State
-------------------------------------------------------------------------------
2147483647 17406:4294967294   MS-PW  192.0.2.1            Up    262136  262138
2147483647 17407:4294967295   MS-PW  192.0.2.8            Up    262137  262137
-------------------------------------------------------------------------------
Number of SDPs : 2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-7#
```

As it can be seen in the preceding output, two PWs (type MS-PW) have been automatically created over two also automatically created SDPs: 17406 and 17407. SDP 17406 is built over an LDP tunnel whereas SDP 17407 runs over a BGP tunnel.

```
*A:PE-7# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
================================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
--------------------------------------------------------------------------------
192.0.2.1/32      sdp       MPLS  17406     5        192.0.2.1      0
192.0.2.1/32      ldp       MPLS  65539     9        192.168.37.1   20
192.0.2.3/32      ldp       MPLS  65538     9        192.168.37.1   10
192.0.2.4/32      ldp       MPLS  65537     9        192.168.47.1   10
192.0.2.8/32      sdp       MPLS  17407     5        192.0.2.8      0
192.0.2.8/32      bgp       MPLS  262145    12       192.168.78.2   1000
--------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
================================================================================
*A:PE-7#

*A:PE-7# show service sdp 17406 detail | match "Active LSP"
Mixed LSP Mode       : Enabled               Active LSP Type    : LDP

*A:PE-7# show service sdp 17407 detail | match "Active LSP"
Mixed LSP Mode       : Enabled               Active LSP Type    : BGP
```

In addition to all of the recommended show commands, **vccv-ping** and **vccv-trace** are two extremely useful commands in this environment. **vccv-trace** can even help to trace the traffic going through the ICBs under failure situations.

Conclusion

Service Providers are always seeking highly scalable VLL services that can be deployed with the lowest operational cost. The SR OS supports MS-PW routing according to the draft-ietf-pwe3-dynamic-ms-pw. MS-PW routing allows the Service Provider to deploy Epipe services without having to provision services in the core of the network. In other words, MS-PW enables end-point provisioning in highly scalable seamless MPLS networks, through the use of BGP. Alternatively, static MS-PW routes or explicit paths can also be used.

The examples used in this chapter illustrate the configuration of MS-PW routing in intra-AS and inter-AS scenarios. Show and OAM commands have also been suggested so that the operator can verify and troubleshoot the MS-PW routing paths and procedures.

P2MP mLDP Tunnels for BUM Traffic in EVPN-MPLS Services

This chapter provides information about P2MP mLDP Tunnels for BUM Traffic in EVPN-MPLS Services.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SR OS Release 14.0.R4, but the CLI in the current edition is based on SR OS Release 15.0.R2.

Point-to-Multipoint (P2MP) multicast Label Distribution Protocol (mLDP) tunnels for Broadcast, Unknown unicast, and Multicast (BUM) traffic in EVPN-MPLS networks are supported in SR OS Release 14.0.R1, and later. IGMP-snooping support for EVPN-MPLS services is supported in SR OS Release 14.0.R4, and later.

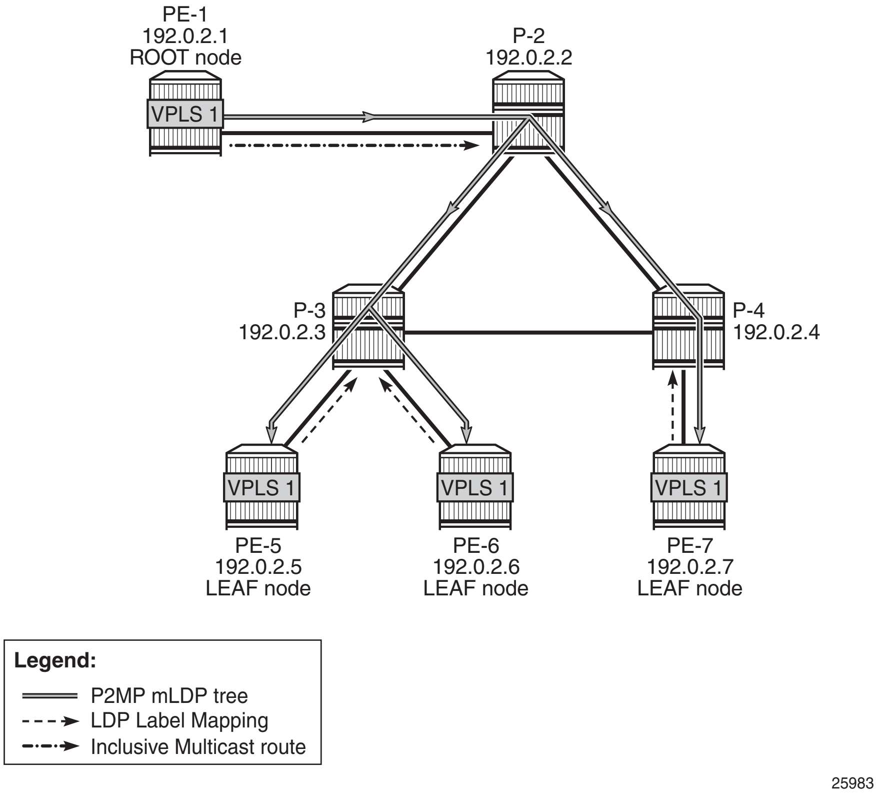
Overview

Service providers are moving their existing VPN services to EVPN. Providers using P2MP LSPs for VPLS services expect the same capabilities in EVPN. Before SR OS release 14.0.R1, only Ingress Replication (IR) was supported. This works well for broadcast and unknown unicast traffic, but it is inefficient for multicast. Ingress replication does not use a multicast mechanism. Instead, the parent node makes n individual copies and unicasts each copy through an MPLS or IP tunnel to each child node.

BUM traffic will be sent from a root node to a number of leaf nodes, but leaf nodes are also allowed to send BUM traffic to root nodes. If most BUM traffic is flowing from a few root nodes to leaf nodes, it would be inefficient to promote all leaf nodes to root-and-leaf nodes due to the amount of P2MP tunnels that would need to be set up. Another solution is to use a combination of P2MP mLDP and ingress replication (IR) tunnels in the service. The root nodes send BUM traffic using P2MP tunnels while the leaf nodes use IR tunnels to send BUM traffic to the root nodes. This avoids the need to set up a P2MP tree from each leaf, while it still allows leaf nodes to send BUM traffic to the root nodes.

Figure 110 shows a multicast mLDP tree with root node PE-1 and leaf nodes PE-5, PE-6, and PE-7.

Figure 110 **P2MP mLDP Tree with Root Node PE-1 and Leaf Nodes PE-5, PE-6, and PE-7**



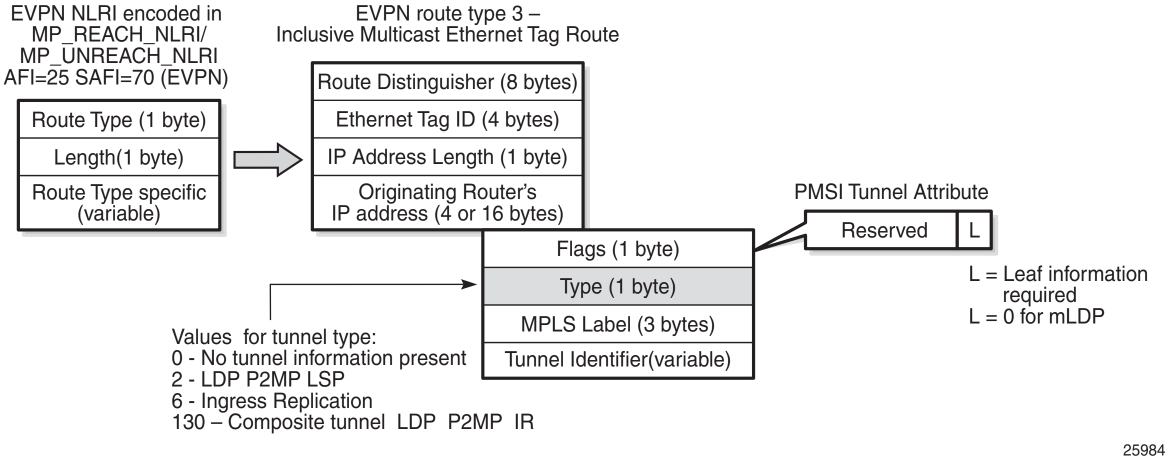
The Inclusive Multicast Ethernet Tag (IMET) route (EVPN route type 3) sent by root node PE-1 contains the required information to set up an mLDP tree, such as the root node IP address and an opaque value. As described in chapter *Multicast Label Distribution Protocol*, the mLDP tree is set up from the leaf nodes toward the root.

The LDP label mapping message contains the root node address, an opaque value, and an MPLS label. The leaf nodes send an LDP label mapping message to their upstream next hop toward the root node of the tree. Each transit node that has received such LDP label mapping message will generate a new LDP label mapping message to its upstream next hop toward the root. This is repeated until the root node receives an LDP label mapping message and the multicast tree is completed.

Figure 110 shows a P2MP mLDP tree rooted in PE-1, which is optimal for multicast traffic. However, no P2MP mLDP tree needs to be rooted in PE-5, PE-6, and PE-7 for the reverse direction. These three PEs can use IR to send traffic to the root (and to the other leaf nodes if needed).

EVPN route type 3 is used for setting up the flooding tree for a specified VPLS service. EVPN route type 3 includes the Provider Multicast Service Interface (PMSI) Tunnel Attribute (PMSI Tunnel Attribute = PTA), which can have different formats depending on the tunnel type; see Figure 111.

Figure 111 **BGP-EVPN Route Type 3 with PTA**



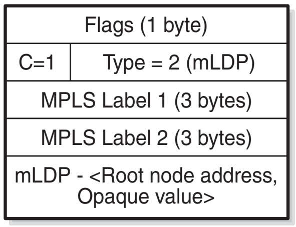
The following route values are used for EVPN-MPLS services:

- The route distinguisher (RD) is taken from the RD of the VPLS service, which can be configured in the BGP context or auto-derived from the BGP-EVPN EVI value. In this case, the RD is auto-derived from the EVI, resulting in a value of 192.0.2.1:1 for VPLS 1 on PE-1.
- The Ethernet tag ID equals 0.
- The IP address length equals 32.
- The originating router's IP address carries the IPv4 system address.

- The PTA can have different formats depending on the tunnel type enabled in the service. The SR OS EVPN-MPLS implementation supports the following tunnel types (SR OS supports different tunnel types for EVPN-VXLAN):
  - Tunnel type 2 - P2MP mLDP
    - The route is referred to as an Inclusive Multicast Ethernet Tag Point-to-Multipoint (IMET-P2MP).
    - Flags: leaf not required.
    - The MPLS label is zero.
    - The tunnel identifier includes the root node address and an opaque number. This is the tunnel identifier that the leaf nodes use to join to the mLDP P2MP tree.
  - Tunnel type 6 - Ingress Replication (IR)
    - The route is referred to as an Inclusive Multicast Ethernet Tag Ingress Replication (IMET-IR).
    - Flags: leaf not required.
    - The MPLS label is a non-zero, downstream allocated label. This MPLS label is allocated to the service and will be the same for unicast MAC/IP routes for the same service, unless **ingress-replication-bum-label** is configured in the service.
    - The tunnel identifier is the tunnel endpoint and is equal to the originating IP address.
  - Tunnel type 130 - Composite tunnel: Type: C-bit (composite) + type 2 (mLDP)
    - The route is referred to as an IMET-P2MP-IR.
    - Flags: leaf not required.
    - MPLS label 1 equals zero.
    - MPLS label 2 is a non-zero, downstream allocated label (as any other IR label). The leaf nodes use the label to set up an EVPN-MPLS binding to the root and add it to the default multicast list.
    - The mLDP tunnel identifier is the root node address and an opaque number. This is the tunnel identifier that the leaf nodes will use to join the mLDP P2MP tree.

Figure 3 shows the PTA for tunnel type 130.

Figure 112 **PTA for Composite Tunnel IMET-P2MP-IR**



25985

The composite bit C is set, indicating that the PTA identifies two tunnels: the transmit tunnel is a P2MP mLDP tunnel and the receive tunnel is an IR tunnel.

IMET-P2MP-IR Routes

The composite tunnel type is an optimized solution that combines mLDP and IR within the same EVPN service so that each root node sends BUM traffic using the P2MP tunnel whereas each leaf-only node sends BUM traffic to the root node using IR.

- PEs configured as **root-and-leaf** can send all BUM traffic over P2MP mLDP tunnels while they receive BUM traffic either over P2MP mLDP tunnels (from other root-and-leaf nodes) or over ingress-replication tunnels (from leaf-only nodes).
- PEs configured as **no root-and-leaf** (default setting) can use IR to send BUM traffic to root nodes and other leaf-only nodes, while receiving BUM traffic over either P2MP mLDP tunnels (from root nodes) or ingress-replication tunnels (from leaf-only nodes).

The root PEs will signal an IMET-P2MP-IR route, indicating that they intend to transmit BUM traffic using an mLDP P2MP tunnel, while they can receive traffic over an IR EVPN-MPLS binding. Composite tunnels reduce the number of P2MP mLDP tunnels that the PE/P routers in the EVI need to handle, because no full mesh of P2MP tunnels among all the PEs in the EVI is required. This is important (in terms of scaling) in services where there are just a pair of root nodes sending BUM in P2MP tunnels and hundreds of leaf nodes that only need to send BUM traffic to the root nodes using IR tunnels.

Configuration

Initial Configuration

The PE and P nodes have the following initial configuration:

- The ports between the routers are configured as network ports and have router interfaces configured.
- IS-IS is enabled on all the router interfaces.
- LDP is enabled on all the router interfaces.
- BGP is enabled on all PEs with route reflector (RR) P-2. The BGP configuration on RR P-2 is as follows:

```
configure
    router
        autonomous-system 64500
        bgp
            vpn-apply-import
            vpn-apply-export
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            split-horizon
            rapid-update evpn
            group "internal"
                family evpn
                cluster 1.1.1.1
                peer-as 64500
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.5
                exit
                neighbor 192.0.2.6
                exit
                neighbor 192.0.2.7
                exit
            exit
        exit
```

Configure EVPN P2MP mLDP in VPLS Service

On the root node PE-1, VPLS 1 is configured as follows:

```
configure
    service
        vpls 1 customer 1 create
            bgp
```

```
exit
bgp-evpn
    ingress-repl-inc-mcast-advertisement       # default setting
    evi 1
    mpls
        auto-bind-tunnel
            resolution any
        exit
        no shutdown
    exit
exit
provider-tunnel
    inclusive
        owner bgp-evpn-mpls
        root-and-leaf
        mldp
        no shutdown
    exit
exit
sap 1/1/3 create
exit
no shutdown
```

The configuration options in the BGP-EVPN context of the VPLS are as follows:

```
*A:PE-1# configure service vpls 1 bgp-evpn
  - bgp-evpn
  - no bgp-evpn

 [no] accept-ivpls-e* - Configure to accept non-zero ethernet-tag MAC routes and
                        process for CMAC flushing
 [no] cfm-mac-advert* - Enable/disable the advertisement of MEP, MIP, and VMEP MAC
                        addresses over the BGP EVPN
 [no] evi             - EVPN Identifier
 [no] incl-mcast-ori* - Configure original IP address
 [no] ingress-repl-i* - Configure BGP EVPN IMET-IR route advertisement
 [no] ip-route-adver* - Configure BGP EVPN IP Route Advertisement
      isid-route-tar* + configure ISID route target information
 [no] mac-advertisem* - Configure BGP EVPN MAC Advertisement
      mac-duplication + Configure BGP EVPN MAC Duplication
      mpls            + Configure BGP EVPN mpls
 [no] unknown-mac-ro* - Configure BGP EVPN Unknown MAC Route
      vxlan           + Configure BGP EVPN vxlan
```

By default, the advertisement of the inclusive multicast route with IR is enabled (**ingress-repl-inc-mcast-advertisement**). However, if it is disabled, the router will not send the IMET-IR or IMET-P2MP-IR routes, regardless of the service being enabled for BGP EVPN-MPLS or BGP EVPN-VXLAN.

For information about the other parameters in the BGP-EVPN context of the VPLS, see chapters EVPN for VXLAN Tunnels (Layer 2) and EVPN for MPLS Tunnels.

The configuration options in the **provider-tunnel inclusive** context are as follows:

```
*A:PE-1# configure service vpls 1 provider-tunnel inclusive
```

```
  - inclusive

 [no] data-delay-int* - Configure data delay interval
 [no] mldp            - Enable/Disable MLDP
 [no] owner           - Configure provider-tunnel owner
 [no] root-and-leaf   - Configure LSP node type
 [no] rsvp            + Configure RSVP parameters
 [no] shutdown        - Administratively enable/disable the service
```

- The **data-delay-interval** is configured in seconds in the range from 3 to 180 seconds. A node configured as root-and-leaf will send all BUM packets (data plane and control plane: ARP, CCMs, and so on) to its provider tunnel after the delay-data-interval has expired. This timer keeps the provider tunnel operationally down until its expiration, and, during that time, the router can use the EVPN-MPLS destinations typically used for IR.
- mLDP is enabled by adding the keyword **mldp** and enabling the provider tunnel (no shutdown).
- The owner must be **bgp-evpn-mpls** if MPLS is enabled in the EVPN.

```
*A:PE-1# configure service vpls 1 provider-tunnel inclusive owner
  - no owner
  - owner {bgp-ad|bgp-vpls|bgp-evpn-mpls}
```

Only one of the three possible owner protocols will support the provider tunnel in the service and needs to be set before the provider tunnel can be enabled. By default, no owner is configured. The following error is raised when a user wants to enable the provider tunnel without an owner:

```
*A:PE-1>config>service>vpls>provider-tunnel>inclusive# no shutdown
INFO: SVCMGR #6732 No owner configured for provider-tunnel
```

After the provider tunnel has an owner and is enabled, the owner can only be changed when the provider tunnel is disabled.

```
*A:PE-1>config>service>vpls>provider-tunnel>inclusive# owner bgp-vpls
INFO: SVCMGR #6721 Provider tunnel is not shutdown
```

After the owner is set, the corresponding protocol is checked to see if it is enabled in the service configuration.

```
*A:PE-1>config>service>vpls>provider-tunnel>inclusive# shutdown
*A:PE-1>config>service>vpls>provider-tunnel>inclusive# owner bgp-vpls
*A:PE-1>config>service>vpls>provider-tunnel>inclusive# no shutdown
MINOR: SVCMGR #6730 provider-tunnel not allowed - bgp-vpls not enabled
```

- If **ingress-repl-inc-mcast-advertisement** is enabled and the PE is root-and-leaf, the router will send an IMET-P2MP-IR route; if no root-and-leaf (default) is configured, the router will send an IMET-IR route. However, if **ingress-repl-inc-mcast-advertisement** is disabled and the PE is root-and-leaf, the router will only send IMET-P2MP routes. Leaf-only nodes will not send any IMET routes at all in case no IR multicast advertisement is allowed.

Root-and-leaf nodes will only send BUM traffic to the P2MP tunnel as long as it is active. If the P2MP tunnel goes operationally down, it will start sending BUM traffic to IR tunnels (EVPN-MPLS destinations shown in the **show service id 1 evpn-mpls** command).

- If a provider tunnel is configured on a node, the router can join P2MP trees as a leaf, by generating an LDP label mapping message including the corresponding P2MP mLDP FEC. If no provider tunnel is configured, the node will not join P2MP mLDP trees, and can only use IR for BUM.
- If one node is configured as root, all other nodes must be configured with provider tunnels; otherwise, they will not receive BUM traffic sent on P2MP tunnels. The configuration of leaf-only node PE-5 is as follows, the main difference with the configuration for the root being the **no root-and-leaf** (default setting):

```
configure
    service
        vpls 1 customer 1 create
            bgp
            exit
            bgp-evpn
                evi 1
                mpls
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            provider-tunnel
                inclusive
                    owner bgp-evpn-mpls
                    mldp
                    no shutdown
                exit
            exit
            sap 1/2/1:1 create
            exit
            no shutdown
```

As described, the tunnel types for BUM traffic are controlled by **ingress-repl-inc-mcast-advertisement** and the provider-tunnel context (**root-and-leaf**). The IMET route sending behavior is summarized in Table 8.

Table 8 **IMET routes and Tunnel Types advertised based on the configuration**

IMET route set	Root + Leaf PE	Leaf-only	No provider-tunnel
IR-mcast advertisement	Composite P2MP + IR	IR	IR
No IR-mcast advertisement	P2MP	-	-

Information about the provider tunnel can be retrieved as follows:

```
*A:PE-1# show service id 1 provider-tunnel

===============================================================================
Service Provider Tunnel Information
===============================================================================
Type               : inclusive          Root and Leaf      : enabled
Admin State        : enabled            Data Delay Intvl   : 15 secs
PMSI Type          : ldp                LSP Template       :
Remain Delay Intvl : 0 secs             LSP Name used      : 8193
PMSI Owner         : bgpEvpnMpls
Oper State         : up                 Root Bind Id       : 17407
===============================================================================
*A:PE-1#
```

Note: The same IMET-P2MP route cannot be imported by two services at the same time. If two VPLS services (where a provider tunnel is enabled) have the same import route-target, only one service will join the mLDP tree (whichever comes first).

EVPN P2MP mLDP Operation

After the root and leaf nodes are configured as shown, the root node sends BGP EVPN composite IMET-P2MP-IR routes, as follows:

```
1 2017/05/08 12:26:10.08 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 100
    Flag: 0x90 Type: 14 Len: 28 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.1
        Type: EVPN-Incl-mcast Len: 17 RD: 192.0.2.1:1, tag: 0,
                              orig_addr len: 32, orig_addr: 192.0.2.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:64500:1
        bgp-tunnel-encap:MPLS
    Flag: 0xc0 Type: 22 Len: 25 PMSI:
        Tunnel-type Composite LDP P2MP IR (130)
        Flags: (0x0)[Type: None BM: 0 U: 0 Leaf: not required]
        MPLS Label1 Ag  0
        MPLS Label2 IR 4194176
        Root-Node 192.0.2.1, LSP-ID 0x2001
"
```

The PTA for tunnel type 130 (composite tunnel) has two MPLS labels, of which MPLS label 1 equals zero. MPLS label 2 is used by the downstream nodes to set up the EVPN-MPLS destination to the root node and add it to the default multicast list. The actual MPLS label only uses the high-order 20 bits out of the 24 bits advertised in the MPLS label. Therefore, the value 4194176 needs to be divided by 16 to have the MPLS label: 4194176/16 = 262136. This is due to the debug message being shown before the router can parse the label field and see whether it corresponds to an MPLS label (20 bits) or a VXLAN VNI (24 bits).

The tunnel identifier field contains the root node address 192.0.2.1 and the opaque value 0x2001, which corresponds to decimal value 8193. With this tunnel identifier, the leaf nodes can join the mLDP multicast tree toward the root node by sending LDP label mapping messages that contain the root node IP address and the opaque value.

Note: When static P2MP mLDP tunnels and dynamic P2MP mLDP tunnels used by BGP-EVPN coexist on the same router, Nokia recommends that the static tunnels to use a tunnel ID less than 8193. If a tunnel ID is statically configured with a value equal to or greater than 8193, BGP-EVPN may attempt to use the same tunnel ID for services with an enabled provider tunnel and fail to set up an mLDP tunnel.

The root node PE-1 receives IMET-IR routes from all leaf nodes, as shown for the BGP update sent by leaf node PE-5 (via RR P-2):

```
2 2017/05/08 12:26:16.08 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 98
    Flag: 0x90 Type: 14 Len: 28 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.5
        Type: EVPN-Incl-mcast Len: 17 RD: 192.0.2.5:1, tag: 0,
                              orig_addr len: 32, orig_addr: 192.0.2.5
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.5
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        1.1.1.1
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:64500:1
        bgp-tunnel-encap:MPLS
    Flag: 0xc0 Type: 22 Len: 9 PMSI:
        Tunnel-type Ingress Replication (6)
        Flags: (0x0)[Type: None BM: 0 U: 0 Leaf: not required]
        MPLS Label 4194176
        Tunnel-Endpoint 192.0.2.5
"
```

The PTA tunnel type 6 for IR has only one MPLS label, which corresponds to the MPLS label 262136 allocated for the service. The tunnel identifier is the tunnel endpoint 192.0.2.5, which is the system address of the originating leaf node.

On leaf node PE-5, three BGP EVPN inclusive multicast routes have been learned and are used, as follows:

```
*A:PE-5# show router bgp routes evpn inclusive-mcast
===============================================================================
 BGP Router ID:192.0.2.5        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN Inclusive-Mcast Routes
===============================================================================
Flag  Route Dist.         OrigAddr
      Tag                 NextHop
-------------------------------------------------------------------------------
u*>i  192.0.2.1:1         192.0.2.1
      0                   192.0.2.1

u*>i  192.0.2.6:1         192.0.2.6
      0                   192.0.2.6

u*>i  192.0.2.7:1         192.0.2.7
      0                   192.0.2.7

-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:PE-5#
```

The details of the BGP EVPN inclusive multicast route sent by root node PE-1 to leaf node PE-5 are as follows:

```
*A:PE-5# show router bgp routes evpn inclusive-mcast rd 192.0.2.1:1 detail
===============================================================================
 BGP Router ID:192.0.2.5        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN Inclusive-Mcast Routes
===============================================================================
Original Attributes

Network        : N/A
Nexthop        : 192.0.2.1
From           : 192.0.2.2
```

```
Res. Nexthop   : 192.168.35.1
Local Pref.    : 100                    Interface Name : int-PE-5-P-3
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : target:64500:1 bgp-tunnel-encap:MPLS
Cluster        : 1.1.1.1
Originator Id  : 192.0.2.1              Peer Router Id : 192.0.2.2
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
EVPN type      : INCL-MCAST
ESI            : N/A
Tag            : 0
Originator IP  : 192.0.2.1
Route Dist.    : 192.0.2.1:1
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 01h48m32s
-------------------------------------------------------------------------------
PMSI Tunnel Attributes :
Tunnel-type    : Composite LDP P2MP IR
Flags          : Type: RNVE(0) BM: 0 U: 0 Leaf: not required
MPLS Label1 Ag : LABEL 0
MPLS Label2 IR : LABEL 262136
Root-Node      : 192.0.2.1              LSP-ID         : 8193
-------------------------------------------------------------------------------
---snip---
```

The MPLS label is 262136, as described. The LSP ID equals 8193, which corresponds to the hexadecimal value 0x2001 in the preceding BGP update message sent by the root node PE-1.

To set up the mLDP tree, leaf node PE-5 has generated an LDP label mapping message to the next hop router toward the root, P-3. The label mapping message includes the root address 192.0.2.1, the opaque value 8193, and MPLS label 262132, as follows:

```
*A:PE-5# show router ldp bindings active p2mp ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.5)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
===============================================================================
LDP Generic IPv4 P2MP Bindings (Active)
```

```
===============================================================================
P2MP-Id                                     Interface
RootAddr                                    Op              IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
-------------------------------------------------------------------------------
8193                                        73740
192.0.2.1                                   Pop             262132      --
  --                                          --

-------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 1
---snip---
```

P-3 has received two label mapping messages: one from PE-5 and one from PE-6. P-3 has sent one label mapping message to its upstream next hop P-2 with label 262134, as follows:

```
*A:P-3# show router ldp bindings active p2mp ipv4 opaque-type generic

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.3)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FE
===============================================================================
LDP Generic IPv4 P2MP Bindings (Active)
===============================================================================
P2MP-Id                                     Interface
RootAddr                                    Op              IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
-------------------------------------------------------------------------------
8193                                        Unknw
192.0.2.1                                   Swap            262134    262132
192.168.35.2                                1/1/3

8193                                        Unknw
192.0.2.1                                   Swap            262134    262128
192.168.36.2                                1/1/4

-------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 2
===============================================================================
*A:P-3#
```

P-2 has received two label mapping messages: one from P-3 and one from P-4. P-2 has sent a label mapping message toward the root node PE-1 with label 262136, as follows:

```
*A:P-2# show router ldp bindings active p2mp ipv4 opaque-type generic

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FE
===============================================================================
LDP Generic IPv4 P2MP Bindings (Active)
===============================================================================
P2MP-Id                                     Interface
RootAddr                                    Op              IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
-------------------------------------------------------------------------------
8193                                        Unknw
192.0.2.1                                   Swap            262136    262134
192.168.23.2                                1/1/2

8193                                        Unknw
192.0.2.1                                   Swap            262136    262135
192.168.24.2                                1/1/1

-------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 2
===============================================================================
*A:P-2#
```

When the LDP label reaches the root node PE-1, the mLDP tree is complete and it can be used for BUM traffic.

The following tools command shows the provider tunnels for VPLS 1 on root and leaf nodes. On root node PE-1, there is one originating inclusive provider tunnel and there are no terminating inclusive provider tunnels, as follows:

```
*A:PE-1# tools dump service id 1 provider-tunnels

===============================================================================
VPLS 1 Inclusive Provider Tunnels Originating
===============================================================================
ipmsi (LDP)                                     P2MP-ID  Root-Addr
-------------------------------------------------------------------------------
8193                                            8193    192.0.2.1

-------------------------------------------------------------------------------

===============================================================================
VPLS 1 Inclusive Provider Tunnels Terminating
===============================================================================
ipmsi (LDP)                                     P2MP-ID  Root-Addr
-------------------------------------------------------------------------------

No Tunnels Found
-------------------------------------------------------------------------------
*A:PE-1#
```

On leaf node PE-5, no originating inclusive provider tunnels are established; only one terminating provider tunnel, as follows:

```
*A:PE-5# tools dump service id 1 provider-tunnels

===============================================================================
VPLS 1 Inclusive Provider Tunnels Originating
===============================================================================
ipmsi (LDP)                                     P2MP-ID  Root-Addr
-------------------------------------------------------------------------------

No Tunnels Found
-------------------------------------------------------------------------------

===============================================================================
VPLS 1 Inclusive Provider Tunnels Terminating
===============================================================================
ipmsi (LDP)                                     P2MP-ID  Root-Addr
-------------------------------------------------------------------------------
                                                8193    192.0.2.1

-------------------------------------------------------------------------------
*A:PE-5#
```

The inclusive provider tunnels are identified by the combination of the P2MP ID (opaque value) and the root address. These parameters are in every label mapping message and they are included in the PTA tunnel identifier for tunnel type 130 (IMET-P2MP-IR) and for tunnel type 2 (IMET-P2MP).

In VPLS 1 on root node PE-1, an SDP of type VplsPmsi is auto-created, as follows:

```
*A:PE-1# show service id 1 sdp

===============================================================================
Services: Service Destination Points
===============================================================================
SdpId            Type     Far End addr    Adm     Opr       I.Lbl     E.Lbl
-------------------------------------------------------------------------------
17407:4294967290 VplsPmsi not applicable  Up      Up        None      3
-------------------------------------------------------------------------------
Number of SDPs : 1
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

The detailed information about this SDP includes the traffic statistics: ingress/egress and forwarding/dropped, as follows:

```
*A:PE-1# show service id 1 sdp detail

===============================================================================
Services: Service Destination Points Details
===============================================================================
-------------------------------------------------------------------------------
 Sdp Id 17407:4294967290  -(not applicable)
```

```
-------------------------------------------------------------------------------
Description     : (Not Specified)
SDP Id             : 17407:4294967290         Type              : VplsPmsi
Split Horiz Grp    : (Not Specified)
Etree Root Leaf Tag: Disabled                 Etree Leaf AC     : Disabled
VC Type            : Ether                    VC Tag            : n/a
Admin Path MTU     : 9194                     Oper Path MTU     : 9194
Delivery           : MPLS
Far End            : not applicable
---snip---
PMSI Owner         : bgpEvpnMpls

Admin State        : Up                       Oper State        : Up
---snip---
Statistics            :
I. Fwd. Pkts.      : 0                        I. Dro. Pkts.     : 0
I. Fwd. Octs.      : 0                        I. Dro. Octs.     : 0
E. Fwd. Pkts.      : 15062951                 E. Fwd. Octets    : 903777580
---snip---
```

IGMP Snooping

When IGMP snooping is disabled and a multicast stream enters VPLS 1 on the root node, this stream is sent to all the leaf nodes, even if no receivers join the multicast group on the leaf nodes. In this example, a receiver connected to PE-5 joins a multicast group, but there are no receivers for any multicast group on PE-6 and PE-7. By default, IGMP is disabled and the multicast stream is flooded to all leaf PEs, as can be verified with the following monitor command on PE-6 where no receivers have joined any multicast stream:

```
*A:PE-6# monitor port 1/2/1 repeat 3

===============================================================================
Monitor statistics for Port 1/2/1
===============================================================================
                                                   Input                 Output
-------------------------------------------------------------------------------
---snip---
-------------------------------------------------------------------------------
At time t = 10 sec (Mode: Delta)
-------------------------------------------------------------------------------
Octets                                                 0               12722732
Packets                                                0                 187099
Errors                                                 0                      0

-------------------------------------------------------------------------------
At time t = 20 sec (Mode: Delta)
-------------------------------------------------------------------------------
Octets                                                 0               12692472
Packets                                                0                 186654
Errors                                                 0                      0

-------------------------------------------------------------------------------
```

```
---snip---
```

This implies that bandwidth is wasted, which can be prevented by enabling IGMP snooping. IGMP snooping ensures that multicast traffic will only be sent to the receivers that joined a multicast group. IGMP snooping can be enabled in VPLS 1 on all PEs, as follows:

```
configure service vpls 1 igmp-snooping no shutdown
```

A receiver connected to PE-5 has sent an IGMP report whereas PE-6 has no receivers that joined a multicast group. The traffic counters are monitored on the outgoing port to the (potential) receivers. On PE-5, traffic is sent to the receiver, as follows:

```
*A:PE-5# monitor port 1/2/1 repeat 3

===============================================================================
Monitor statistics for Port 1/2/1
===============================================================================
                                                   Input                 Output
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
---snip---
-------------------------------------------------------------------------------
At time t = 10 sec (Mode: Delta)
-------------------------------------------------------------------------------
Octets                                                 0               13053892
Packets                                                0                 191969
Errors                                                 0                      0
-------------------------------------------------------------------------------
---snip---
```

On PE-6, no traffic is sent to any receiver, as follows:

```
*A:PE-6# monitor port 1/2/1 repeat 3

===============================================================================
Monitor statistics for Port 1/2/1
===============================================================================
                                                   Input                 Output
-------------------------------------------------------------------------------
---snip---
-------------------------------------------------------------------------------
At time t = 10 sec (Mode: Delta)
-------------------------------------------------------------------------------
Octets                                                 0                      0
Packets                                                0                      0
Errors                                                 0                      0
-------------------------------------------------------------------------------
---snip---
```

IGMP snooping can be enabled in EVPN-MPLS services with IR or provider-tunnel mLDP trees. When IGMP snooping is enabled on the VPLS, all the EVPN-MPLS destinations are added to the MFIB as a single router interface. IGMP queries and reports are properly forwarded to and from EVPN-MPLS destinations.

The following shows the EVPN-MPLS destinations as part of the MFIB when IGMP-snooping is enabled:

```
*A:PE-5# show service id 1 mfib

===============================================================================
Multicast FIB, Service 1
===============================================================================
Source Address  Group Address         Port Id                      Svc Id   Fwd
                                                                            Blk
-------------------------------------------------------------------------------
*               *                     sap:1/2/1:1                  Local    Fwd
                                      eMpls:192.0.2.1:262136       Local    Fwd
                                      eMpls:192.0.2.6:262136       Local    Fwd
                                      eMpls:192.0.2.7:262136       Local    Fwd
-------------------------------------------------------------------------------
Number of entries: 1
===============================================================================
*A:PE-5#
```

Connected to SAP 1/2/1:1, PE-5 has a receiver that joined the multicast stream. EVPN-MPLS is added as a single logical IGMP snooping interface and treated as an mrouter, also on the other leaf nodes, as follows:

```
*A:PE-5# show service id 1 igmp-snooping base

===============================================================================
IGMP Snooping Base info for service 1
===============================================================================
Admin State : Up
Querier     : 172.16.0.5 on SAP 1/2/1:1
-------------------------------------------------------------------------------
Port                      Oper MRtr Pim  Send Max   Max  Max   MVR       Num
Id                        Stat Port Port Qrys Grps  Srcs Grp   From-VPLS Grps
                                                         Srcs
-------------------------------------------------------------------------------
sap:1/2/1:1               Up   Yes  No   No   None  None None  Local     0
evpn-mpls                 Up   Yes  N/A  N/A  N/A   N/A  N/A   N/A       N/A
===============================================================================
*A:PE-5#
```

On leaf node PE-5, the receiving host connected to SAP 1/2/1:1 has IP address 172.16.0.5, as follows:

```
*A:PE-5# show service id 1 igmp-snooping mrouters

===============================================================================
IGMP Snooping Multicast Routers for service 1
===============================================================================
MRouter          Port Id                    Up Time        Expires   Version
-------------------------------------------------------------------------------
172.16.0.5       1/2/1:1                    0d 04:58:11    238s      3
-------------------------------------------------------------------------------
Number of mrouters: 1
===============================================================================
*A:PE-5#
```

On leaf node PE-6, SAP 1/2/1:1 has no receiving host connected, but EVPN-MPLS is always added as an mrouter, as follows:

```
*A:PE-6# show service id 1 igmp-snooping base

===============================================================================
IGMP Snooping Base info for service 1
===============================================================================
Admin State : Up
Querier     : 172.16.0.5 on evpn-mpls
-------------------------------------------------------------------------------
Port                      Oper MRtr Pim  Send Max   Max  Max   MVR       Num
Id                        Stat Port Port Qrys Grps  Srcs Grp   From-VPLS Grps
                                                         Srcs
-------------------------------------------------------------------------------
sap:1/2/1:1               Up   No   No   No   None  None None  Local      0
evpn-mpls                 Up   Yes  N/A  N/A  N/A   N/A  N/A   N/A       N/A
===============================================================================
*A:PE-6#
```

On PE-6, the only mrouter in the list is the receiving host connected to PE-5, with port ID EVPN-MPLS instead of a local SAP, as follows:

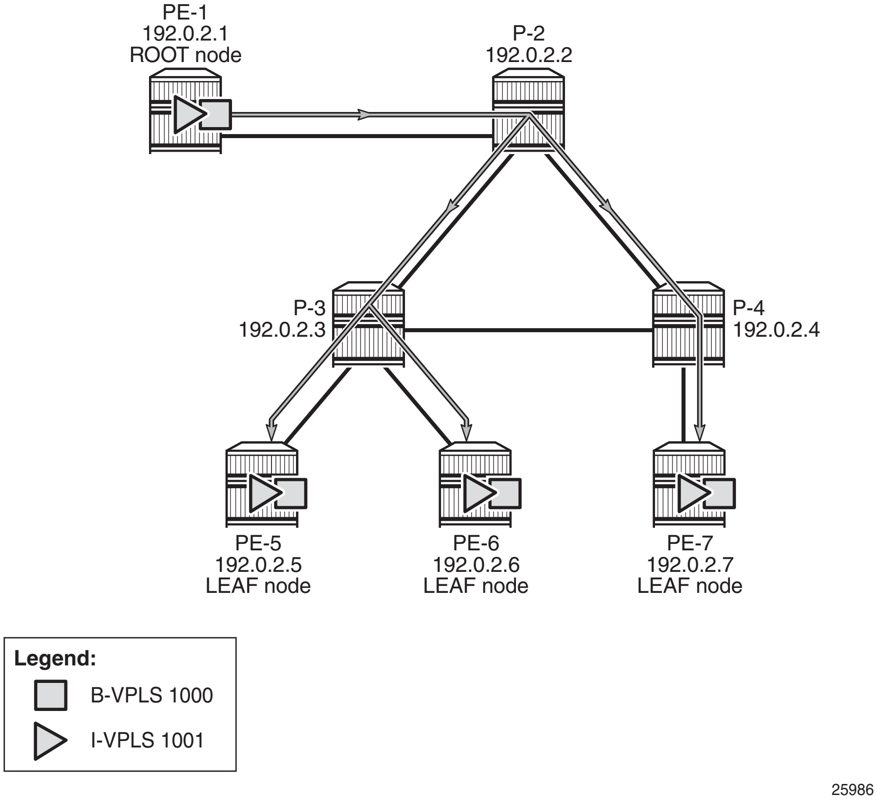
```
*A:PE-6# show service id 1 igmp-snooping mrouters

===============================================================================
IGMP Snooping Multicast Routers for service 1
===============================================================================
MRouter          Port Id                    Up Time        Expires   Version
-------------------------------------------------------------------------------
172.16.0.5       evpn-mpls                  0d 04:45:40    241s      3
-------------------------------------------------------------------------------
Number of mrouters: 1
===============================================================================
*A:PE-6#
```

PBB-EVPN and P2MP mLDP

PBB-EVPN is described in chapter EVPN for PBB over MPLS (PBB-EVPN).

Figure 113 shows the setup for P2MP mLDP in PBB-EVPN.

Figure 113 **P2MP mLPD in PBB-EVPN**



P2MP mLDP tunnels can also be used in PBB-EVPN services. In release 14.0, the use of **provider-tunnel inclusive mldp** is only for the default multicast list; no per-ISID IMET-P2MP routes are supported.

The B-VPLS still uses Multicast Forwarding Information Bases (MFIBs) for ISIDs using IR.

If an ISID policy is configured in the B-VPLS, a range of ISIDs configured with **use-def-mcast** will use the P2MP tree, and a range of ISIDs configured with **advertise-local** will make the router advertise IMET-IR routes for the local ISIDs in the range.

PE-1 is configured as root-and-leaf. The configuration for B-VPLS and I-VPLS is as follows:

```
configure
    service
        vpls 1000 customer 1 b-vpls create
            service-mtu 2000
            pbb
                source-bmac 00:00:00:00:00:01
            exit
            bgp
            exit
            bgp-evpn
                evi 1000
                mpls
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            provider-tunnel
                inclusive
                    owner bgp-evpn-mpls
                    root-and-leaf
                    mldp
                    no shutdown
                exit
            exit
            isid-policy
                entry 10 create
                    use-def-mcast
                    no advertise-local
                    range 1001 to 2000
                exit
            exit
            no shutdown
        exit
        vpls 1001 customer 1 i-vpls create
            pbb
                backbone-vpls 1000
                exit
            exit
            sap 1/1/3 create
            exit
            no shutdown
        exit
```

In this example, ISIDs in the range from 1001 to 2000 will use the P2MP tree (**use-def-mcast**) and the router does not advertise the IMET-IR routes for the local ISIDs included in that range (**no advertise-local**). Any other local ISID will advertise an IMET-IR and will use the MFIB to forward BUM packets to the remote EVPN-MPLS bindings created by IMET-IR routes.

The configuration on the leaf nodes PE-5, PE-6, and PE-7 is similar to the one for the root node, except for the **no root-and-leaf** setting (which is default), as follows:

```
configure
    service
        vpls 1000 customer 1 b-vpls create
            service-mtu 2000
            pbb
                source-bmac 00:00:00:00:00:05
            bgp
            exit
            bgp-evpn
                evi 1000
                mpls
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            provider-tunnel
                inclusive
                    owner bgp-evpn-mpls
                    mldp
                    no shutdown
                exit
            exit
            isid-policy
                entry 10 create
                    use-def-mcast
                    no advertise-local
                    range 1001 to 2000
                exit
            exit
            no shutdown
        exit
        vpls 1001 customer 1 i-vpls create
            pbb
                backbone-vpls 1000
                exit
            exit
            sap 1/2/1:1001 create
            exit
            no shutdown
        exit
```

A VPLS-PMSI SDP is auto-created in the B-VPLS at the root node, as follows:

```
*A:PE-1# show service id 1000 sdp

===============================================================================
Services: Service Destination Points
===============================================================================
SdpId            Type     Far End addr    Adm     Opr       I.Lbl     E.Lbl
-------------------------------------------------------------------------------
17407:4294967289 VplsPmsi not applicable  Up      Up        None      3
-------------------------------------------------------------------------------
Number of SDPs : 1
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

The default multicast list for the B-VPLS 1000 can be retrieved on root and leaf nodes, for instance for leaf node PE-5, as follows:

```
*A:PE-5# tools dump service id 1000 evpn-mpls default-multicast-list
-------------------------------------------------------------------------
TEP Address                              Egr Label
```

```
                                         Transport
-------------------------------------------------------------------------
192.0.2.1                                262140
                                         ldp
192.0.2.6                                262142
                                         ldp
192.0.2.7                                262139
                                         ldp
-------------------------------------------------------------------------
*A:PE-5#
```

IGMP snooping can be enabled in the I-VPLS 1001 on all PEs, as follows:

```
configure service vpls 1001 igmp-snooping no shutdown
```

After IGMP snooping is enabled, the multicast stream is not flooded anymore to any receivers until they send an IGMP report for the multicast stream.

On each PE, the logical interface B-EVPN-MPLS is added as a single IGMP snooping interface and treated as an mrouter, as follows:

```
*A:PE-5# show service id 1001 igmp-snooping base

===============================================================================
IGMP Snooping Base info for service 1001
===============================================================================
Admin State : Up
Querier     : 172.16.0.55 on SAP 1/2/1:1001
-------------------------------------------------------------------------------
Port                      Oper MRtr Pim  Send Max   Max  Max  MVR       Num
Id                        Stat Port Port Qrys Grps  Srcs Grp  From-VPLS Grps
                                                         Srcs
-------------------------------------------------------------------------------
b-evpn-mpls               Up   Yes  N/A  N/A  N/A   N/A  N/A  N/A       N/A
sap:1/2/1:1001            Up   Yes  No   No   None  None None Local     0
===============================================================================
*A:PE-5#
```

PE-5 has a receiver that sent an IGMP report for a multicast group in I-VPLS 1001 on SAP 1/2/1:1001 and this SAP is an mrouter port. On PE-6, there is no receiver that sent IGMP reports; therefore, the only mrouter port corresponds to the B-EVPN-MPLS logical interface, as follows:

```
*A:PE-6# show service id 1001 igmp-snooping base

===============================================================================
IGMP Snooping Base info for service 1001
===============================================================================
Admin State : Up
Querier     : 172.16.0.55 on evpn-mpls
-------------------------------------------------------------------------------
Port                      Oper MRtr Pim  Send Max   Max  Max  MVR       Num
Id                        Stat Port Port Qrys Grps  Srcs Grp  From-VPLS Grps
                                                         Srcs
-------------------------------------------------------------------------------
```

```
b-evpn-mpls               Up   Yes  N/A  N/A  N/A   N/A  N/A   N/A       N/A
sap:1/2/1:1001            Up   No   No   No   None  None None  Local     0
===============================================================================
*A:PE-6#
```

PE-5 has a local mrouter 172.16.0.55 on SAP 1/2/1:1001, as follows:

```
*A:PE-5# show service id 1001 igmp-snooping mrouters

===============================================================================
IGMP Snooping Multicast Routers for service 1001
===============================================================================
MRouter          Port Id                    Up Time        Expires   Version
-------------------------------------------------------------------------------
172.16.0.55      1/2/1:1001                 0d 01:15:41    213s      3
-------------------------------------------------------------------------------
Number of mrouters: 1
===============================================================================
*A:PE-5#
```

On PE-6, mrouter 172.16.0.55 is not local; therefore, the EVPN-MPLS logical interface is used, as follows:

```
*A:PE-6# show service id 1001 igmp-snooping mrouters

===============================================================================
IGMP Snooping Multicast Routers for service 1001
===============================================================================
MRouter          Port Id                    Up Time        Expires   Version
-------------------------------------------------------------------------------
172.16.0.55      evpn-mpls                  0d 00:02:22    237s      3
-------------------------------------------------------------------------------
Number of mrouters: 1
===============================================================================
*A:PE-6#
```

Conclusion

Service providers are migrating their existing VPN services to EVPN and expect at least the same capabilities in EVPN, including the forwarding of BUM traffic. Ingress replication is a good mechanism for broadcast and unknown unicast traffic in EVPN networks, but not efficient for multicast applications. EVPN P2MP mLDP offers efficiency for multicast, using composite tunnels combining the benefits of P2MP mLDP and IR.

PBB-Epipe

This chapter provides information about Provider Backbone Bridging (PBB) — Ethernet Virtual Leased Line in an MPLS-based network which is applicable to SR OS.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SROS Release 7.0.R5. The CLI in the current edition corresponds to SR OS Release 15.0.R2. There are no specific prerequisites.

Overview

The draft-ietf-l2vpn-pbb-vpls-pe-model-00*, Extensions to VPLS PE model for Provider Backbone Bridging,* describes the PBB-VPLS model supported by SR OS. This model expands the VPLS PE model to support PBB as defined by the IEEE 802.1ah.

The PBB model is organized around a B-component (backbone instance) and an I-component (customer instance). In Nokia's implementation of the PBB model, the use of an Epipe as I-component is allowed for point-to-point services. Multiple I-VPLS and Epipe services can be all mapped to the same B-VPLS (backbone VPLS instance).

The use of Epipe scales the E-Line services because no MAC switching, learning, or replication is required in order to deliver the point-to-point service. All packets ingressing the customer SAP are PBB-encapsulated and unicasted through the B-VPLS tunnel using the backbone destination MAC of the remote PBB PE. All the packets egressing the B-VPLS destined for the Epipe are PBB de-encapsulated and forwarded to the customer SAP.

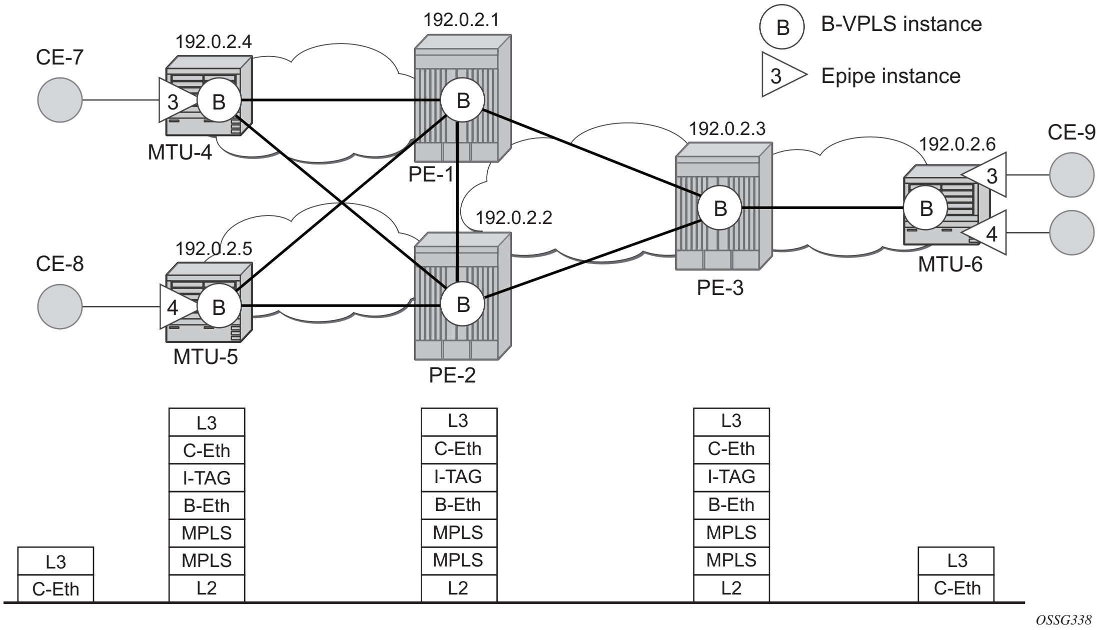
Some use cases for PBB-Epipe are:

- Get a more efficient and scalable solution for point-to-point services:
  - Up to 8K VPLS services per box are supported (including I-VPLS or B-VPLS) and using I-VPLS for point-to-point services takes VPLS resources as well as unnecessary customer MAC learning. A better solution is to connect a PBB-Epipe to a B-VPLS instance, where there is no customer MAC switching/learning.
- Take advantage of the pseudowire aggregation in the M:1 model:
  - Many Epipe services may use only a single service and set of pseudowires over the backbone.
- Have a uniform provisioning model for both point-to-point (Epipe) and multipoint (VPLS) services.
  - Using the PBB-Epipe, the core MPLS/pseudowire infrastructure does not need to be modified: the new Epipe inherits the existing pseudowire and MPLS structure already configured on the B-VPLS and there is no need for configuring new tunnels or pseudowire switching instances at the core.

Knowledge of the PBB-VPLS architecture and functionality on the service router family is assumed throughout this section. For additional information, see the relevant Nokia user documentation.

The following network setup will be used throughout the rest of the chapter.

Figure 114 **Network Topology**



The setup consists of a three 7x50 SR/ESS (PE-1, PE-2 and PE-3) core and three Multi-Tenant Unit (MTU) nodes connected to the core. A backbone VPLS instance (B-VPLS 101) will be defined in all the six nodes, whereas two Epipe services will be defined as illustrated in Figure 114 (Epipe 3 in nodes MTU-4 and MTU-6, Epipe 4 in nodes MTU-5 and MTU-6). Those Epipe services will be multiplexed into the common B-VPLS 101, using the I-Service ID (ISID) field within the I-TAG as the demultiplexer field required at the egress MTU to differentiate each specific customer. I-VPLS and Epipe services can be mapped to the same B-VPLS.

The B-VPLS domain constitutes a H-VPLS network itself, with spoke SDPs from the MTUs to the core PE layer. Active/standby (A/S) spoke SDPs can be used from the MTUs to the PEs (like in the MTU-4 and MTU-5 cases) or single non-redundant spoke SDPs (like MTU-6).

The protocol stack being used along the path between the CEs is represented in Figure 114.

Configuration

This section describes all the relevant PBB-Epipe configuration tasks for the setup shown in Figure 114. The appropriate B-VPLS and associated IP/MPLS configuration is out of the scope of this document. In this particular example, the following protocols will be configured beforehand in the core:

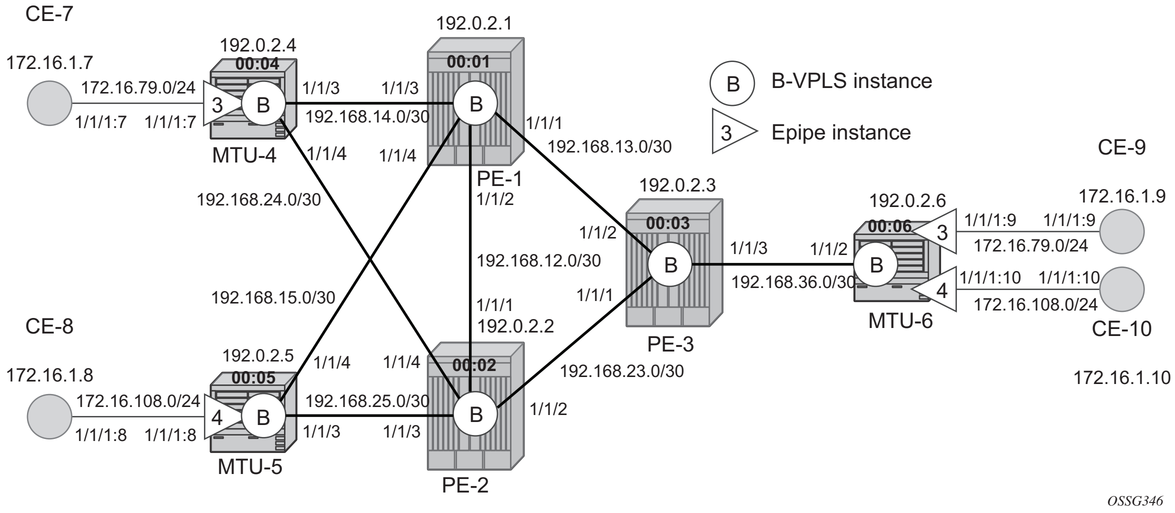
- ISIS-TE as IGP with all the interfaces being level-2. Alternatively, OSPF could have been used.
- RSVP-TE as the MPLS protocol to signal the transport tunnels.
- LSPs between core PEs will be fast re-route protected (facility bypass tunnels) whereas LSP tunnels between MTUs and PEs will not be protected.
- The protection between MTU-4, MTU-5 and PE-1, PE-2 will be based on the A/S pseudowire protection configured in the B-VPLS.
- BGP is configured for auto-discovery—BGP-AD (Layer 2 VPN family), because FEC 129 will be used to establish the pseudowires between PEs in the core (FEC 128 between MTU and PE nodes).

Once the IP/MPLS infrastructure is up and running, the service configuration tasks described in the following sections can be implemented.

PBB Epipe Service Configuration

In this particular example, the Epipes 3 and 4 are using the B-VPLS 101 in the core. The same B-VPLS which is multiplexing the Epipe services into a common service provider infrastructure can also be used to connect the I-VPLS instances existing in the network for multipoint services.

Figure 115 **Setup Detailed View**



B-VPLS and PBB Configuration

First, configure the B-VPLS instance that will carry the PBB traffic. There is no specific requirement on the B-VPLS to support Epipes. The following shows the B-VPLS configuration on MTU-4 and PE-1.

On MTU-4:

```
configure
    service
        vpls 101 customer 1 b-vpls create
            service-mtu 2000
            pbb
                source-bmac 00:04:04:04:04:04
            exit
            endpoint "core" create
                no suppress-standby-signaling
            exit
            spoke-sdp 41:101 endpoint "core" create
                precedence primary
            exit
            spoke-sdp 42:101 endpoint "core" create
            exit
            no shutdown
        exit
```

On PE-1:

```
configure
    service
        pw-template 1 use-provisioned-sdp create
            split-horizon-group "CORE"
            exit
        exit
        vpls 101 customer 1 b-vpls create
            service-mtu 2000
            pbb
                source-bmac 00:01:01:01:01:01
            exit
            bgp
                route-target export target:65000:101 import target:65000:101
                pw-template-binding 1
                exit
            exit
            bgp-ad
                vpls-id 65000:101
                no shutdown
            exit
            spoke-sdp 14:101 create
            exit
            spoke-sdp 15:101 create
            exit
            no shutdown
        exit
```

The relevant B-VPLS commands are in **bold**.

The keyword **b-vpls** is given at creation time and therefore it cannot be added to an existing regular VPLS instance. Besides the **b-vpls** keyword, the B-VPLS is a regular VPLS instance in terms of configuration, with the following exceptions:

- The B-VPLS service MTU must be at least 18 bytes greater than the Epipe MTU of the multiplexed instances. In this example, the I-VPLS instances will have the default service MTU (1514 bytes), therefore, any MTU equal or greater than 1532 bytes must be configured. In this particular example, an MTU of 2000 bytes is configured in the B-VPLS instance throughout the network.
- The source B-MAC is the MAC that will be used as a source when the PBB traffic is originated from that node. It is possible to configure a source B-MAC per B-VPLS instance (if there are more than one B-VPLS) or a common source B-MAC that will be shared by all the B-VPLS instances in the node. A common B-MAC is configured as follows:

```
*A:MTU-4# configure service pbb source-bmac 00:04:04:04:04:04
*A:MTU-5# configure service pbb source-bmac 00:05:05:05:05:05
*A:MTU-6# configure service pbb source-bmac 00:06:06:06:06:06
```

The following considerations will be taken into account when configuring the B-VPLS:

- B-VPLS SAPs:
  - Ethernet DOT1Q and NULL encapsulations are supported.
  - Default SAP types are blocked in the CLI for the B-VPLS SAP.
- B-VPLS SDPs:
  - For MPLS, both mesh and spoke SDPs with split horizon groups are supported.
  - Similar to regular pseudowire, the outgoing PBB frame on an SDP (for example, Bpseudowire) contains a BVID Qtag only if the pseudowire type is Ethernet VLAN (vc-type=vlan). If the pseudowire type is Ethernet (vc-type=ether), the BVID qtag is stripped before the frame goes out.
  - BGP-AD is supported in the B-VPLS, therefore, spoke SDPs in the B-VPLS can be signaled using FEC 128 or FEC 129. In this example, BGP-AD and FEC 129 are used. A split-horizon group has been configured to emulate the behavior of mesh SDPs in the core.
- While Multiple MAC Registration Protocol (MMRP) is useful to optimize the flooding in the B-VPLS domain and build a flooding tree on a per I-VPLS basis, it does not have any effect for Epipes because the destination B-MAC used for Epipes is always the destination B-MAC configured in the Epipe and never the group B-MAC corresponding to the ISID.
- If a local Epipe instance is associated with the B-VPLS, local frames originated or terminated on local Epipe(s) are PBB encapsulated or de-encapsulated using the PBB Etype provisioned under the related port or SDP component.

By default, the PBB Etype is 0x88e7 (which is the standard one defined in the 802.1ah, indicating that there is an I-TAG in the payload) but this PBB Etype can be changed if required due to interoperability reasons. This is the way to change it at port and/or SDP level:

```
A:MTU-4# configure port 1/1/3 ethernet pbb-etype
  - pbb-etype <0x0600..0xffff>
  - no pbb-etype

 <0x0600..0xffff>     : [1536..65535] - accepts in decimal or hex

A:MTU-4# configure service sdp 41 pbb-etype
  - no pbb-etype [<0x0600..0xffff>]
  - pbb-etype <0x0600..0xffff>

 <0x0600..0xffff>     : [1536..65535] - accepts in decimal or hex
```

The following commands are useful to check the actual PBB etype.

```
A:MTU-4# show service sdp 41 detail | match PBB
```

```
Bw BookingFactor     : 100                  PBB Etype          : 0x88e7
A:MTU-4#

A:MTU-4# show port 1/1/3 | match PBB
PBB Ethertype      : 0x88e7
A:MTU-4#
```

Before configuring the Epipe itself, the operator can optionally configure MAC names under the PBB context. MAC names will simplify the Epipe provisioning later on and in case of any change on the remote node MAC address, only one configuration modification is required as opposed as one change per affected Epipe (potentially thousands of Epipes which are terminated onto the same remote node). The MAC names are configured in the service PBB CLI context:

```
*A:MTU-4# configure service pbb mac-name
  - mac-name <name> <ieee-address>
  - no mac-name <name>

 <name>               : 32 char max
 <ieee-address>       : xx:xx:xx:xx:xx:xx or xx-xx-xx-xx-xx-xx
```

```
configure
    service
        pbb
            mac-name "MTU-4" 00:04:04:04:04:04
            mac-name "MTU-5" 00:05:05:05:05:05
            mac-name "MTU-6" 00:06:06:06:06:06
```

It is not required to configure a node with its own MAC address, so on MTU-4, the line defining the mac-name MTU-4 can be omitted.

Epipe Configuration

Once the common B-VPLS is configured, the next step is the provisioning of the customer Epipe instances. For PBB-Epipes, the I-component or Epipe is composed of an I-SAP and a PBB tunnel endpoint which points to the backbone destination MAC address (B-DA).

The following outputs show the relevant CLI configuration for the two Epipe instances represented in Setup Detailed View. The Epipe instances are configured on the MTU devices, whereas the core PEs are kept as customer-unaware nodes.

Epipes 3 and 4 are configured on MTU-6 as follows :

```
configure
    service
        epipe 3 customer 1 create
            description "pbb epipe number 3"
            pbb
```

```
                tunnel 101 backbone-dest-mac "MTU-4" isid 3
            exit
            sap 1/1/1:9 create
            exit
            no shutdown
        exit
        epipe 4 customer 1 create
            description "pbb epipe number 4"
            pbb
                tunnel 101 backbone-dest-mac "MTU-5" isid 4
            exit
            sap 1/1/1:10 create
            exit
            no shutdown
        exit
```

The following shows the relevant configuration on MTU-4 and MTU-5.

On MTU-4:

```
configure
    service
        epipe 3 customer 1 create
            description "pbb epipe number 3"
            pbb
                tunnel 101 backbone-dest-mac "MTU-6" isid 3
            exit
            sap 1/1/1:7 create
            exit
            no shutdown
        exit
```

On MTU-5:

```
configure
    service
        epipe 4 customer 1 create
            description "pbb epipe number 4"
            pbb
                tunnel 101 backbone-dest-mac "MTU-6" isid 4
            exit
            sap 1/1/1:8 create
            exit
            no shutdown
        exit
```

All Ethernet SAPs supported by a regular Epipe are also supported in the PBB Epipe. Spoke SDPs are not supported in PBB-Epipes, for example, no spoke SDP is allowed when PBB tunnels are configured on the Epipe.

The PBB tunnel links the SAP configured to the B-VPLS 101 existing in the core. The following parameters are accepted in the PBB tunnel configuration:

```
*A:MTU-5# configure service epipe 4 pbb tunnel
```

```
 - no tunnel
 - tunnel <service-id> backbone-dest-mac <mac-name> isid <ISID>
 - tunnel <service-id> backbone-dest-mac <ieee-address> isid <ISID>

<service-id>         : [1..2148007978]|<svc-name:64 char max>
<mac-name>           : 32 char max
<ieee-address>       : xx:xx:xx:xx:xx:xx or xx-xx-xx-xx-xx-xx
<ISID>               : [0..16777215]
```

Where:

- The service-id matches the B-VPLS ID.
- The **backbone-dest-mac** can be given by a MAC name (as in this configuration example) or the MAC address itself. It is recommended to use MAC names, as explained in the previous section.
- The ISID must be specified.

Flood Avoidance in PBB-Epipes

As already discussed in the previous section, when provisioning a PBB Epipe, the remote **backbone-dest-mac** must be explicitly configured on the PBB tunnel so that the ingress PBB node can build the 802.1ah encapsulation.

If the configured remote backbone-destination-mac is not known in the local FDB, the Epipe customer frames will be 802.1ah encapsulated and flooded into the B-VPLS until the MAC is learned. As previously stated, MMRP does not help to minimize the flooding because the PBB Epipes always use the configured **backbone-destination-mac** for flooding traffic as opposed to the group B-MAC derived from the ISID.

Flooding could be indefinably prolonged in the following cases:

- Configuration mistake of the **backbone-destination-mac**. The service will not work, but the operator will not detect the mistake, because the customer traffic is not dropped at the source node. Every single frame is turned into an unknown unicast PBB frame and therefore flooded into the B-VPLS domain.
- Change the **backbone-smac** in the remote PE B-VPLS instance.
- There is only unidirectional traffic in the Epipe service. In this case, the backbone-dest-mac will never be learned in the local SFIB and the frames will always be flooded into the B-VPLS domain.
- The remote node owning the **backbone-destination-mac** simply goes down.

In any of those cases, the operator can easily check whether the PBB Epipe is flooding into the B-VPLS domain, just by looking at the flood flag in the following command output:

```
*A:MTU-4# show service id 3 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 3                   Vpn Id            : 0
Service Type      : Epipe
---snip---

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/1:7                              q-tag        1518    1518    Up   Up

-------------------------------------------------------------------------------
PBB Tunnel Point
-------------------------------------------------------------------------------
B-vpls     Backbone-dest-MAC Isid     AdmMTU OperState Flood Oper-dest-MAC
-------------------------------------------------------------------------------
101        MTU-6             3        2000   Up        Yes   00:06:06:06:06:06
-------------------------------------------------------------------------------
Last Status Change: 04/21/2017 09:11:55
Last Mgmt Change  : 04/21/2017 09:11:55
===============================================================================
*A:MTU-4#
```

In this particular example, the PBB Epipe 3 is flooding into the B-VPLS 101, as the flood flag indicates. The operator can also confirm that the operational destination B-MAC for the pbb-tunnel, MTU-6, has not been learned in the B-VPLS FDB:

```
*A:MTU-4# show service id 101 fdb pbb

=======================================================================
Forwarding Database, b-Vpls Service 101
=======================================================================
MAC               Source-Identifier     iVplsMACs  Epipes     Type/Age
-----------------------------------------------------------------------
No Matching Entries
=======================================================================
A:MTU-4#
```

In small B-VPLS environments (up to 20 B-VPLSs, each with 10 MC-LAGs), it is possible to configure the PBB V-VPLS MAC notification mechanism to send notification messages at regular intervals (using the renotify parameter), rather than being only event-driven. This can avoid flooding into the B-VPLS.

Flooding Cases 1 and 2 — Wrong backbone-dest-mac

Flooding cases 1 and 2 should be fixed after detecting the flooding (see previous commands) and checking the FDBs and PBB tunnel configurations.

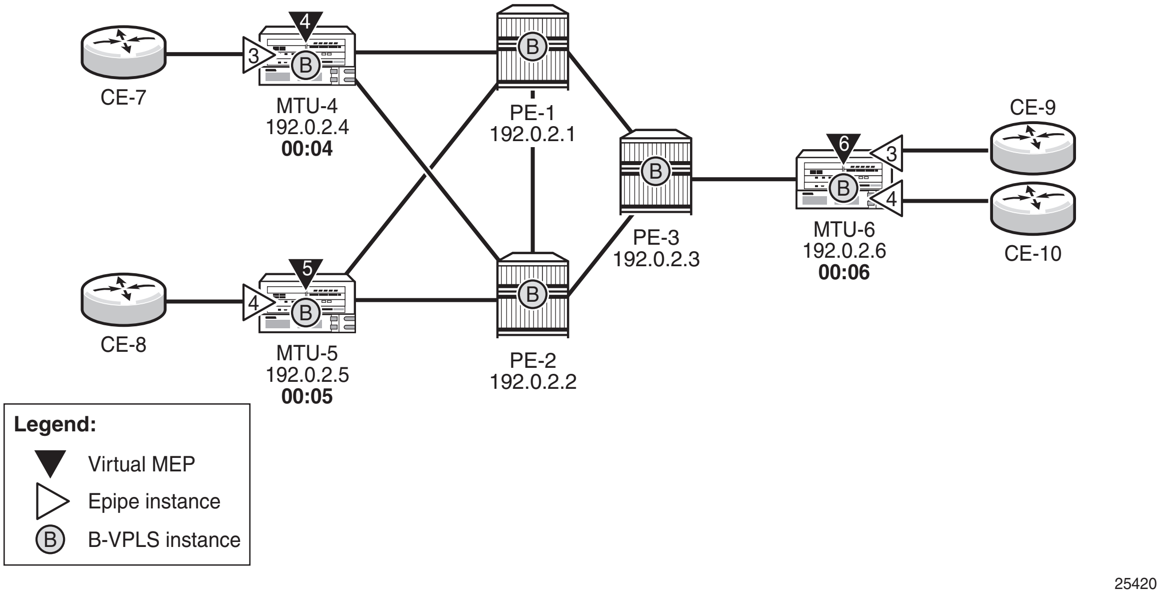
Flooding Case 3 — Unidirectional Traffic: Virtual MEP and CCM Configuration

For flooding case 3 (unidirectional traffic), Nokia recommends the use of ETH-CFM (802.1ag/Y.1731 Connectivity Fault Management) virtual Maintenance End Points (MEPs). By defining a virtual MEP per node terminating a PBB-Epipe, configuring the MEP mac-address to be the source-bmac value and activating continuity check messages (ccm), a twofold effect is achieved:

- The **pbb-tunnel backbone-destination-mac** will always be learned at the local FDB, as long as the remote virtual MEP is active and sending **cc** messages. As a result, there will not be flooding even if we have unidirectional traffic.
- An automatic proactive OAM mechanism exists to detect failures on remote nodes, which ultimately cause unnecessary flooding in the B-VPLS domain.

In the following network example, the virtual MEPs in B-VPLS 101: MEP4, MEP5 and MEP6 are configured.

Figure 116 **Virtual MEPs for Flooding Avoidance**



The following configuration example uses MTU-4. First, the general ETH-CFM configuration is made:

```
configure
    eth-cfm
        domain 1 format none level 3
            association 1 format icc-based name "B-VPLS-000101"
                bridge-identifier 101
                exit
                remote-mepid 5
                remote-mepid 6
            exit
        exit
    exit
```

Then the actual virtual MEP configuration is made:

```
configure
    service
        vpls 101
            eth-cfm
                mep 4 domain 1 association 1
                    ccm-enable
                    mac-address 00:04:04:04:04:04
                    no shutdown
                exit
            exit
        exit
```

The MAC address configured for the MEP4 matches the MAC address configured as the **source-bmac** on MTU-4, which is the **backbone-destination-mac** configured on the Epipe 3 pbb-tunnel on MTU-6. The source-bmac address on MTU-4 is 00:04:04:04:04:04, as follows:

```
configure
    service
        pbb
            source-bmac 00:04:04:04:04:04
            mac-name "MTU-4" 00:04:04:04:04:04
            mac-name "MTU-5" 00:05:05:05:05:05
            mac-name "MTU-6" 00:06:06:06:06:06
        exit
```

The backbone-dest-mac configured on MTU-6 uses MAC name “MTU-4”, which corresponds to MAC address 00:04:04:04:04:04, as follows:

```
configure
    service
        pbb
            source-bmac 00:06:06:06:06:06
            mac-name "MTU-4" 00:04:04:04:04:04
            mac-name "MTU-5" 00:05:05:05:05:05
            mac-name "MTU-6" 00:06:06:06:06:06
        exit
        epipe 3 customer 1 create
            description "pbb epipe number 3"
            pbb
                tunnel 101 backbone-dest-mac "MTU-4" isid 3
            exit
            sap 1/1/1:9 create
            exit
            no shutdown
        exit
```

Once MEP4 has been configured, check that MTU-6 is receiving CC messages from MEP4 with the following command:

```
*A:MTU-6# show eth-cfm mep 6 domain 1 association 1 all-remote-mepids

=============================================================================
Eth-CFM Remote-Mep Table
=============================================================================
R-mepId AD Rx CC RxRdi Port-Tlv If-Tlv Peer Mac Addr     CCM status since
-----------------------------------------------------------------------------
4          True  False Absent   Absent 00:04:04:04:04:04 04/21/2017 09:14:46
5          True  False Absent   Absent 00:05:05:05:05:05 04/21/2017 09:14:46
=============================================================================
Entries marked with a 'T' under the 'AD' column have been auto-discovered.
*A:MTU-6#
```

As a result of the **CC** messages coming from MEP4, the MTU-4 MAC is permanently learned in the B-VPLS 101 FDB on node MTU-6 and no flooding takes place. The following output shows that the flooding flag is not set.

```
*A:MTU-6# show service id 3 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 3                   Vpn Id            : 0
Service Type      : Epipe
---snip---

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/1:9                              q-tag        1518    1518    Up   Up

-------------------------------------------------------------------------------
PBB Tunnel Point
-------------------------------------------------------------------------------
B-vpls     Backbone-dest-MAC Isid     AdmMTU OperState Flood Oper-dest-MAC
-------------------------------------------------------------------------------
101        MTU-4             3        2000   Up        No    00:04:04:04:04:04
-------------------------------------------------------------------------------
Last Status Change: 04/21/2017 09:13:00
Last Mgmt Change  : 04/21/2017 09:13:00
===============================================================================
*A:MTU-6#
```

Flooding Case 4 — Remote Node Failure

If the node owner of the **backbone-dest-mac** fails or gets isolated, the node where the PBB Epipe is initiated will not detect the failure; that is, if MTU-4 fails, the Epipe 3 remote end will also fail but MTU-6 will not detect the failure and as a result of that, MTU-6 will flood the traffic to the network (flooding will occur after MTU-4 MAC is removed from the B-VPLS FDBs, due to either the B-VPLS flushing mechanisms or aging).

In order to avoid/reduce flooding in this case, the following mechanisms are recommended:

- Provision virtual MEPs in the B-VPLS instances terminating PBB Epipes, as already explained. This will guarantee there is no unknown B-MAC unicast being flooded under normal operation.
- CCM timers should be provisioned based on how long the service provider is willing to accept flooding.

```
*A:MTU-6# configure eth-cfm domain 1 association 1 ccm-interval
 - ccm-interval <interval>
 - no ccm-interval

 <interval>           : {10ms|100ms|1|10|60|600} - default 10 seconds
```

- It is possible to provision **discard-unknown** in the B-VPLS, so that flooded traffic due to the destination MAC being unknown in the B-VPLS is discarded immediately. This can be configured on the PEs and the MTUs. On the MTUs, it is important to configure this in conjunction with the CC messages from the virtual MEPs to ensure that the remote B-MACs are learned in both directions. If for any reason the remote B-MACs are not in the MTU B-VPLS, no traffic will be forwarded at all on the PBB-Epipe.

```
*A:PE-1# configure service vpls 101 discard-unknown
*A:PE-2# configure service vpls 101 discard-unknown
*A:PE-3# configure service vpls 101 discard-unknown
*A:MTU-4# configure service vpls 101 discard-unknown
*A:MTU-5# configure service vpls 101 discard-unknown
*A:MTU-6# configure service vpls 101 discard-unknown
```

As soon as the MTU node recovers, it will start sending CC messages and the backbone-mac will be learned on the backbone nodes and MTU nodes again.

With the recommended configuration in place, in case MTU-4 fails, the **backbone-dest-mac** configured on the pbb-tunnel for Epipe 3 on MTU-6 will be removed from the B-VPLS 101 on all the nodes (either by MAC flush mechanisms on the B-VPLS or by aging). From that point on, traffic originated from CE-9 will be discarded at MTU-6 and won’t be flooded further.

As soon as MTU-4 comes back up, MEP4 will start sending CCM and as such the MTU-4 MAC will be learned throughout the B-VPLS 101 domain and in particular in PE-1, PE-3 and MTU-6 (CCM PDUs use a multicast address). From the moment MTU-4 MAC is known on the backbone nodes and MTU-6, the traffic won’t be discarded any more, but forwarded to MTU-4.

PBB-Epipe Show Commands

The following commands can help to check the PBB Epipe configuration and their related parameters.

For the B-VPLS service:

```
*A:MTU-4# show service id 101 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 101                 Vpn Id            : 0
Service Type      : b-VPLS
Name              : (Not Specified)
```

```
Description       : (Not Specified)
Customer Id       : 1                   Creation Origin   : manual
Last Status Change: 04/21/2017 09:09:37
Last Mgmt Change  : 04/21/2017 09:15:29
Etree Mode        : Disabled
Admin State       : Up                  Oper State        : Up
MTU               : 2000                Def. Mesh VC Id   : 101
SAP Count         : 0                   SDP Bind Count    : 2
Snd Flush on Fail : Disabled            Host Conn Verify  : Disabled
SHCV pol IPv4     : None
Propagate MacFlush: Disabled            Per Svc Hashing   : Disabled
Allow IP Intf Bind: Disabled
Fwd-IPv4-Mcast-To*: Disabled            Fwd-IPv6-Mcast-To*: Disabled
Mcast IPv6 scope  : mac-based
Temp Flood Time   : Disabled            Temp Flood        : Inactive
Temp Flood Chg Cnt: 0
SPI load-balance  : Disabled
TEID load-balance : Disabled
VSD Domain        : <none>
Oper Backbone Src : 00:04:04:04:04:04
Use SAP B-MAC     : Disabled
i-Vpls Count      : 0
Epipe Count       : 1
Use ESI B-MAC     : Disabled

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sdp:41:101 S(192.0.2.1)                  Spok         8000    8000    Up   Up
sdp:42:101 S(192.0.2.2)                  Spok         8000    8000    Up   Up
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:MTU-4#
```

For the Epipe service:

```
*A:MTU-4# show service id 3 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 3                   Vpn Id            : 0
Service Type      : Epipe
Name              : (Not Specified)
Description       : pbb epipe number 3
Customer Id       : 1                   Creation Origin   : manual
Last Status Change: 04/21/2017 09:11:55
Last Mgmt Change  : 04/21/2017 09:11:55
Test Service      : No
Admin State       : Up                  Oper State        : Up
MTU               : 1514
Vc Switching      : False
SAP Count         : 1                   SDP Bind Count    : 0
Per Svc Hashing   : Disabled
Vxlan Src Tep Ip  : N/A
Force QTag Fwd    : Disabled
```

```
-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/1:7                              q-tag        1518    1518    Up   Up

-------------------------------------------------------------------------------
PBB Tunnel Point
-------------------------------------------------------------------------------
B-vpls     Backbone-dest-MAC Isid     AdmMTU OperState Flood Oper-dest-MAC
-------------------------------------------------------------------------------
101        MTU-6             3        2000   Up        No    00:06:06:06:06:06
-------------------------------------------------------------------------------
Last Status Change: 04/21/2017 09:11:55
Last Mgmt Change  : 04/21/2017 09:11:55
===============================================================================
*A:MTU-4#
```

The following command shows all the Epipe instances multiplexed into a particular B-VPLS and its status.

```
*A:MTU-4# show service id 101 epipe

===============================================================================
Related Epipe services for b-Vpls service 101
===============================================================================
Epipe SvcId         Oper ISID           Admin               Oper
-------------------------------------------------------------------------------
3                   3                   Up                  Up
-------------------------------------------------------------------------------
Number of Entries : 1
-------------------------------------------------------------------------------
===============================================================================
*A:MTU-4#
```

To check the virtual MEP information, show the local virtual MEPs configured on the node:

```
*A:MTU-4# show eth-cfm cfm-stack-table all-virtuals
===============================================================================
CFM Stack Table Defect Legend:
R = Rdi, M = MacStatus, C = RemoteCCM, E = ErrorCCM, X = XconCCM
A = AisRx, L = CSF LOS Rx, F = CSF AIS/FDI rx, r = CSF RDI rx
G = receiving grace PDU (MCC-ED or VSM) from at least one peer

===============================================================================
CFM Virtual Stack Table
===============================================================================
Service            Lvl Dir Md-index   Ma-index  MepId  Mac-address     Defect G
-------------------------------------------------------------------------------
101                  3 U          1          1    4 00:04:04:04:04:04 ------- -
===============================================================================
*A:MTU-4#
```

The following command shows all the information related to the remote MEPs configured in the association, for example, the remote virtual MEPs configured in MTU-5 and MTU-6:

```
*A:MTU-4# show eth-cfm mep 4 domain 1 association 1 all-remote-mepids

=============================================================================
Eth-CFM Remote-Mep Table
=============================================================================
R-mepId AD Rx CC RxRdi Port-Tlv If-Tlv Peer Mac Addr     CCM status since
-----------------------------------------------------------------------------
5          True  False Absent   Absent 00:05:05:05:05:05 04/21/2017 09:13:39
6          True  False Absent   Absent 00:06:06:06:06:06 04/21/2017 09:13:39
=============================================================================
Entries marked with a 'T' under the 'AD' column have been auto-discovered.
*A:MTU-4#
```

The following command shows the detail information and status of the local virtual MEP configured in MTU-4:

```
*A:MTU-4# show eth-cfm mep 4 domain 1 association 1
===============================================================================
Eth-Cfm MEP Configuration Information
===============================================================================
Md-index         : 1                        Direction         : Up
Ma-index         : 1                        Admin             : Enabled
MepId            : 4                        CCM-Enable        : Enabled
SvcId            : 101
Description      : (Not Specified)
FngAlarmTime     : 0                        FngResetTime      : 0
FngState         : fngReset                 ControlMep        : False
LowestDefectPri  : macRemErrXcon            HighestDefect     : none
Defect Flags     : None
Mac Address      : 00:04:04:04:04:04        Collect LMM Stats : disabled
LMM FC Stats     : None
LMM FC In Prof   : None
TxAis            : noTransmit               TxGrace           : noTransmit
Facility Fault   : disabled
CcmLtmPriority   : 7                        CcmPaddingSize    : 0 octets
CcmTx            : 16                       CcmSequenceErr    : 0
CcmTxIfStatus    : Absent                   CcmTxPortStatus   : Absent
CcmTxRdi         : False                    CcmTxCcmStatus    : transmit
CcmIgnoreTLVs    : (Not Specified)
Fault Propagation: disabled                 FacilityFault     : n/a
MA-CcmInterval   : 10                       MA-CcmHoldTime    : 0ms
MA-Primary-Vid   : Disabled
Eth-1Dm Threshold: 3(sec)                   MD-Level          : 3
Eth-1Dm Last Dest: 00:00:00:00:00:00
Eth-Dmm Last Dest: 00:00:00:00:00:00
Eth-Ais          : Disabled
Eth-Ais Tx defCCM: allDef
Eth-Tst          : Disabled
Eth-CSF          : Disabled

Eth-Cfm Grace Tx : Enabled                  Eth-Cfm Grace Rx  : Enabled
Eth-Cfm ED Tx    : Disabled                 Eth-Cfm ED Rx     : Enabled
Eth-Cfm ED Rx Max: 0
```

```
Eth-Cfm ED Tx Pri: CcmLtmPri (7)

Redundancy:
    MC-LAG State : n/a

CcmLastFailure Frame:
    None

XconCcmFailure Frame:
    None
===============================================================================
*A:MTU-4#
```

When there is a failure on a remote Epipe node, as described, the source node keeps sending traffic. The 802.1ag/Y.1731 virtual MEP configured can help to detect and troubleshoot the problem. For instance, when a failure happens in MTU-6 (node goes down or the B-VPLS instance is shut down), the virtual MEP show commands will show the following information:

```
*A:MTU-6# configure service vpls 101 shutdown


*A:MTU-4# show eth-cfm mep 4 domain 1 association 1
===============================================================================
Eth-Cfm MEP Configuration Information
===============================================================================
Md-index         : 1                        Direction         : Up
Ma-index         : 1                        Admin             : Enabled
MepId            : 4                        CCM-Enable        : Enabled
SvcId            : 101
Description      : (Not Specified)
FngAlarmTime     : 0                        FngResetTime      : 0
FngState         : fngDefectReported        ControlMep        : False
LowestDefectPri  : macRemErrXcon            HighestDefect     : defRemoteCCM
Defect Flags     : bDefRDICCM bDefRemoteCCM
Mac Address      : 00:04:04:04:04:04        Collect LMM Stats : disabled
LMM FC Stats     : None
LMM FC In Prof   : None
TxAis            : noTransmit               TxGrace           : noTransmit
Facility Fault   : disabled
CcmLtmPriority   : 7                        CcmPaddingSize    : 0 octets
CcmTx            : 39                       CcmSequenceErr    : 0
CcmTxIfStatus    : Absent                   CcmTxPortStatus   : Absent
CcmTxRdi         : True                     CcmTxCcmStatus    : transmit
CcmIgnoreTLVs    : (Not Specified)
Fault Propagation: disabled                 FacilityFault     : n/a
MA-CcmInterval   : 10                       MA-CcmHoldTime    : 0ms
MA-Primary-Vid   : Disabled
Eth-1Dm Threshold: 3(sec)                   MD-Level          : 3
Eth-1Dm Last Dest: 00:00:00:00:00:00
Eth-Dmm Last Dest: 00:00:00:00:00:00
Eth-Ais          : Disabled
Eth-Ais Tx defCCM: allDef
Eth-Tst          : Disabled
Eth-CSF          : Disabled

Eth-Cfm Grace Tx : Enabled                  Eth-Cfm Grace Rx  : Enabled
Eth-Cfm ED Tx    : Disabled                 Eth-Cfm ED Rx     : Enabled
```

```
Eth-Cfm ED Rx Max: 0
Eth-Cfm ED Tx Pri: CcmLtmPri (7)

Redundancy:
    MC-LAG State : n/a

CcmLastFailure Frame:
    None

XconCcmFailure Frame:
    None
===============================================================================
*A:MTU-4#
```

The bDefRemoteCCMdefect flag clearly shows that there is a remote MEP in the association which has stopped sending CCMs. In order to find out which node is affected, see the following output:

```
*A:MTU-4# show eth-cfm mep 4 domain 1 association 1 all-remote-mepids

=============================================================================
Eth-CFM Remote-Mep Table
=============================================================================
R-mepId AD Rx CC RxRdi Port-Tlv If-Tlv Peer Mac Addr     CCM status since
-----------------------------------------------------------------------------
5          True  True  Absent   Absent 00:05:05:05:05:05 04/21/2017 09:13:39
6          False False Absent   Absent 00:00:00:00:00:00 04/21/2017 09:17:58
=============================================================================
Entries marked with a 'T' under the 'AD' column have been auto-discovered.
*A:MTU-4#
```

CCMs are no longer received from virtual MEP 6 (the one defined in MTU-6) since 04/21/2017 09:17:58. This conveys which node has failed and when it failed.

Conclusion

Point-to-Point Ethernet services can use the same operational model followed by PBB VPLS for multipoint services. In other words, Epipes can be linked to the same B-VPLS domain being used by I-VPLS instances and use the existing H-VPLS network infrastructure in the core. The use of PBB Epipes reduces dramatically the number of services and pseudowires in the core and therefore allows the service provider to scale the number of E-Line services in the network.

The example used in this document shows the configuration of the PBB Epipes as well as all the related features which are required for this environment. Show commands have also been suggested so that the operator can verify and troubleshoot the service.

PBB-EVPN ISID-based CMAC Flush

This chapter provides information about PBB-EVPN ISID-based CMAC Flush.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

The information and configuration in this chapter is based on SR OS Release 15.0.R4. PBB-EVPN ISID-based CMAC flush is supported on the following objects in an I-VPLS:

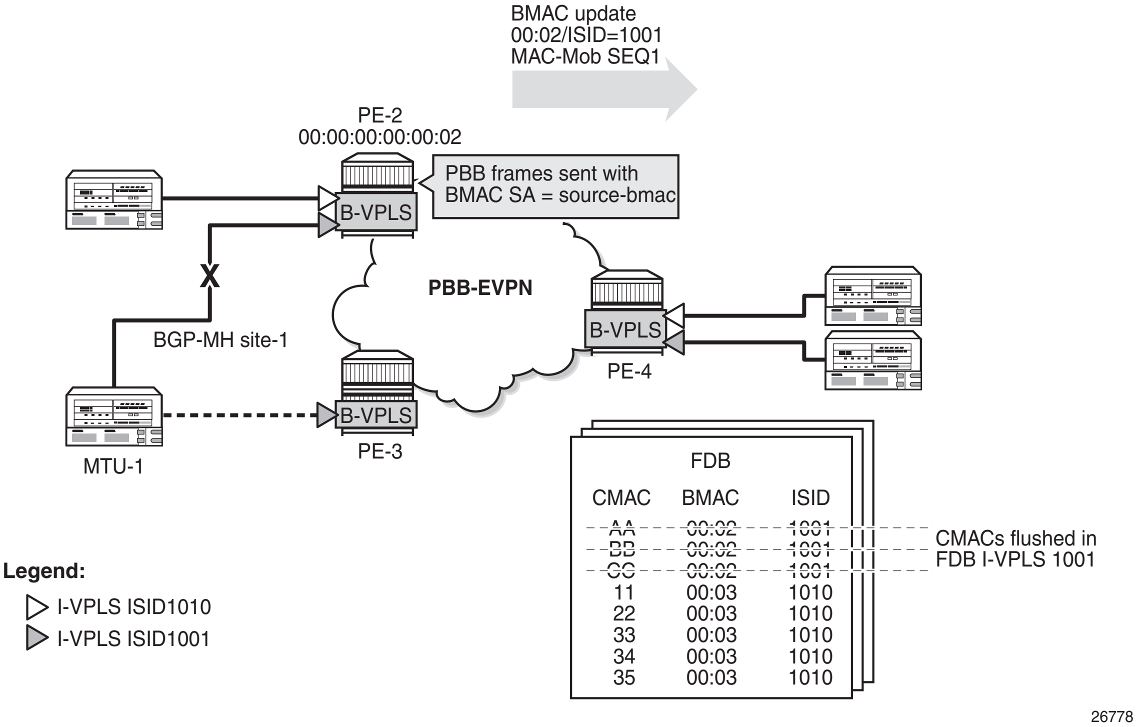
- SAPs in a BGP multi-homing site (no Ethernet Segment (ES))-supported in SR OS Release 14.0.R4, and later
- SAPs in ESs or virtual ESs (vESs)-SR OS Release 15.0.R1, and later
- Spoke-SDPs (that may be part of an ES/vES or not)-SR OS Release 15.0.R4, and later.

Chapter EVPN for PBB over MPLS (PBB-EVPN) is prerequisite reading.

Overview

Figure 117 shows an example topology with PBB-EVPN where a CMAC flush is triggered after a SAP in a BGP multi-homing site fails.

Figure 117 **CMAC flush when SAP in BGP Multi-homing Site Fails**



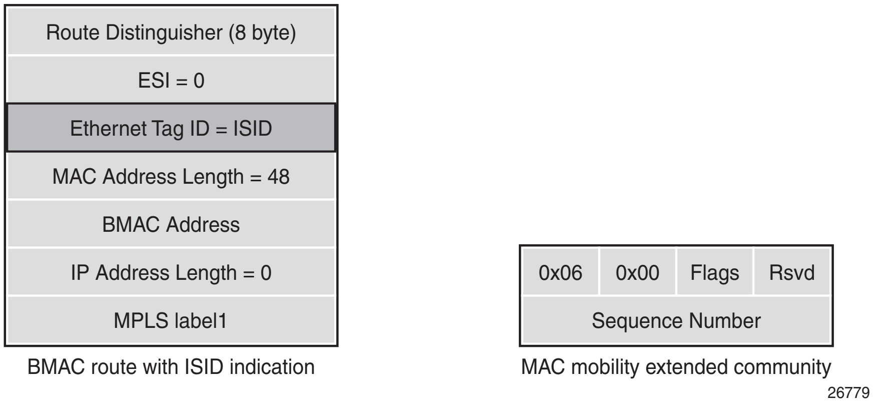
FDB

CMAC	BMAC	ISID
~~AA~~	~~00:02~~	~~1001~~
~~BB~~	~~00:02~~	~~1001~~
~~CC~~	~~00:02~~	~~1001~~
11	00:03	1010
22	00:03	1010
33	00:03	1010
34	00:03	1010
35	00:03	1010

I-VPLS 1001 is configured in PE-2 and PE-3 with **send-bvpls-evpn-flush** and connected to MTU-1. In the example, the SAP goes operationally down in I-VPLS 1001 on PE-2. To speed up convergence without flushing CMAC addresses in other I-VPLS services, PE-2 sends a BGP-EVPN BMAC route for ISID 1001 with increased sequence number to trigger a MAC-flush for I-VPLS 1001 on the remote PEs. All CMAC addresses in the FDB for other I-VPLS services, such as I-VPLS 1010 in this example, will be preserved. When PE-4 needs to send traffic to one of the flushed CMAC addresses in I-VPLS 1001, it will flood the frames until the CMAC address is learned again (via PE-3).

When SAPs or SDP-bindings-associated with ESs, vESs, or BGP-MH sites-in an I-VPLS service fail, a BGP-EVPN BMAC route (route type 2) can trigger an ISID-based CMAC flush on the remote PEs. For the CMAC addresses to be flushed from the FDB of the I-VPLS, the existing EVPN BMAC routes will be used with the Ethernet tag equal to the ISID. Figure 118 shows the EVPN BMAC route with ISID indication (BMAC/ISID). A BMAC/ISID update may trigger a selective MAC-flush for a specific I-VPLS, whereas a BMAC/0 update (BMAC/ISID route where ISID=0) may trigger a MAC-flush for all I-VPLS services. This procedure is based on *draft-snr-bess-pbb-evpn-isid-cmacflush*.

Figure 118 **EVPN BMAC Route with ISID Indication**



By default, ISID-based CMAC flush is disabled: no I-VPLS will send a B-VPLS EVPN flush message and no B-VPLS will accept any I-VPLS EVPN flush messages. The router only installs CMAC entries corresponding to a zero Ethernet tag and ignores non-zero Ethernet tag MAC routes. However, when the B-VPLS is configured to accept BMAC/ISID routes, non-zero Ethernet tag BMAC routes can be processed for CMAC flush. The CMAC flush trigger will be an EVPN BMAC/ISID route with a sequence number that is higher than before. The receiving PE will then flush all CMACs associated with this BMAC address in the I-VPLS.

The first time that a BMAC/ISID route is received, it is added to the database as a baseline. It does not cause a CMAC flush. Only subsequent BMAC/ISID updates with increased sequence number or withdrawals will cause CMAC flush.

The following command shows that B-VPLS 1000 does not accept any I-VPLS EVPN flush messages. This is the default behavior.

```
*A:PE-2# show service id 1000 bgp-evpn | match "Accept IVPLS Flush"
Accept IVPLS Flush : Disabled
```

At the receiving node, B-VPLS 1000 will accept BMAC/ISID routes when the following command is configured:

```
*A:PE-2# configure service vpls 1000 bgp-evpn accept-ivpls-evpn-flush
```

By default, I-VPLS 1001 will not send any B-VPLS EVPN flush messages, as follows:

```
*A:PE-2# show service id 1001 base | match SendBvplsEvpnFlush
SendBvplsEvpnFlush: Disabled
```

The following configuration allows I-VPLS 1001 to send B-VPLS EVPN flush messages when a SAP or SDP-binding fails:

```
*A:PE-2# configure service vpls 1001 pbb send-bvpls-evpn-flush
```

When enabled, the I-VPLS will send a BMAC/ISID route and subsequent updates with a higher sequence number whenever a SAP fails in the I-VPLS on the node. The default setting for a SAP allows a B-VPLS EVPN flush message to be sent (when enabled in the I-VPLS itself):

```
*A:PE-2# show service id 1001 sap 1/2/1:1001 detail | match SendBvplsEvpnFlush
SendBvplsEvpnFlush : Enabled
```

When no alternative route via another node is available for specific SAPs (single-homed SAPs), no CMAC flush should be triggered. When no B-VPLS EVPN flush messages need to be sent from PE-4 when SAP 1/2/1:1001 goes down, the configuration is as follows:

```
*A:PE-4# configure service vpls 1001 sap 1/2/1:1001 disable-send-bvpls-evpn-flush
```

The router only installs the BMACs received in MAC routes that have Ethernet tag zero. When CMAC flush is enabled, MAC routes with Ethernet tag equal to the ISID (always non-zero) are for CMAC flush, but not for installing the conveyed BMACs.

BMAC/ISID routes have the following characteristics:

- BMAC/ISID routes are sent with the static bit flag set as for any other BMAC route. The static bit is ignored at reception because this route is never used to install a BMAC in the FDB.
- BMAC/ISID routes received with non-zero ESI and non-zero Ethernet tag are treated as withdraw by the router at application level. Route Reflectors (RRs) treat such BMAC/ISID routes as valid routes that can be forwarded.
- BMAC/ISID routes are shown as valid in the **show router bgp routes evpn mac** commands, as in the following output, even though they are not used to populate the FDB. This shows that BGP is sending the routes to the application layer for CMAC flush processing. The BMAC/0 route should be sent before the BMAC/ISID routes for the same BMAC. Also, when the B-VPLS goes operationally down, the BMAC/0 should be withdrawn before the BMAC/ISID routes.

```
*A:PE-2# show router bgp routes evpn mac
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN MAC Routes
===============================================================================
Flag  Route Dist.         MacAddr           ESI
```

```
      Tag                 Mac Mobility      Label1
                          Ip Address
                          NextHop
-------------------------------------------------------------------------------
u*>i  192.0.2.3:1000      00:00:00:00:00:03 ESI-0
      1001                Static            LABEL 262138
                          N/A
                          192.0.2.3
u*>i  192.0.2.3:1000      00:00:00:00:00:03 ESI-0
      0                   Static            LABEL 262138
                          N/A
                          192.0.2.3
---snip---
```

When **send-bvpls-evpn-flush** is enabled in an I-VPLS that is associated with a B-VPLS, BGP-EVPN BMAC/ISID updates will be sent when certain events take place in the I-VPLS or B-VPLS. Table 1 shows the CMAC flush transmission behavior at the egress PE.

Table 9 **CMAC Flush Transmission Behavior**

Local Event	**Send-bvpls-evpn-flush**	**SAP disable-bvpls-evpn-flush**	**Action**
Reconfigure I-VPLS: enable or disable send-bvpls-evpn-flush	Enable or disable	N/A	Send update/withdraw source BMAC/ISID with Seq=0
Associate/disassociate I-VPLS to/from B-VPLS	Enabled	N/A	Send update/withdraw source BMAC/ISID with Seq=0
I-VPLS oper-up/oper-down	Enabled	N/A	Send update/withdraw source BMAC/ISID with Seq=0
B-VPLS oper-up/oper-down	Enabled	N/A	Send update/withdraw source BMAC/ISID with Seq=0 Note: All BMACs are also advertised/withdrawn.
B-VPLS bgp-evpn mpls no shut/shut	Enabled	N/A	Send update/withdraw source BMAC/ISID with Seq=0
B-VPLS operational source BMAC change	Enabled	N/A	Send update/withdraw source BMAC/ISID with Seq=0
SAP oper-up	Enabled	N/A	No operation

Table 9 **CMAC Flush Transmission Behavior (Continued)**

Local Event	Send-bvpls-evpn-flush	SAP disable-bvpls-evpn-flush	Action
SAP oper-down	Enabled	No disable	Send update source BMAC/ISID Seq=Seq+1
	Enabled	Disable	No operation

Table 2 shows the reception behavior at the ingress PE. For the CMAC flush triggered by a BMAC/ISID update with increased sequence number, the B-VPLS in the receiving PE must be configured with **accept-ivpls-evpn-flush**. BMAC/0 refers to a BMAC route where the Ethernet Tag is 0.

Table 10 **CMAC Flush Reception Behavior**

Received Route	Action
BMAC/0 withdraw	Flush all CMACs for that BMAC
BMAC/ISID withdraw	Flush all CMACs for that BMAC and ISID
BMAC/0 update + Seq change	Flush all CMACs for that BMAC
BMAC/ISID update + Seq change	Flush all CMACs for that BMAC and ISID
BMAC/0 update + PE (NHop) change	No CMAC-flush
BMAC/ISID update + PE (NHop) change	Flush all CMACs for that BMAC and ISID

BMAC/ISID updates will trigger CMAC flush procedures regardless of the Termination Endpoint (TEP) or Route Distinguisher (RD) with which the update is received. CMAC flush will be processed even if the BMAC-ISID comes from a TEP or RD different from the BMAC/0 route. Even when the sequence number is the same as in the previous BMAC/ISID update, CMAC flush will happen when the TEP is different. When the same BMAC/ISID is received from two PEs, both are accepted and any change in sequence number causes a MAC flush. However, when the same BMAC/ISID route is received from two PEs with the same RD, BGP will select only one, so the router only sees one.

CMAC Flush for ES/vES

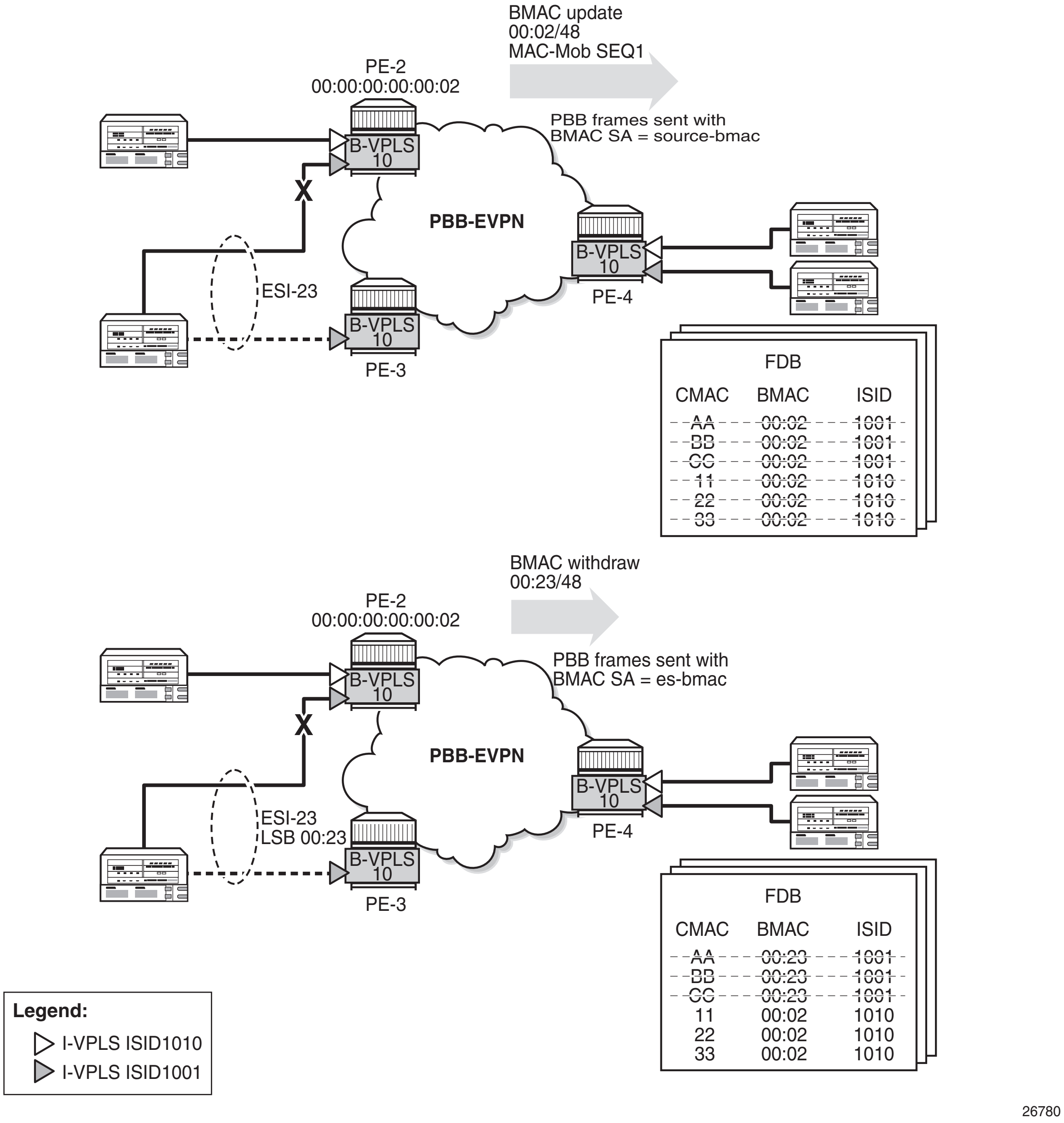
RFC7623 (PBB-EVPN) defines the following CMAC Flush notification mechanisms for single-active multi-homing. These notifications do not include the local ISIDs:

- When ES-BMACs are used and the ES goes operationally down, the ES-BMAC will be withdrawn.
- When source-BMACs are used and the ES goes operationally down, a BGP-EVPN BMAC/0 is sent with a higher sequence number.

Figure 119 shows the following two scenarios for ISID-independent CMAC flush that are supported in SR OS release 13.0.R4, and later:

- PBB frames are sent with the source-BMAC. When the ES goes operationally down, a BMAC update is sent with an increased sequence number, triggering a CMAC flush for all CMACs associated with the BMAC in I-VPLS, regardless of the ISID.
- PBB frames are sent with the ES-BMAC. When the ES goes operationally down, a BMAC withdraw message is sent, triggering the remote PEs to flush all CMACs associated to the ES-BMAC, regardless of the ISID.

Figure 119 **ISID-independent CMAC Flush when ES Fails**



In addition to the preceding ISID-independent CMAC flush mechanisms, ISID-based CMAC flush is also supported in I-VPLS services with SAP or spoke-SDPs that are part of an ES or vES. ISID-based CMAC flush is enabled in the I-VPLS with the **send-bvpls-evpn-flush** command. An I-VPLS that is configured with **send-bvpls-evpn-flush** requires one of the following conditions to be met:

- The SAP or spoke-SDP has **disable-send-bvpls-evpn-flush** configured.

- The SAP or spoke-SDP has **no disable-send-bvpls-evpn-flush** configured (default) and one of the following conditions is met:
  - The SAP or spoke-SDP is not on an ES.
  - The SAP or spoke-SDP is on an ES or vES with **no src-bmac-lsb** configured.
  - The B-VPLS has **no use-es-bmac** configured.

For ES SAPs with **no disable-send-bvpls-evpn-flush** in I-VPLS services that have **send-bvpls-evpn-flush** configured, the ISID-based CMAC flush replaces the RFC7623-based CMAC flush mechanism.

For each ES/vES and B-VPLS, the system will check whether all I-VPLS services in the ES/B-VPLS have ISID-based MAC-flush enabled.

- If all I-VPLSs have **send-bvpls-evpn-flush** enabled:
  - No BMAC/0 updates with increased sequence number will be triggered when the ES/vES goes operationally down.
  - Only BMAC/ISID updates with increased sequence number will be sent when the I-VPLS attachment circuit goes operationally down.
- If at least one I-VPLS has **no send-bvpls-evpn-flush** enabled:
  - BMAC/0 updates with increased sequence number will be triggered when the ES/vES goes operationally down.
  - Also, BMAC/ISID updates with increased sequence number will be generated for those I-VPLS services that have **send-bvpls-evpn-flush** enabled.

The number of CMACs that may be flushed at the remote nodes can be reduced by enabling ISID-based MAC-flush for all the I-VPLS services in the ES/vES.

When attempting to set **use-es-bmac** in B-VPLS 1000 on PE-4 when the SAP/SDP-binding has default settings (and **send-bvpls-evpn-flush** is enabled in the I-VPLS), the following error is raised:

```
*A:PE-4# configure service vpls 1000 pbb use-es-bmac
MINOR: SVCMGR #1433 Cannot set use-es-bmac - spoke 46:1001 on ethernet-segment ESI-
45 has "no disable-send-bvpls-evpn-flush"
```

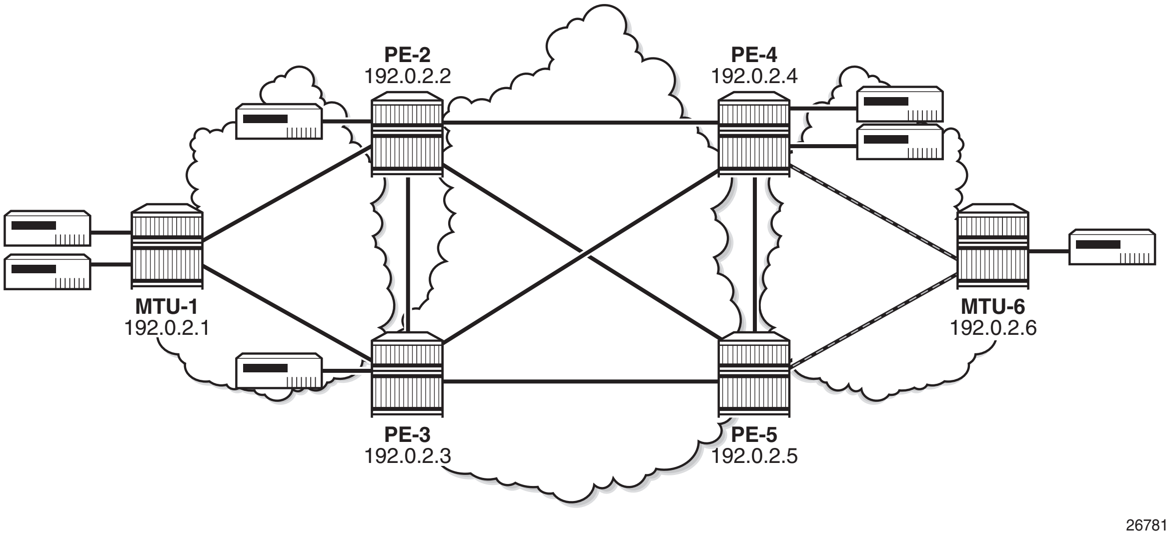
When the ES is shut down, the B-VPLS can be configured with **use-es-bmac**. When attempting to enable the ES afterward, the following error is raised.

```
*A:PE-4# configure service system bgp-evpn ethernet-segment "ESI-45" shutdown
*A:PE-4# configure service vpls 1000 pbb use-es-bmac
*A:PE-4# configure service system bgp-evpn ethernet-segment "ESI-
45" no shutdown MINOR: SVCMGR #8057 Ethernet segment cannot change admin state -
 spoke 46:1001 has "no disable-send-bvpls-evpn-flush"
```

Configuration

Figure 120 shows the example topology.

***Figure 120* Example Topology**



The initial configuration includes the following:

- Cards, MDAs
- Ports: the ports between the MTUs and the PEs are hybrid or access ports with dot1q encapsulation; the ports between the PEs are network ports with null encapsulation
- Router interfaces
- IS-IS on all router interfaces (alternatively, OSPF could be used)
- LDP on all router interfaces

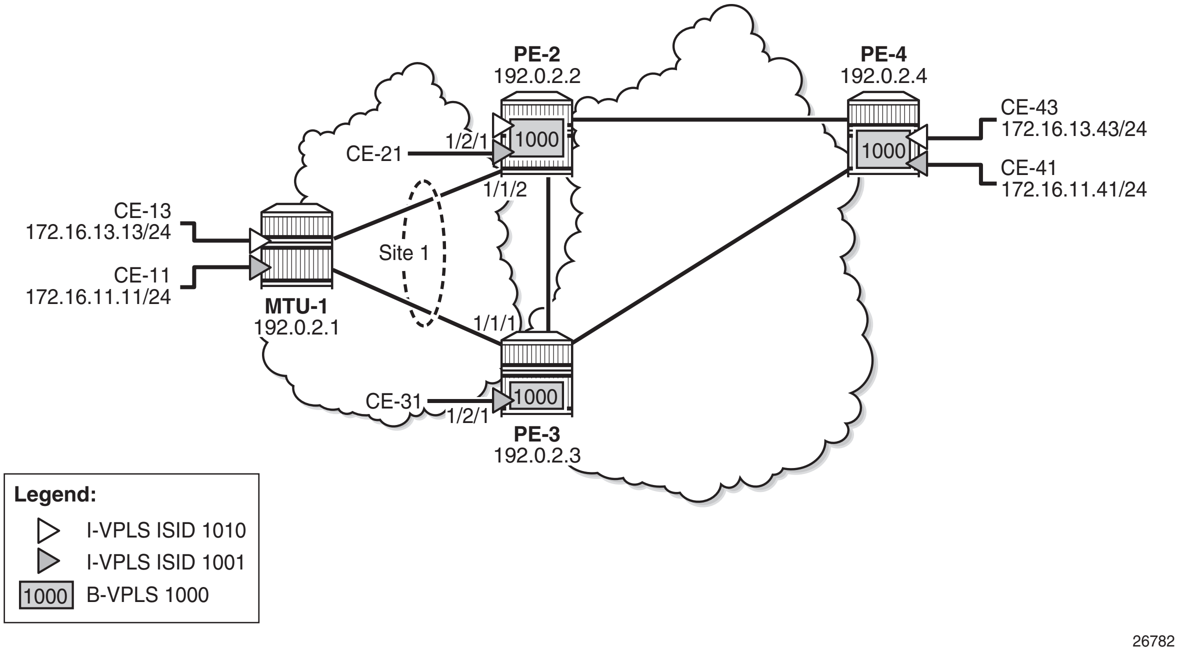
The following use cases are described in this section:

- ISID-based CMAC flush for BGP non-EVPN multi-homing (no ES)
- ISID-based CMAC flush for BGP-EVPN in a single-active ES

ISID-based CMAC Flush for BGP Multi-homing

Figure 121 shows the example topology with BGP multi-homing site 1 between PE-2 and PE-3. B-VPLS 1000 is configured on all the core nodes (PEs) and I-VPLS 1001 and I-VPLS 1010 are associated with this B-VPLS in the PEs. On MTU-1, regular VPLSs are configured. For more information about BGP non-EVPN multi-homing, see chapter BGP Multi-Homing for VPLS Networks.

Figure 121 **Example Topology with BGP Multi-homing**



BGP is configured for address family EVPN on all PEs with PE-2 as RR. For BGH multi-homing, address family L2-VPN is enabled between PE-2 and PE-3. The BGP configuration on PE-2 is as follows:

```
configure
    router
        autonomous-system 64500
        bgp
            vpn-apply-import
            vpn-apply-export
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            split-horizon
            rapid-update l2-vpn evpn
            group "internal"
                cluster 1.1.1.1
                peer-as 64500
                neighbor 192.0.2.3
                family l2-vpn evpn
            exit
            neighbor 192.0.2.4
                family evpn
            exit
        exit
    exit
```

The BGP configuration on PE-4 is as follows:

```
configure
    router
        autonomous-system 64500
        bgp
            vpn-apply-import
            vpn-apply-export
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            split-horizon
            rapid-update evpn
            group "internal"
                family evpn
                peer-as 64500
                neighbor 192.0.2.2
                exit
            exit
        exit
```

The configuration of B-VPLS 1000 and I-VPLS 1001 on PE-2 is as follows. ISID-based CMAC flush is disabled by default. BGP multi-homing site "site 1" is configured on PE-2 with SAP 1/1/2:1001 associated with it, whereas SAP 1/2/1:1001 is not associated to the MH site. CE-21 is attached to I-VPLS 1001 with SAP 1/2/1:1001.

```
configure
    service
        system
            bgp-auto-rd-range 192.0.2.2 comm-val 1 to 999
        exit
        vpls 1000 customer 1 b-vpls create
            service-mtu 2000
            pbb
                source-bmac 00:00:00:00:00:02
            exit
            bgp
            exit
            bgp-evpn
                evi 1000
                mpls
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
```

```
                no shutdown
            exit
            vpls 1001 customer 1 i-vpls create
                pbb
                    backbone-vpls 1000
                    exit
                exit
                bgp
                    route-distinguisher auto-rd
                    route-target export target:64500:1001 import target:64500:1001
                exit
                site "MH-site-1" create
                    site-id 1
                    1/1/2:1001
                    no shutdown
                exit
                sap 1/1/2:1001 create
                exit
                sap 1/2/1:1001 create
                exit
                no shutdown
            exit
            vpls 1010 customer 1 i-vpls create
                pbb
                    backbone-vpls 1000
                    exit
                exit
                bgp
                    route-distinguisher auto-rd
                    route-target export target:64500:1010 import target:64500:1010
                exit
                sap 1/1/2:1010 create
                exit
                no shutdown
            exit
        exit
```

I-VPLS 1010 is configured without multi-homing. The configuration of VPLS 1001 on PE-3 is similar, but without I-VPLS 1010.

ISID-based CMAC flush is not enabled yet. The PEs exchange BGP-EVPN MAC routes with Ethernet tag zero. PE-3 has received BMAC/0 routes from PE-2 and PE-4, as follows:

```
*A:PE-3# show router bgp routes evpn mac
===============================================================================
 BGP Router ID:192.0.2.3        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN MAC Routes
===============================================================================
Flag  Route Dist.         MacAddr           ESI
```

```
      Tag                 Mac Mobility      Label1
                          Ip Address
                          NextHop
-------------------------------------------------------------------------------
u*>i  192.0.2.2:1000      00:00:00:00:00:02 ESI-0
      0                   Static            LABEL 262138
                          N/A
                          192.0.2.2
u*>i  192.0.2.4:1000      00:00:00:00:00:04 ESI-0
      0                   Static            LABEL 262138
                          N/A
                          192.0.2.4
-------------------------------------------------------------------------------
Routes : 2
```

PE-2 and PE-4 have also received BMAC/0 routes from the other PEs.

ISID-based CMAC flush is enabled in I-VPLS 1001 on PE-2 and PE-3. PE-4 has no multi-homing in I-VPLS 1001, so it should not send any CMAC flush. I-VPLS 1010 has no multi-homing in any PE, so ISID-based MAC-flush should not be enabled in I-VPLS 1010.

```
*A:PE-2# configure service vpls 1001 pbb send-bvpls-evpn-flush
*A:PE-3# configure service vpls 1001 pbb send-bvpls-evpn-flush
```

PE-2 and PE-3 will send BMAC/1001 updates with sequence number 0 to the other two PEs. As an example, the following EVPN-MAC route for BMAC 00:00:00:00:00:03 with tag 1001 is sent by PE-3:

```
44 2017/08/08 09:25:39.792 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 96
    Flag: 0x90 Type: 14 Len: 44 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.3
        Type: EVPN-MAC Len: 33 RD: 192.0.2.3:1000 ESI: ESI-0, tag: 1001, mac len: 48
                       mac: 00:00:00:00:00:03, IP len: 0, IP: NULL, label1: 4194208
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:1000
        bgp-tunnel-encap:MPLS
        mac-mobility:Seq:0/Static
"
```

PE-4 has received the following BMAC routes from PE-2 and PE-3, with Ethernet tag zero and Ethernet tag 1001. BMAC routes are always static (received with the sticky bit set).

```
*A:PE-4# show router bgp routes evpn mac
```

```
===============================================================================
 BGP Router ID:192.0.2.4        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN MAC Routes
===============================================================================
Flag  Route Dist.         MacAddr           ESI
      Tag                 Mac Mobility      Label1
                          Ip Address
                          NextHop
-------------------------------------------------------------------------------
u*>i  192.0.2.2:1000      00:00:00:00:00:02 ESI-0
      1001                Static            LABEL 262138
                          N/A
                          192.0.2.2
u*>i  192.0.2.2:1000      00:00:00:00:00:02 ESI-0
      0                   Static            LABEL 262138
                          N/A
                          192.0.2.2
u*>i  192.0.2.3:1000      00:00:00:00:00:03 ESI-0
      1001                Static            LABEL 262138
                          N/A
                          192.0.2.3
u*>i  192.0.2.3:1000      00:00:00:00:00:03 ESI-0
      0                   Static            LABEL 262138
                          N/A
                          192.0.2.3
-------------------------------------------------------------------------------
Routes : 4
```

When a failure occurs on PE-2, PE-3, and PE-4 should accept the BMAC/ISID with increased sequence number; for a failure on PE-3, PE-2, and PE-4 should accept the BMAC/ISID update. Therefore, the B-VPLS on all PEs should accept the CMAC flush message for ISID 1001, and this is configured as follows:

```
configure service vpls 1000 bgp-evpn accept-ivpls-evpn-flush
```

The FDB for VPLS 1001 on PE-4 includes MAC address 00:00:11:11:11:11 with source-identifier 192.0.2.2:262138, so PE-4 will forward traffic toward that MAC address to PE-2.

```
*A:PE-4# show service id 1001 fdb detail

===============================================================================
Forwarding Database, Service 1001
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1001      00:00:11:11:11:11 b-eMpls:                 L/60     08/08/17 11:53:22
                            192.0.2.2:262138
```

```
1001      00:00:41:41:41:41 sap:1/2/1:1001          L/60     08/08/17 11:53:22
-------------------------------------------------------------------------------
No. of MAC Entries: 2
```

A failure is simulated on SAP 1/1/2:1001 in multi-homing site 1 on PE-2 as follows:

```
*A:PE-2# configure service vpls 1001 sap 1/1/2:1001 shutdown
```

SAP 1/1/2:1001 has the default **no disable-send-bvpls-evpn-flush** and I-VPLS 1001 is configured with **send-bvpls-evpn-flush**, so PE-2 will send BMAC/ISID updates for BMAC 00:00:00:00:00:02, ISID 1001, and sequence number 1 to its BGP peers. The following BGP update is sent by PE-2 to PE-4:

```
24 2017/08/08 12:01:46.614 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.4
"Peer 1: 192.0.2.4: UPDATE
Peer 1: 192.0.2.4 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 96
    Flag: 0x90 Type: 14 Len: 44 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.2
        Type: EVPN-MAC Len: 33 RD: 192.0.2.2:1000 ESI: ESI-
0, tag: 1001, mac len: 48 mac: 00:00:00:00:00:02, IP len: 0, IP: NULL, label1: 41942
08
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:1000
        bgp-tunnel-encap:MPLS
        mac-mobility:Seq:1/Static
"
```

This BMAC/ISID with sequence number 1 triggers a CMAC flush in the FDB for VPLS 1001, so the entry for 00:00:11:11:11:11 will be flushed, along with all other MAC addresses associated with BMAC 00:00:00:00:00:02. The FDB on PE-4 does not contain any entries with source-identifier BMAC 00:00:00:00:00:02, as follows:

```
*A:PE-4# show service id 1001 fdb detail

===============================================================================
Forwarding Database, Service 1001
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1001      00:00:41:41:41:41 sap:1/2/1:1001          L/0      08/08/17 11:01:58
-------------------------------------------------------------------------------
No. of MAC Entries: 1
```

When the MAC address 00:00:11:11:11:11 is learned via PE-3, the FDB is as follows:

```
*A:PE-4# show service id 1001 fdb detail

===============================================================================
Forwarding Database, Service 1001
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1001      00:00:11:11:11:11 b-eMpls:                 L/0      08/08/17 11:02:23
                            192.0.2.3:262138
1001      00:00:41:41:41:41 sap:1/2/1:1001           L/90     08/08/17 11:01:58
-------------------------------------------------------------------------------
No. of MAC Entries: 2
```

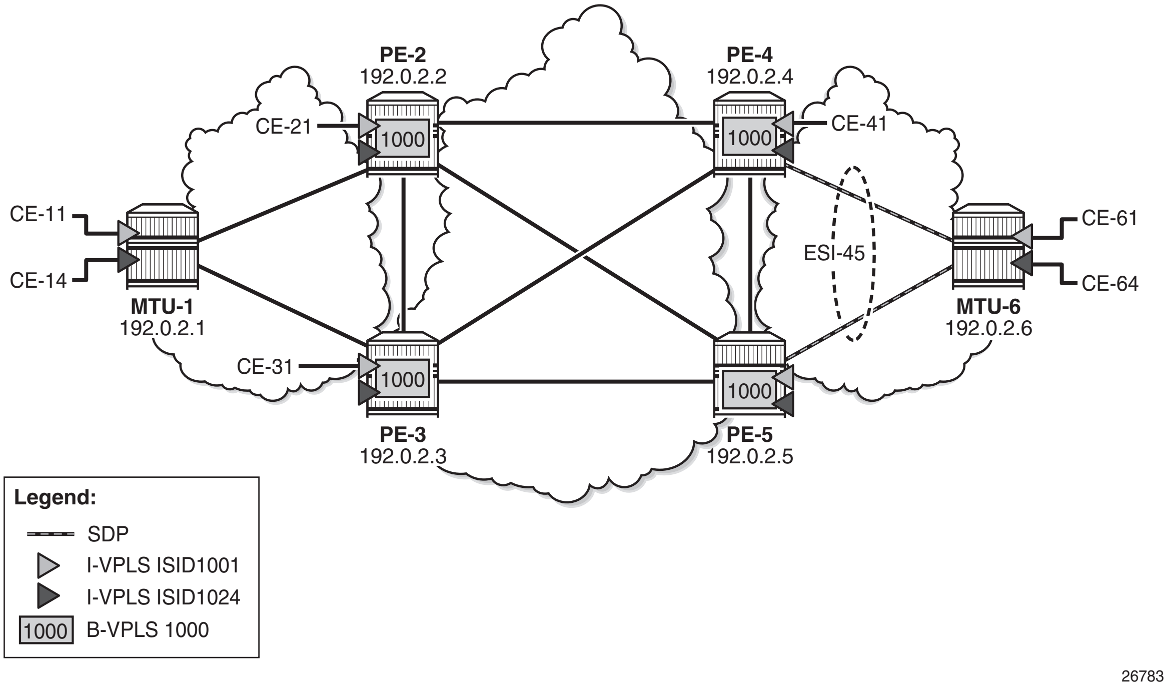
The CMAC flush is only applied for VPLS 1001, so the FDB for VPLS 1010 on PE-4 will keep entries learned from PE-2, as follows:

```
*A:PE-4# show service id 1010 fdb detail

===============================================================================
Forwarding Database, Service 1010
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1010      00:00:13:13:13:13 b-eMpls:                 L/0      08/08/17 11:28:54
                            192.0.2.2:262138
1010      00:00:43:43:43:43 sap:1/2/1:1010           L/0      08/08/17 12:02:04
-------------------------------------------------------------------------------
No. of MAC Entries: 2
```

ISID-based CMAC flush in Single-active ES

CMAC flush only makes sense for single-active multi-homing. Also, CMAC flush only works for single-active multi-homing; not for all-active multi-homing, because ES-BMAC is required in all-active multi-homing. Figure 122 shows the example topology with a single-active ES "ESI-45" configured in PE-4 and PE-5.

Figure 122 **Example Topology with Single-Active ES**



The multi-homing configuration has been removed from PE-2 and PE-3, so no CMAC flush should be sent by PE-2 or PE-3. VPLS 1001 is configured as follows on PE-2 and PE-3:

```
configure
    service
        vpls 1001 customer 1 i-vpls create
            pbb
                backbone-vpls 1000
                exit
            exit
            bgp
                route-distinguisher auto-rd
                route-target export target:64500:1001 import target:64500:1001
            exit
            sap 1/2/1:1001 create
            exit
            sap lag-1:1001 create
            exit
            no shutdown
        exit
```

SDPs are configured between PE-4 and MTU-6, and between PE-5 and MTU-6. These SDPs are associated with the single-active ES "ESI-45".

The configuration of B-VPLS 1000 on PE-4 is as follows. The B-VPLS configuration on the other PEs is similar, but with a different source BMAC.

```
configure
    service
        vpls 1000 customer 1 b-vpls create
            service-mtu 2000
            pbb
                source-bmac 00:00:00:00:00:04
            exit
            bgp
            exit
            bgp-evpn
                accept-ivpls-evpn-flush
                evi 1000
                mpls
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            no shutdown
```

The service configuration on PE-4 includes an SDP toward PE-6 and a single-active multi-homing ES, as follows:

```
configure
    service
        sdp 46 mpls create
            far-end 192.0.2.6
            ldp
            no shutdown
        exit
        system
            bgp-evpn
                ethernet-segment "ESI-45" create
                    esi 01:00:00:00:00:45:00:00:00:01
                    source-bmac-lsb 45-45 es-bmac-table-size 8
                    es-activation-timer 3
                    service-carving
                        mode auto
                    exit
                    multi-homing single-active
                    sdp 46
                    no shutdown
                exit
            exit
        exit
```

The configuration on PE-5 is similar, but with a different SDP. The configuration of B-VPLS 1000 is similar to the one for PE-2, with only a different BMAC. The configuration of I-VPLS 1001 on PE-4 is as follows:

```
configure
    service
```

```
        vpls 1001 customer 1 i-vpls create
            pbb
                backbone-vpls 1000
                exit
                send-bvpls-evpn-flush
            exit
            sap 1/2/1:1001 create
                no shutdown
            exit
            spoke-sdp 46:1001 create
                no shutdown
            exit
            no shutdown
        exit
```

ISID-based MAC-flush is enabled in B-VPLS 1000 and I-VPLS 1001 on all PEs.

I-VPLS 1024 is also associated with B-VPLS 1000 and contains one object (SAP or spoke-SDP) in each PE. The configuration of I-VPLS 1024 is identical on PE-2 and PE-3, as follows:

```
configure
    service
        vpls 1024 customer 1 i-vpls create
            pbb
                backbone-vpls 1000
                exit
            exit
            sap lag-1:1024 create
                no shutdown
            exit
            no shutdown
        exit
```

The configuration of I-VPLS 1024 on PE-4 has **send-bvpls-evpn-flush** enabled and contains a spoke-SDP instead of a SAP, as follows. The configuration on PE-5 is similar, but with a different SDP.

```
configure
    service
        vpls 1024 customer 1 i-vpls create
            pbb
                backbone-vpls 1000
                exit
                send-bvpls-evpn-flush
            exit
            spoke-sdp 46:1024 create
                no shutdown
            exit
            no shutdown
        exit
```

ISID-based MAC-flush is enabled on PE-4 and PE-5 for both I-VPLS 1001 and I-VPLS 1024, and BMAC/ISID updates are sent for ISID 1001 and ISID 1024, as follows:

```
*A:PE-3# show router bgp routes evpn mac
===============================================================================
 BGP Router ID:192.0.2.4        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN MAC Routes
===============================================================================
Flag  Route Dist.         MacAddr           ESI
      Tag                 Mac Mobility      Label1
                          Ip Address
                          NextHop
-------------------------------------------------------------------------------
---snip---
u*>i  192.0.2.4:1000      00:00:00:00:00:04 ESI-0
      1024                Static            LABEL 262136
                          N/A
                          192.0.2.4
u*>i  192.0.2.4:1000      00:00:00:00:00:04 ESI-0
      1001                Static            LABEL 262136
                          N/A
                          192.0.2.4
u*>i  192.0.2.4:1000      00:00:00:00:00:04 ESI-0
      0                   Static            LABEL 262136
                          N/A
                          192.0.2.4
---snip---
```

PE-5 is the DF for VPLS 1001 in the single-active ES "ESI-45", but not for VPLS 1024, as follows:

```
*A:PE-5# show service id 1001 ethernet-segment
No sap entries
===============================================================================
SDP Ethernet-Segment Information
===============================================================================
SDP                   Eth-Seg                           Status
-------------------------------------------------------------------------------
56:1001               ESI-45                            DF
===============================================================================
No vxlan instance entries
*A:PE-5# show service id 1024 ethernet-segment
No sap entries
===============================================================================
SDP Ethernet-Segment Information
===============================================================================
SDP                   Eth-Seg                           Status
-------------------------------------------------------------------------------
56:1024               ESI-45                            NDF
```

```
===============================================================================
No vxlan instance entries
*A:PE-5#
```

The following FDB for VPLS 1001 on PE-5 shows that traffic toward CMAC 00:00:11:11:11:11 (CE-11) in VPLS 1001 will be forwarded to PE-3:

```
*A:PE-5# show service id 1001 fdb detail

===============================================================================
Forwarding Database, Service 1001
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1001      00:00:11:11:11:11 b-eMpls:                 L/0      08/08/17 13:43:52
                            192.0.2.3:262138
1001      00:00:41:41:41:41 b-eMpls:                 L/0      08/08/17 13:14:30
                            192.0.2.4:262138
1001      00:00:61:61:61:61 sdp:56:1001              L/0      08/08/17 13:14:30
-------------------------------------------------------------------------------
No. of MAC Entries: 3
```

The following FDB for VPLS 1024 on PE-4 shows that traffic toward CMAC 00:00:14:14:14:14 (CE-14) will be forwarded to PE-3:

```
*A:PE-4# show service id 1024 fdb detail

===============================================================================
Forwarding Database, Service 1024
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1024      00:00:14:14:14:14 b-eMpls:                 L/0      08/08/17 13:41:35
                            192.0.2.3:262138
1024      00:00:64:64:64:64 sdp:46:1024              L/0      08/08/17 13:41:35
-------------------------------------------------------------------------------
No. of MAC Entries: 2
```

The following FDB for VPLS 1001 on PE-3 shows that traffic toward CMAC 00:00:61:61:61:61 (CE-61) will be forwarded to PE-5:

```
*A:PE-3# show service id 1001 fdb detail

===============================================================================
Forwarding Database, Service 1001
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1001      00:00:11:11:11:11 sap:lag-1:1001           L/0      08/08/17 13:12:27
1001      00:00:41:41:41:41 b-eMpls:                 L/0      08/08/17 13:12:27
                            192.0.2.4:262138
1001      00:00:61:61:61:61 b-eMpls:                 L/0      08/08/17 13:12:27
                            192.0.2.5:262135
```

```
-------------------------------------------------------------------------------
No. of MAC Entries: 3
```

The following FDB for VPLS 1024 on PE-3 shows that traffic toward CMAC 00:00:64:64:64:64 (CE-64) will be forwarded to PE-4:

```
*A:PE-3# show service id 1024 fdb detail

===============================================================================
Forwarding Database, Service 1024
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1024      00:00:14:14:14:14 sap:lag-1:1024           L/0      08/08/17 13:12:17
1024      00:00:64:64:64:64 b-eMpls:                 L/0      08/08/17 13:12:17
                            192.0.2.4:262136
-------------------------------------------------------------------------------
No. of MAC Entries: 2
```

PE-5 is the DF for VPLS 1001 in "ESI-45". A failure is simulated by disabling the SDP toward PE-5 on MTU-6, as follows:

```
*A:MTU-6# configure service sdp 65 shutdown
```

PE-5 sends the following BMAC/ISID with increased sequence number for ISID 1001 to the RR PE-2:

```
74 2017/08/08 14:29:08.788 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 96
    Flag: 0x90 Type: 14 Len: 44 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.5
        Type: EVPN-MAC Len: 33 RD: 192.0.2.5:1000 ESI: ESI-0, tag: 1001, mac len: 48
              mac: 00:00:00:00:00:05, IP len: 0, IP: NULL, label1: 4194160
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:1000
        bgp-tunnel-encap:MPLS
        mac-mobility:Seq:1/Static
"
```

When PE-3 receives this BMAC/ISID, all MAC routes with next-hop PE-3 are flushed and the FDB will contain the following MAC entries:

```
*A:PE-3# show service id 1001 fdb detail

===============================================================================
Forwarding Database, Service 1001
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1001      00:00:11:11:11:11 sap:lag-1:1001           L/0      08/08/17 14:27:00
1001      00:00:41:41:41:41 b-eMpls:                 L/0      08/08/17 14:20:14
                            192.0.2.4:262138
-------------------------------------------------------------------------------
No. of MAC Entries: 2
```

The configuration is restored as follows:

```
*A:MTU-6# configure service sdp 65 no shutdown
```

No CMAC/ISID update will be sent when the last SAP/SDP-binding in a service goes operationally down. VPLS 1024 only has one SAP/SDP-binding in DF PE-4: spoke-SDP 46:1024. A failure of the spoke-SDP is simulated as follows:

```
*A:MTU-6# configure service sdp 64 shutdown
```

When the last SAP/SDP-binding is down, the service will be operationally down, as follows:

```
*A:PE-4# show service id 1024 base | match "Oper State"
Admin State       : Up                  Oper State        : Down
```

PE-4 sends the following withdrawal message instead of a CMAC/ISID:

```
87 2017/08/08 14:19:47.276 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 61
    Flag: 0x90 Type: 15 Len: 57 Multiprotocol Unreachable NLRI:
        Address Family EVPN
        Type: EVPN-Incl-
mcast Len: 17 RD: 192.0.2.4:1000, tag: 1024, orig_addr len: 32, orig_addr: 192.0.2.4
        Type: EVPN-MAC Len: 33 RD: 192.0.2.4:1000 ESI: ESI-
0, tag: 1024, mac len: 48 mac: 00:00:00:00:00:04, IP len: 0, IP: NULL, label1: 0
"
```

The configuration is restored as follows:

```
*A:MTU-6# configure service sdp 64 no shutdown
```

ISID-based and Regular CMAC Flush in ES

When ISID-based CMAC flush is not enabled in all I-VPLS services using the ES, a failure in the ES will trigger BMAC/0 updates and BMAC/ISID updates with increased sequence number. An additional I-VPLS is configured on the nodes with **no send-bvpls-evpn-flush** (default). The configuration of I-VPLS 1021 on PE-5 is as follows:

```
configure
    service
        vpls 1021 customer 1 i-vpls create
            pbb
                backbone-vpls 1000
                exit
            exit
            sap 1/2/1:1021 create
            exit
            spoke-sdp 56:1021 create
            exit
            no shutdown
        exit
```

The configuration on PE-4 is similar; PE-2 and PE-3 have SAP lag-1:1021 instead of the spoke-SDP.

On MTU-6, SDP 65 is disabled, which will cause an ES failure on PE-5:

```
*A:MTU-6# configure service sdp 65 shutdown
```

The following BMAC updates are sent by PE-5:

- BMAC/0 with increased sequence number, which will trigger a CMAC flush for all entries received from PE-5 for all I-VPLS services (ISID-independent)
- BMAC/ISID with increased sequence number, which will trigger a CMAC flush for all entries received from PE-5 for VPLS 1001

```
97 2017/08/08 14:17:08.862 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 96
    Flag: 0x90 Type: 14 Len: 44 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.5
        Type: EVPN-MAC Len: 33 RD: 192.0.2.5:1000 ESI: ESI-0, tag: 0, mac len: 48
              mac: 00:00:00:00:00:05, IP len: 0, IP: NULL, label1: 4194160
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:1000
        bgp-tunnel-encap:MPLS
        mac-mobility:Seq:1/Static
```

```
"

98 2017/08/08 14:17:08.862 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 96
    Flag: 0x90 Type: 14 Len: 44 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.5
        Type: EVPN-MAC Len: 33 RD: 192.0.2.5:1000 ESI: ESI-
0, tag: 1001, mac len: 48
              mac: 00:00:00:00:00:05, IP len: 0, IP: NULL, label1: 4194160
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 24 Extended Community:
        target:64500:1000
        bgp-tunnel-encap:MPLS
        mac-mobility:Seq:4/Static
"
```

Conclusion

ISID-based MAC-flush speeds up convergence after a SAP or spoke-SDP failure, triggering a selective CMAC flush on the receiving nodes, which flushes all CMAC entries associated with that ISID and BMAC. The feature can be enabled per I-VPLS and disabled for those SAPs or spoke-SDPs for which no alternative route is available, or for those SAPs that are contained in an all-active Ethernet Segment. The BMAC/ISID update always contains the source-BMAC, not the ES-BMAC. CMAC flush based on ES-BMAC is not performed per ISID.

PBB-VPLS

This chapter provides information about Provider Backbone Bridging (PBB) in a Multi-Protocol Label Switching (MPLS) based network.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter is applicable to SR OS and was initially written for SR OS Release 7.0.R6. The CLI in this edition is based on release 15.0.R2.

➡ **Note:** Although it can be used in an MPLS-based PBB network as explained in this document, the MAC notification feature for dual-homed access is normally used in native PBB networks.

Summary

The *draft-ietf-l2vpn-pbb-vpls-pe-model-00, Extensions to LDP Signaling for PBB-VPLS*, describes the PBB-VPLS model supported by SR OS. This model expands the VPLS PE model to support PBB as defined by the IEEE 802.1ah.

PBB-VPLS combines the best of the PBB and VPLS technologies to deliver the most scalable multi-point Layer 2 VPN in the market. PBB-VPLS inherits all the benefits derived from MPLS (for example, sub-50ms Fast Reroute (FRR) protection, Traffic Engineering (TE), no need for Multiple Spanning Tree Protocol (MSTP) in the backbone) while greatly increasing the scalability of the network by providing MAC hiding, service multiplexing, and pseudowire aggregation.

The SR OS PBB-VPLS implementation also includes support for:

- Multiple MAC Registration Protocol (MMRP), application within IEEE 802.1ak for flood containment in the backbone instances, as specified in Section 6 of the *draft-ietf-l2vpn-pbb-vpls-pe-model*.
- Extensions to LDP signaling for PBB-VPLS, according to *draft-balus-l2vpn-pbb-ldp-ext-00*. These extensions will avoid network black-hole issues, as described in the Section 3 of the mentioned draft.

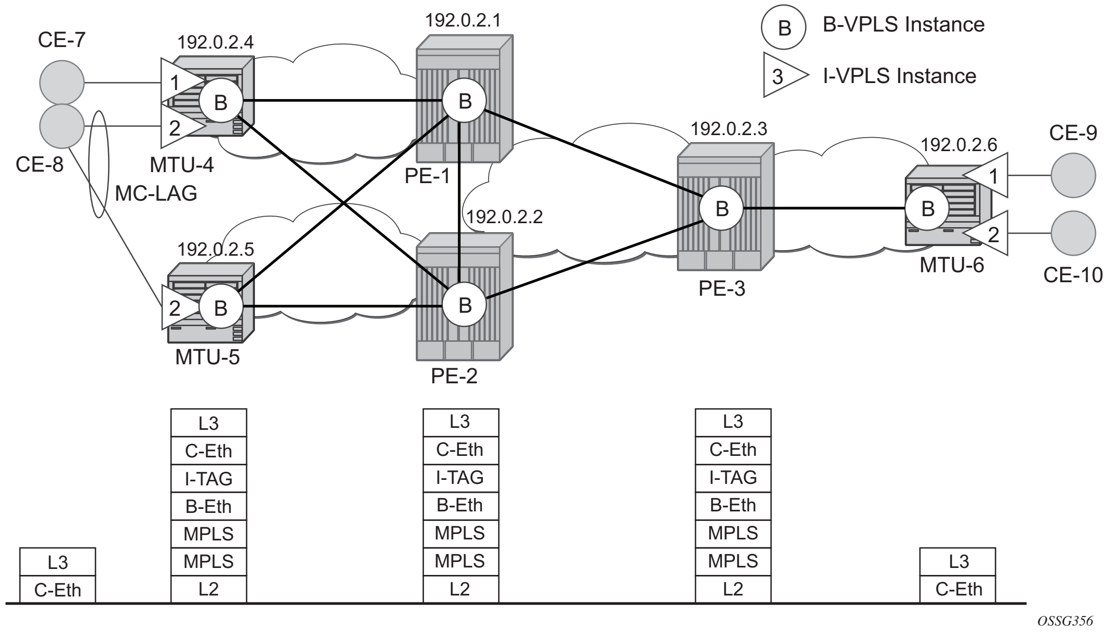
This chapter describes how to configure and troubleshoot a PBB-VPLS network.

Knowledge of the VPLS and H-VPLS (RFC 4762, *Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signalin*g) architecture and functionality is assumed throughout this chapter. The most relevant concepts will be briefly explained throughout the chapter, taking the example topology shown in the next section as an example. For further information, see the relevant Nokia documentation.

Overview

The following example topology will be used throughout the rest of the chapter.

Figure 123 **Example Topology**



The topology consists of three core nodes (PE-1, PE-2, and PE-3) and three MTU (Multi-Tenant Unit) nodes connected to the core. A backbone VPLS instance (B-VPLS 100) will be defined in all the six nodes, whereas a few customer I-VPLS instances will be defined on the three MTU nodes.

Those I-VPLS instances will be multiplexed into the common B-VPLS, using the ISID field within the I-TAG as the demultiplexer field at the egress MTU to differentiate each specific customer.

The B-VPLS domain constitutes an H-VPLS network itself, with spoke SDPs from the MTUs to the core PE layer. Active/standby spoke SDPs can be used from the MTUs to the PEs (for example, in the MTU-4 and MTU-5 cases) or single non-redundant spoke SDPs (for example, MTU-6). CE-8 is dual-connected to the service provider network through MC-LAG.

The protocol stack being used along the path between the CEs is shown in Figure 123.

Configuration

This section describes all the relevant PBB-VPLS configuration tasks for the setup shown in Figure 123. The appropriate associated IP/MPLS configuration is out of the scope of this example. In this particular example, the following protocols will be configured beforehand:

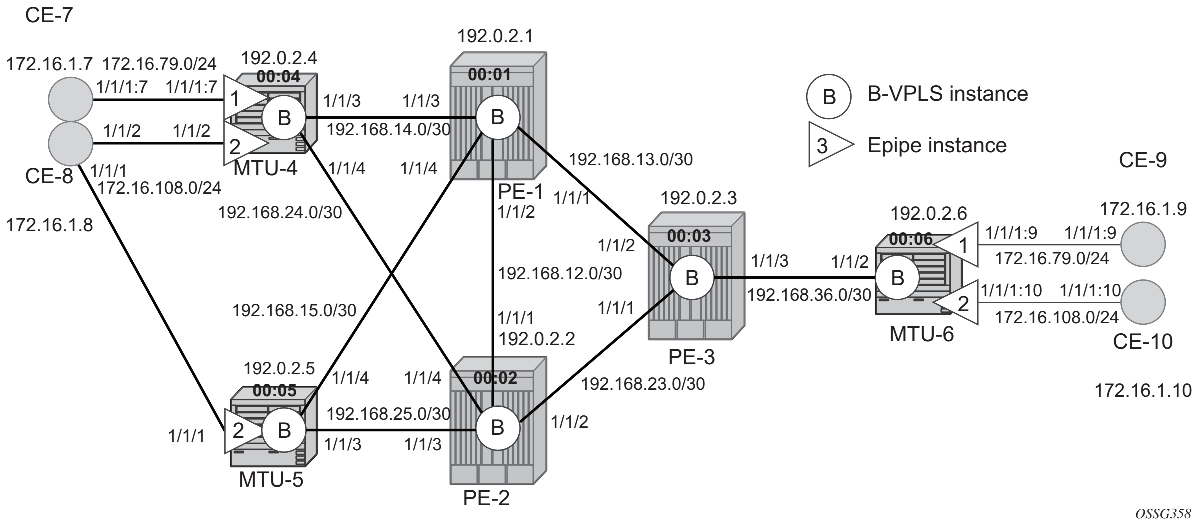
- ISIS-TE as IGP with all the interfaces being Level-2 (OSPF-TE could have been used instead).
- RSVP-TE as the MPLS protocol to signal the transport tunnels (LDP could have been used instead).
- LSPs between core PEs will be fast reroute protected (facility bypass tunnels) whereas LSP tunnels between MTUs and PEs will not be protected.
- The protection between MTU-4, MTU-5 and PE-1, PE-2 will be based on the A/S pseudowire protection configured in the B-VPLS.
- BGP is configured for auto-discovery (Layer 2-VPN family), because FEC 129 will be used for the pseudowires between PEs in the core.

Once the IP/MPLS infrastructure is up and running, the service configuration tasks described in the following sections can be implemented.

PBB-VPLS M:1 Service Configuration

This section explains the process to configure PBB-VPLS services in a M:1 fashion, M being the number of customer I-VPLS services multiplexed into the same B-VPLS instance (instance 100). An alternative configuration is 1:1, where each customer I-VPLS has its own B-VPLS. MTU-4 and PE-1 will be picked to show the relevant CLI configuration commands. The bold digits separated by colons **00:xx** are abbreviations for the backbone MAC addresses.

Figure 124 **MTU-4 and PE-1 Nodes as Configuration Examples**



B-VPLS Configuration

The first step is to configure the B-VPLS instance that will carry the PBB traffic. The following shows the B-VPLS configuration on MTU-4 and PE-1.

The configuration for B-VPLS 100 on MTU-4 is as follows:

```
configure
    service
        vpls 100 customer 1 b-vpls create
            service-mtu 2000
            pbb
                source-bmac 00:04:04:04:04:04
            exit
            endpoint "core" create
                no suppress-standby-signaling
            exit
            spoke-sdp 41:100 endpoint "core" create
                precedence primary
            exit
            spoke-sdp 42:100 endpoint "core" create
            exit
            no shutdown
        exit
```

On PE-1, B-VPLS 100 is configured as follows:

```
configure
    service
        pw-template 1 use-provisioned-sdp create
            split-horizon-group "CORE"
            exit
        exit
        vpls 100 customer 1 b-vpls create
            service-mtu 2000
            pbb
                source-bmac 00:01:01:01:01:01
            exit
            bgp
                route-target export target:65000:100 import target:65000:100
                pw-template-binding 1
                exit
            exit
            bgp-ad
                vpls-id 65000:100
                no shutdown
            exit
            spoke-sdp 14:100 create
            exit
            spoke-sdp 15:100 create
            exit
            no shutdown
        exit
```

The relevant B-VPLS commands are in bold.

The keyword **b-vpls** is given at creation time and therefore it cannot be added to a regular existing VPLS instance. Besides the **b-vpls** keyword, the B-VPLS is a regular VPLS instance in terms of configuration, with the following exceptions:

- The B-VPLS service MTU must be at least 18 bytes greater than the I-VPLS MTU of the multiplexed instances. In this example, the I-VPLS instances will have the default service MTU (1500 bytes); therefore, any MTU equal to or greater than 1518 bytes must be configured. In this particular example, a MTU of 2000 bytes is configured in the B-VPLS instance throughout the network.

- The source B-MAC is the MAC that will be sourced when the PBB traffic is originated from that node. A source B-MAC per B-VPLS instance can be configured (if there are more than one B-VPLS) or a common source B-MAC that will be shared by all the B-VPLS instances in the box. If no specific source B-MAC is provisioned, the system MAC address is used as the source B-MAC. When using the access multi-homing feature for native PBB, the source B-MAC must be a configured one and never the chassis mac address. The way to configure a common B-MAC for all the B-VPLS instances on MTU-4 is as follows:

```
configure
    service
        pbb
            source-bmac 00:04:04:04:04:04
```

The following considerations will be taken into account when configuring the B-VPLS:

- B-VPLS SAPs:
  - Ethernet dot1q and null encapsulations are supported
  - Default SAP (:*) types are blocked in the CLI for the B-VPLS SAP
- B-VPLS SDPs:
  - For MPLS, both mesh and spoke SDPs with split horizon groups are supported.
  - Similar to regular pseudowires, the outgoing PBB frame on an SDP (for example, B-pseudowire) contains a BVID qtag only if the pseudowire type is Ethernet VLAN. If the pseudowire type is **Ethernet**, the BVID qtag is stripped before the frame goes out.
  - BGP-AD is supported in the B-VPLS; therefore, spoke SDPs in the B-VPLS can be signaled using FEC 128 or FEC 129. In this example, BGP-AD and FEC 129 are used. A split-horizon group has been configured to emulate the behavior of mesh-SDPs in the core.
- If a local I-VPLS instance is associated with the B-VPLS, “local frames” originated/terminated on local I-VPLS(s) are PBB encapsulated/de-encapsulated using the PBB Ethertype provisioned under the related port or SDP component.

By default, the PBB Ethertype is 0x88e7 (which is the standard one defined in 802.1ah for the I-TAG) but this PBB Ethertype can be changed if required due to interoperability reasons. This is the way to change it at port and/or SDP level:

```
A:MTU-4# configure port 1/1/3 ethernet pbb-etype
  - pbb-etype <0x0600..0xffff>
  - no pbb-etype

 <0x0600..0xffff>     : [1536..65535] - accepts in decimal or hex
```

```
A:MTU-4# configure service sdp 41 pbb-etype
  - no pbb-etype [<0x0600..0xffff>]
  - pbb-etype <0x0600..0xffff>

 <0x0600..0xffff>     : [1536..65535] - accepts in decimal or hex
```

The following commands are useful to check the actual PBB Ethertype.

```
A:MTU-4# show service sdp 41 detail | match PBB
Bw BookingFactor     : 100                  PBB Etype          : 0x88e7

A:MTU-4# show port 1/1/3 | match PBB
PBB Ethertype      : 0x88e7
```

I-VPLS Configuration

Once the common B-VPLS is configured, the next step is to provision the customer I-VPLS instances. The following shows the relevant configuration on MTU-4 for the two I-VPLS instances represented in Figure 124. The I-VPLS instances are configured on the MTU devices, whereas the core PEs are customer-unaware nodes.

```
configure
    service
        vpls 1 customer 1 i-vpls create
            pbb
                backbone-vpls 100
                exit
            exit
            sap 1/1/1:7 create
            exit
            no shutdown
        exit
        vpls 2 customer 1 i-vpls create
            pbb
                backbone-vpls 100 isid 2
                exit
            exit
            sap lag-1 create
            exit
            no shutdown
        exit
```

The relevant I-VPLS commands are in bold.

The keyword **i-vpls** is given at creation time and therefore it cannot be added to a regular existing VPLS instance. After creating the I-VPLS instance, it has to be linked to its corresponding transport B-VPLS instance. That link is given by the **backbone-vpls** *b-vpls isid* isid command. If no ISID (20 bit customer identification in the ITAG) is specified, the system will take the VPLS instance identifier as the ISID value.

The following considerations will be taken into account when configuring the I-VPLS:

- I-VPLS SAPs:
  - SAPs can be defined on ports with any Ethernet encapsulation type (null, dot1q, and qinq)
  - The I-VPLS SAPs can coexist on the same port with SAPs for other business services, for example, VLL and VPLS SAPs.
- I-VPLS SDPs:
  - GRE and MPLS SDPs are supported.
  - No mesh SDPs are supported, only spoke SDP. Mesh SDPs can be emulated by using Split Horizon Groups (SHGs).

Existing SAP processing rules still apply for the I-VPLS case; the SAP encapsulation definition on Ethernet ingress ports defines which VLAN tags are used to determine the service that the packet belongs to:

- Null encapsulation defined on ingress — Any VLAN tags are ignored and the packet goes to a default service for the SAP;
- Dot1q encapsulation defined on ingress — only first VLAN tag is considered;
- QinQ encapsulation defined on ingress — both VLAN tags are considered; wildcard for the inner VLAN tag is supported.
- For dot1q/qinq encapsulations, traffic encapsulated with VLAN tags for which there is no definition is discarded.
- Any VLAN tag used for service selection on the I-SAP is stripped before the PBB encapsulation is added. Appropriate VLAN tags are added at the remote PBB PE when sending the packet out on the egress SAP.

Up to 8000 VPLS instances can be defined per system. That number includes I-VPLS, B-VPLS and regular VPLS.

MMRP for Flooding Optimization

When the M:1 model is used (as in this example), any I-VPLS broadcast/multicast/unknown frame is flooded throughout the B-VPLS domain regardless of the nodes where the originating I-VPLS is defined. In other words, in our example in Figure 123, any broadcast/multicast/unknown frame coming from CE-7 would be flooded in the B domain and would reach PE-2 and MTU-5, even though that traffic only needs to go to PE-3 and MTU-6. In order to build customer-based flooding trees and optimize the flooding, Multiple MAC Registration Protocol (MMRP) must be configured on the B-VPLS.

MMRP can be enabled with its default settings just by executing a **mrp no shutdown** command on all nodes:

```
configure service vpls 100 mrp no shutdown
```

There are certain B-VPLS MRP settings that can be modified. These are the default values:

```
*A:MTU-4>config>service>vpls>mrp# info detail
----------------------------------------------
                mmrp
                    no end-station-only
                    attribute-table-size 2048
                    attribute-table-low-wmark 90
                    attribute-table-high-wmark 95
                    no flood-time
                    no shutdown
                exit
                no shutdown
----------------------------------------------
*A:MTU-4>config>service>vpls>mrp#
```

These attributes can be changed in order to control the number of MMRP attributes per B-VPLS and optimize the convergence time in case of failures in the B-VPLS:

- Controlling the number of attributes per B-VPLS

  The MMRP exchanges create one entry per attribute (group B-MAC) in the B-VPLS where MMRP protocol is running. PBB uses a group B-MAC address—built using a specific OUI (00:1e:83) with the multicast bit set, and the ISID value for the last 24 bits—as a destination MAC address for flooding any Broadcast, Unknown unicast, and Multicast (BUM) frame into the B-domain.

  When the first registration is received for an attribute, an MFIB entry is created for it. The *attribute-table-size* allows the user to control the number of MMRP attributes (group B-MACs) created on a per B-VPLS basis, between 1 and 2048. Based on the configured size, high and low watermarks can be set (in percentage) so that alarms can be triggered upon exceeding the watermarks. This ensures that no B-VPLS will take up all the resources from the total pool. The maximum number of attributes per B-VPLS is 2048 and 4000 can be configured globally on the system.

- Optimizing the convergence time

  Assuming that MMRP is used in a certain B-VPLS, under failure conditions the time it takes for the B-VPLS forwarding to resume may depend on the data plane and control plane convergence plus the time it takes for MMRP exchanges to stabilize the flooding trees on a per ISID basis. In order to minimize the convergence time, the PBB SR OS implementation offers the selection of a mode where B-VPLS forwarding reverts for a short time to flooding so that MMRP has enough time to converge. This mode can be selected through configuration using the **flood-time** *value* command where value represents the amount of time in seconds (between 3 and 600) that flooding will be enabled. If this behavior is selected, the forwarding plane starts with B-VPLS flooding for a configurable time period, then it reverts back to the MFIB entries installed by MMRP. The following B-VPLS events initiate the switch from per I-VPLS (MMRP) MFIB entries to BVPLS flooding:

- Reception or local triggering of a Spanning Tree Topology Change Notification (TCN)
- B-SAP failure
- Failure of a B-SDP binding
- Pseudowire activation in a primary/standby H-VPLS resiliency solution
- SF/CPM switchover due to STP reconvergence

The IEEE 802.1ak standard, which defines MRP, requires the implementation of different state machines with associated timers that can be tuned. A full MRP participant maintains the following state machines:

- Registrar state machine
- Applicant state machine
- LeaveAll state machine
- PeriodicTransmission state machine

The two first state machines are maintained for each attribute in which the participant is interested, whereas the two latter are global to all the attributes.

The job of the registrar function is to record declarations of the attribute made by other participants on the LAN. A registrar does not send any protocol messages, because the applicant looks after the interests of all would-be participants.

The job of the applicant is twofold: first, to ensure that this participant’s declaration is correctly registered by other participants’ registrars, and next, to prompt other participants to register again after one withdraws a declaration.

The associated timers can be tuned on a per SAP/SDP basis:

```
A:MTU-4>config>service>vpls>spoke-sdp# mrp
  - mrp

 [no] join-time       - Configure timer value in 10th of seconds for sending
                        join-messages
 [no] leave-all-time  - Configure timer value in 10th of seconds for refreshing
                        all attributes
 [no] leave-time      - Configure timer value in 10th of seconds to hold
                        attribute in leave-state
 [no] mrp-policy      - Configure mrp-policy
 [no] periodic-time   - Configure timer value in 10th of seconds for
                        re-transmission of attribute declarations
 [no] periodic-timer  - Control re-transmission of attribute declarations
```

```
A:MTU-4>config>service>vpls>spoke-sdp>mrp# info detail
----------------------------------------------
                    join-time 2
                    leave-time 30
                    leave-all-time 100
                    periodic-time 10
                    no periodic-timer
                    no mrp-policy
----------------------------------------------
A:MTU-4>config>service>vpls>spoke-sdp>mrp#
```

A brief description of the MRP SAP/SDP attributes follows:

- Join-time — This command controls the interval between transmit opportunities that are applied to the applicant state machine. An instance of this join period timer is required on a per-port, per-MRP participant basis. For additional information, see IEEE 802.1ak-2007 section 10.7.4.1.
- Leave-time — This command controls the period of time that the registrar state machine will wait in the leave state before transitioning to the MT state when it is removed. An instance of the timer is required for each state machine that is in the leave state. The leave period timer is set to the value leave-time when it is started. A registration is normally in “in” state where there is an MFIB entry and traffic being forwarded. When a “leave all” is performed (periodically around every 10-15 seconds per SAP/SDP binding – see leave-all-time below), a node sends a message to its peer indicating a leave all is occurring and puts all of its registrations in leave state. The peer refreshes its registrations based on the leave all PDU it receives and sends a PDU back to the originating node with the state of all its declarations. See IEEE 802.1ak-2007 section 10.7.4.2.
- Leave-all-time — This command controls the frequency with which the leaveall state machine generates leaveall PDUs. The timer is required on a per-port, per-MRP participant basis. The leaveall period timer is set to a random value, T, in the range leavealltime<T<1.5*leave-all-time when it is started. See IEEE 802.1ak-2007, section 10.7.4.3.
- Periodic-time — This command controls the frequency the periodic transmission state machine generates periodic events if the periodic transmission timer is enabled. The timer is required on a per-port basis. The periodic transmission timer is set to one second when it is started.
- Periodic-timer — This command enables or disables the periodic transmission timer.

The following command shows the MRP configuration and statistics on a per SAP/SDP basis within the B-VPLS:

```
*A:MTU-4# show service id 100 all | match MRP post-lines 10
Sdp Id 41:100 MRP Information
-------------------------------------------------------------------------------
Join Time          : 0.2 secs                 Leave Time        : 3.0 secs
Leave All Time     : 10.0 secs                Periodic Time     : 1.0 secs
```

```
Periodic Enabled   : false
Mrp Policy         : N/A
Rx Pdus            : 2542                 Tx Pdus           : 2963
Dropped Pdus       : 0
Rx New Event       : 0                    Rx Join-In Event  : 1986
Rx In Event        : 0                    Rx Join Empty Evt : 1533543
Rx Empty Event     : 3                    Rx Leave Event    : 126
SDP MMRP Information
-------------------------------------------------------------------------------
MAC Address       Registered        Declared
-------------------------------------------------------------------------------
01:1e:83:00:00:01 Yes               Yes

01:1e:83:00:00:02 Yes               Yes
-------------------------------------------------------------------------------
Number of MACs=2 Registered=2 Declared=2
-------------------------------------------------------------------------------
Sdp Id 42:100 MRP Information
-------------------------------------------------------------------------------
Join Time          : 0.2 secs             Leave Time        : 3.0 secs
Leave All Time     : 10.0 secs            Periodic Time     : 1.0 secs
Periodic Enabled   : false
Mrp Policy         : N/A
Rx Pdus            : 0                    Tx Pdus           : 0
Dropped Pdus       : 0
Rx New Event       : 0                    Rx Join-In Event  : 0
Rx In Event        : 0                    Rx Join Empty Evt : 0
Rx Empty Event     : 0                    Rx Leave Event    : 0
SDP MMRP Information
-------------------------------------------------------------------------------
MAC Address       Registered        Declared
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Number of MACs=0 Registered=0 Declared=0
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Number of SDPs : 2
-------------------------------------------------------------------------------
* indicates that the corresponding row element may have been truncated.
Service MRP Information
===============================================================================
Admin State          : enabled
-------------------------------------------------------------------------------
MMRP
-------------------------------------------------------------------------------
Admin Status         : enabled              Oper Status      : up
Register Attr Cnt    : 2                    Declared Attr Cnt: 2
End-station-only     : disabled
Max Attributes       : 2048                 Attribute Count  : 2
Hi Watermark         : 95%                  Low Watermark    : 90%
Failed Registers     : 0                    Flood Time       : Off
-------------------------------------------------------------------------------
MVRP
-------------------------------------------------------------------------------
MRP SAP Table
=============================================================================
SAP                             Join      Leave     Leave All Periodic
                                Time(sec) Time(sec) Time(sec) Time(sec)
-----------------------------------------------------------------------------
```

```
=============================================================================
=============================================================================
MRP SDP-BIND Table
=============================================================================
SDP-BIND                        Join      Leave     Leave All Periodic
                                Time(sec) Time(sec) Time(sec) Time(sec)
-----------------------------------------------------------------------------
41:100                          0.2       3.0       10.0      1.0
42:100                          0.2       3.0       10.0      1.0
=============================================================================
===============================================================================

-------------------------------------------------------------------------------
*A:MTU-4#
```

The following command is useful to check the MRP configuration and status.

```
*A:MTU-4# show service id 100 mrp

===============================================================================
Service MRP Information
===============================================================================
Admin State          : enabled
-------------------------------------------------------------------------------
MMRP
-------------------------------------------------------------------------------
Admin Status         : enabled             Oper Status      : up
Register Attr Cnt    : 2                   Declared Attr Cnt: 2
End-station-only     : disabled
Max Attributes       : 2048                Attribute Count  : 2
Hi Watermark         : 95%                 Low Watermark    : 90%
Failed Registers     : 0                   Flood Time       : Off
-------------------------------------------------------------------------------
MVRP
-------------------------------------------------------------------------------
Admin Status         : disabled            Oper Status      : down
Max Attr             : 4095                Failed Register  : 0
Register Attr Count  : 0                   Declared Attr    : 0
Hi Watermark         : 95%                 Low Watermark    : 90%
Hold Time            : disabled            Attr Count       : 0
-------------------------------------------------------------------------------
=============================================================================
MRP SAP Table
=============================================================================
SAP                             Join      Leave     Leave All Periodic
                                Time(sec) Time(sec) Time(sec) Time(sec)
-----------------------------------------------------------------------------
=============================================================================
=============================================================================
MRP SDP-BIND Table
=============================================================================
SDP-BIND                        Join      Leave     Leave All Periodic
                                Time(sec) Time(sec) Time(sec) Time(sec)
-----------------------------------------------------------------------------
41:100                          0.2       3.0       10.0      1.0
42:100                          0.2       3.0       10.0      1.0
=============================================================================
===============================================================================
*A:MTU-4#
```

In the example throughout the chapter, as soon as MMRP is enabled, an optimized flooding tree will be built for ISID 1, because the I-VPLS 1 is only defined in MTU-4 and MTU-6, but not in MTU-5. A good way to track the flooding tree for a particular ISID is the following command:

```
*A:MTU-4# show service id 100 mmrp mac
-------------------------------------------------------------------------------
SAP/SDP                                 MAC Address       Registered  Declared
-------------------------------------------------------------------------------
sdp:41:100                              01:1e:83:00:00:01 Yes         Yes
sdp:41:100                              01:1e:83:00:00:02 Yes         Yes
-------------------------------------------------------------------------------
Number of Entries=2 SAPs=0 SDPs=2
-------------------------------------------------------------------------------
*A:MTU-4#
```

```
*A:MTU-5# show service id 100 mmrp mac
-------------------------------------------------------------------------------
SAP/SDP                                 MAC Address       Registered  Declared
-------------------------------------------------------------------------------
sdp:52:100                              01:1e:83:00:00:01 Yes         No
sdp:52:100                              01:1e:83:00:00:02 Yes         No
-------------------------------------------------------------------------------
Number of Entries=2 SAPs=0 SDPs=2
-------------------------------------------------------------------------------
*A:MTU-5#
```

The group B-MAC ending in **01** corresponds to the I-VPLS 1 whereas the one ending in **02** to the I-VPLS 2. MMRP PDUs for the two attributes are sent throughout the loop-tree topology (not over STP blocked ports or standby spoke SDPs and observing the split horizon rules). The two attributes are registered on every B-VPLS virtual port; however, the tree is only built on those ports where the attribute is also declared, and not only registered. For instance, the spoke SDP 52:100 in MTU-5 will not be part of the ISID 1 or ISID 2 flooding trees. Neither attribute is declared because I-VPLS 1 doesn’t exist on MTU-5 and I-VPLS 2 is operationally down on MTU-5 (MC-LAG SAP is in standby state, so the I-VPLS down).

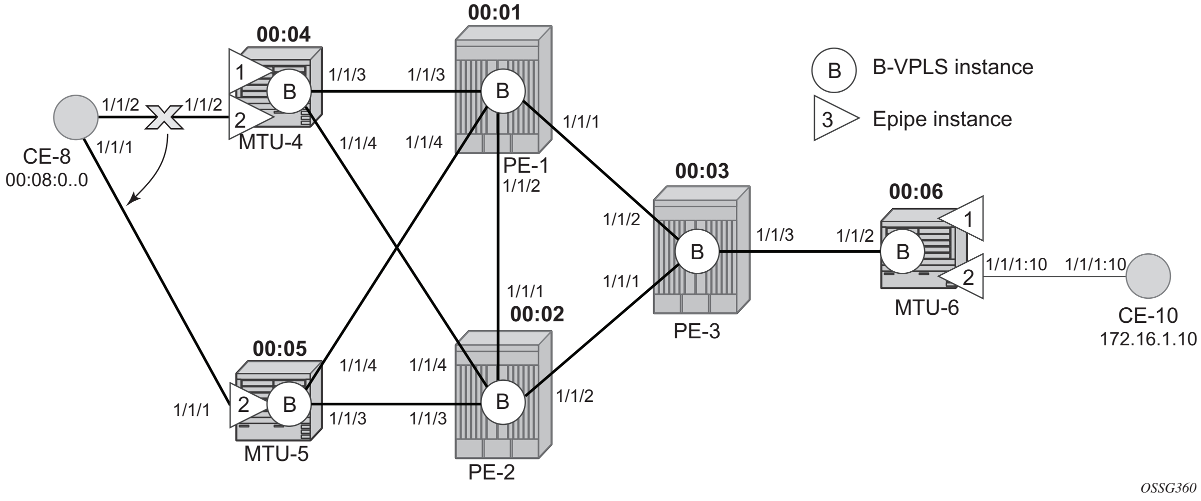
As soon as a group B-MAC attribute is registered on a particular port, an MFIB entry is added for that B-MAC on that port, regardless of the declaration state for that attribute on the port. For instance, neither B-MAC is declared on MTU-5 however the two MFIB entries are created as soon as the attributes are registered:

```
*A:MTU-5# show service id 100 mfib

===============================================================================
Multicast FIB, Service 100
===============================================================================
Source Address  Group Address         Sap/Sdp Id               Svc Id   Fwd/Blk
-------------------------------------------------------------------------------
*               01:1e:83:00:00:01     b-sdp:52:100             Local    Fwd
*              01:1e:83:00:00:02     b-sdp:52:100             Local     Fwd
-------------------------------------------------------------------------------
Number of entries: 2
===============================================================================
*A:MTU-5#
```

MAC Flush: Avoiding black-holes

Both the I-VPLS and B-VPLS components inherit the MAC flush capabilities of a regular VPLS clearing the related C-MAC and respectively B-MAC FIBs. All types of MAC flush—**flush-all-but-mine** and **flush-all-from-me**—are supported together with the related CLI. In addition to these features, some extensions have been added so that MAC flush can be triggered on the B-VPLS based on some events happening on the I-VPLS. The following diagram shows a potential scenario where black-holes can occur if the proper configuration is not added.

Figure 125 **Black-hole**



Under normal conditions the I-VPLS 2 FIB on MTU-6 shows that CE-8 MAC address is learned through B-MAC 00:04 (MTU-4’s B-MAC):

```
*A:MTU-6# show service id 2 fdb pbb

==============================================================================
Forwarding Database, i-Vpls Service 2
==============================================================================
MAC               Source-Identifier     B-Svc     b-Vpls MAC        Type/Age
------------------------------------------------------------------------------
00:08:00:00:00:00 b-sdp:63:100          100       00:04:04:04:04:04 L/60
00:10:00:00:00:00 sap:1/1/1:10          100       N/A               L/60
==============================================================================
*A:MTU-6#
```

When a failure happens in the CE-8 MC-LAG active link, the link to MTU-5 takes over. However, the FIB on MTU-6 still points at MTU-4's B-MAC and that will still be the B-MAC used in the PBB encapsulation. Therefore, a black-hole occurs until either bidirectional traffic is sent or the FIB aging timer expires.

The configuration in the I-VPLS can be modified to trigger a MAC flush in the B-VPLS with the following command:

```
*A:MTU-4# configure service vpls 2 pbb send-bvpls-flush
  - send-bvpls-flush {[all-but-mine] [all-from-me]}
  - no send-bvpls-flush

 <all-but-mine>       : keyword
 <all-from-me>        : keyword
```

The following command is executed on all MTUs to solve the black-hole:

```
configure service vpls 2 pbb send-bvpls-flush all-from-me
```

By enabling **send-bvpls-flush all-from-me** on I-VPLS 2, a failure on the MC-LAG active link on I-VPLS 2 will trigger an LDP MAC flush **flush-all-from-me** into the B-VPLS that will flush the FIB in MTU-6 for I-VPLS 2, avoiding the black-hole. A MC-LAG failure is emulated by disabling the LAG on MTU-4, as follows:

```
*A:MTU-4# configure lag 1 shutdown
```

MTU-4 sends the following LDP MAC flush for all MAC addresses learned from MTU-4:

```
1 2017/04/24 06:03:05.19 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Send Address Withdraw packet (msgId 263) to 192.0.2.1:0
Protocol version = 1
MAC Flush (All MACs learned from me)
Service FEC PWE3: ENET(5)/100 Group ID = 0 cBit = 0
Number of PBB-BMACs = 1
BMAC 1 = 00:04:04:04:04:04
Number of PBB-ISIDs = 1
ISID 1 = 2
Number of Path Vectors : 1
Path Vector(  1) = 192.0.2.4
"
```

On MTU-6:

```
1 2017/04/24 06:03:17.21 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Recv Address Withdraw packet (msgId 206) from 192.0.2.3:0
Protocol version = 1
MAC Flush (All MACs learned from me)
Service FEC PWE3: ENET(5)/100 Group ID = 0 cBit = 0
Number of PBB-BMACs = 1
```

```
BMAC 1 = 00:04:04:04:04:04
Number of PBB-ISIDs = 1
ISID 1 = 2
Number of Path Vectors : 3
Path Vector(  1) = 192.0.2.4
Path Vector(  2) = 192.0.2.1
Path Vector(  3) = 192.0.2.3
```

Immediately after receiving the MAC flush, the CE-8 MAC is flushed:

```
*A:MTU-6# show service id 2 fdb pbb

===============================================================================
Forwarding Database, i-Vpls Service 2
===============================================================================
MAC               Source-Identifier     B-Svc     b-Vpls MAC        Type/Age
-------------------------------------------------------------------------------
00:10:00:00:00:00 sap:1/1/1:10          100       N/A               L/30
===============================================================================
*A:MTU-6#
```

The CE-8 MAC is learned again, but this time linked to the B-MAC 00:05, which is the B-MAC of MTU-5:

```
*A:MTU-6# show service id 2 fdb pbb

===============================================================================
Forwarding Database, i-Vpls Service 2
===============================================================================
MAC               Source-Identifier     B-Svc     b-Vpls MAC        Type/Age
-------------------------------------------------------------------------------
00:08:00:00:00:00 b-sdp:63:100          100       00:05:05:05:05:05 L/0
00:10:00:00:00:00 sap:1/1/1:10          100       N/A               L/120
===============================================================================
*A:MTU-6#
```

The following I-VPLS events are propagated into the B-VPLS depending on the **flush-all-but-mine** or **flush-all-from-me** keywords used in the configuration:

If the **flush-all-but-mine** keyword is configured (positive flush), the following events in the I-VPLS trigger a MAC flush into the B-VPLS:

1. TCN event in one or more of the related I-VPLS/M-VPLS.
2. Pseudowire/SDP binding activation with active/standby pseudowire (standby to active or down to up).
3. Reception of an LDP MAC withdraw flush-all-but-mine in the related I-VPLS.

If the **flush-all-from-me** keyword is configured (negative flush) the following events in the I-VPLS trigger a MAC flush into the B-VPLS:

1. MC-LAG active link failure (in our example).

2. Failure of a local SAP – requires **send-flush-on-failure** to be enabled in I-VPLS.
3. Failure of a local pseudowire/SDP binding – requires **send-flush-on-failure** to be enabled in I-VPLS.
4. Reception of an LDP MAC withdraws flush-all-from-me in the related I-VPLS.

In addition to this and regardless of what type, MAC flush has been optimized to avoid flushing in the core PEs, flushing only the C-MACs mapped to a certain B-MAC (belonging to a specific ISID FIB) and the ability to indicate to core PEs which messages should always be forwarded endpoint-to-endpoint toward all PBB PEs regardless of the propagate-mac-flush setting in B-VPLS. All of this is implemented without the need of any additional CLI commands and it is part of *draft-balus-l2vpn-pbb-ldp-ext-00*.

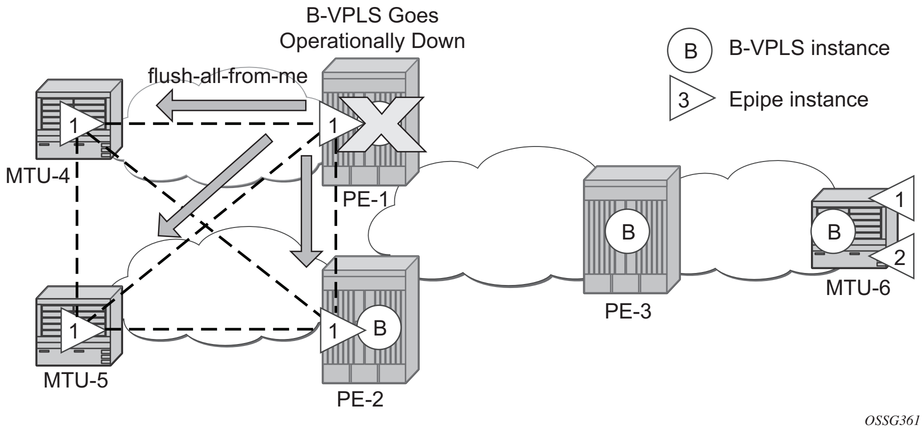
Another extension supported to avoid black-holes within this mix of I- and B-VPLS environments is the **block-on-mesh-failure** feature in PBB. When the VPLS mesh exists only in I-VPLS or in B-VPLS, and the **block-on-mesh-failure** feature is enabled, the regular VPLS behavior will apply (when all the mesh SDPs go down an LDP notification with pseudowire status bits = 0x01—Pseudo Wire Not Forwarding—is sent over the spoke SDPs). When the active/standby pseudowire resiliency is implemented in I-VPLS such that the PBB PE performs the role of a PE-rs, the B-VPLS core replaces the pseudowire (SDP binding) mesh. The block-on-mesh notification (LDP notification indicating pseudowire not forwarding) will be sent to the MTUs only when the related B-VPLS is operationally down. The B-VPLS core is operationally down only when all of its SAPs and SDPs are down.

The final feature that can be enabled in an I-VPLS with CLI is the **send-flush-on-bvpls-failure** feature.

```
*A:MTU-4# configure service vpls 2 pbb send-flush-on-bvpls-failure
  - no send-flush-on-bvpls-failure
  - send-flush-on-bvpls-failure
```

This feature is required to avoid black-holes when there is a full mesh of pseudowires in the I-VPLS domain and the B-VPLS instance can go operationally down. The following figure shows a typical scenario where this feature is needed (normally when PBB-VPLS and multi-chassis end point are combined together).

Figure 126 **Send Flush on BVPLS Failure Example**



Access Dual-Homing and MAC Notification

Although this section is focused on PBB in a MPLS based network, Nokia PBB implementation also allows the operator to use a native Ethernet infrastructure in the PBB core. Native Ethernet tunneling can be emulated using Ethernet SAPs to interconnect the related B-VPLS instances. In those cases, there is no LDP signaling available; therefore, there is no MAC flush sent when the active link in a multi-homed access device fails.

The SR OS supports a mechanism to avoid potential black-holes in native Ethernet PBB networks. In addition to the source B-MAC associated with each B-VPLS, an additional B-MAC is associated with each MC-LAG supporting Multi-homed I-VPLS SAPs. The nodes that are in a multi-homed MC-LAG configuration share a common B-MAC on the related MC-LAG interfaces. When the **mac-notification no shutdown** command is executed, an Ethernet CFM notification message is sent from the node holding the active link. That message will be flooded in the B-VPLS domain using the MC-LAG SAP B-MAC as the source MAC address. The remote nodes will learn the customer MAC addresses behind the MC-LAG and will link them to this new SAP B-MAC. MC-LAG will keep track of the active link for each particular LAG associated to a SAP B-MAC. Should MC-LAG detect any new active link in a node, a new CFM notification message will be flooded from the new active node.

The following caveats and considerations must be taken into account:

- Only MC-LAG is supported as dual-home mechanism.

- This mechanism is supported for native PBB and/or MPLS-based PBB-VPLS. Although it is mostly beneficial when native PBB is used in the core, it can also help to optimize the re-learning process in a MPLS-based core in case of MC-LAG failures, in addition to the existing LDP MAC flush procedures.

The example of this configuration shows the setup being used in this configuration example. MAC-notification will be configured in MTU-4 and MTU-5 for the dual-homed CE-8.

The first step is to configure the SAP B-MAC that will be used for the mac-notification messages. The **source-bmac-lsb** (source backbone MAC least significant bits) command has been added to the mc-lag branch so that the operator can decide the two last octets to be used in the SAP B-MAC. Those two last octets can be derived from the LACP key (if the **use-lacp-key** statement is used) or can be specifically defined.

```
*A:MTU-4>config>redundancy>mc>peer>mc-lag# lag
  - lag <lag-id> lacp-key <admin-key> system-id <system-id> [remote-lag <remote-lag-
    id>] system-priority <system-priority> source-bmac-lsb use-lacp-key
  - lag <lag-id> lacp-key <admin-key> system-id <system-id> [remote-lag <remote-lag-
    id>] system-priority <system-priority> source-bmac-lsb <MAC-Lsb>
  - lag <lag-id> lacp-key <admin-key> system-id <system-id> [remote-lag <remote-lag-
    id>] system-priority <system-priority>
  - lag <lag-id> [remote-lag <remote-lag-id>]
  - no lag <lag-id>

 <lag-id>             : [1..800]
 <admin-key>          : [1..65535]
 <system-id>          : xx:xx:xx:xx:xx:xx     - xx [00..FF]
 <remote-lag-id>      : [1..800]
 <system-priority>    : [1..65535]
 <MAC-Lsb>            : [1..65535] or xx-xx or xx:xx
```

There must be a different SAP B-MAC per MC-LAG. The use of the LACP key as a default for two least significant octets makes the operations simpler. In this example, the sap-bmac last two octets will come from the lacp-key. The configuration on MTU-4 is as follows:

```
configure
    redundancy
        multi-chassis
            peer 192.0.2.5 create
                mc-lag
                    lag 1 lacp-key 15 system-id 00:00:00:00:00:01
                            system- priority 65535 source-bmac-lsb use-lacp-key
                    no shutdown
                exit
                no shutdown
```

Therefore, the SAP B-MAC will be formed in the following way:

[sap-bmac = 4 first bytes of the source bmac + 2 bytes from source-bmac-lsb]

Enable the mac-notification in B-VPLS 100 on all MTUs as follows:

```
configure service vpls 100 mac-notification no shutdown
```

The **mac-notification** command activates the described mechanism and has the following parameters:

```
*A:MTU-4# configure service vpls 100 mac-notification
  - mac-notification

 [no] count           - Configure count for MAC-notification messages
 [no] interval        - Configure interval for MAC-notification messages
 [no] renotify        - Configure re-notify interval for MAC-notification messages
 [no] shutdown        - Configure admin state for MAC-notification messages
```

Where:

- interval <value> controls how often the subsequent MAC notification messages are sent. Default = 100 ms. Required values: 100 ms – 10 sec, in increments of 100 ms.
- count <value> controls how often the MAC notification messages are sent. Default: 3. Range: 1–10.

The “count” and “interval” parameters can also be configured at the service context. The settings configured at the B-VPLS service context take precedence though.

```
*A:MTU-4# configure service mac-notification
  - mac-notification

 [no] count           - Configure count for MAC-notification messages
 [no] interval        - Configure interval for MAC-notification messages
```

Finally, the B-VPLS is instructed to use the SAP B-MAC. The **use-sap-bmac** statement enables the use of the source B-MAC allocated to the multi-homed SAPs (assigned to the MC-LAG) in the related I-VPLS service (could be Epipe service as well). The command will fail if the value of the source B-MAC assigned to the B-VPLS is the hardware (chassis) B-MAC. In other words, the source B-MAC must be a configured one. The **use-sap-bmac** statement is by default off.

```
*A:MTU-4# configure
    service
        vpls 100
            pbb
                source-bmac 00:aa:aa:aa:aa:04
                use-sap-bmac
            exit


*A:MTU-5# configure
    service
        vpls 100
            pbb
```

```
                source-bmac 00:aa:aa:aa:aa:05
                use-sap-bmac
            exit

*A:MTU-6# show service id 2 fdb pbb

===============================================================================
Forwarding Database, i-Vpls Service 2
===============================================================================
MAC               Source-Identifier     B-Svc     b-Vpls MAC        Type/Age
-------------------------------------------------------------------------------
00:08:00:00:00:00 b-sdp:63:100          100       00:aa:aa:aa:00:0f L/0
00:10:00:00:00:00 sap:1/1/1:10          100       N/A               L/0
===============================================================================
*A:MTU-6#
```

As soon as the **mac-notification no shutdown** command is executed, an Ethernet CFM notification message is sent from MTU-4, which is the node where the active MC-LAG link resides. The CFM message will have the source mac “00:aa:aa:aa:00:0f” (4 first bytes of the configured source bmac + 2 bytes from the configured source-bmac-lsb, which is 15 in hex) and will be flooded throughout the B-VPLS domain. Should the link between CE-8 and MTU-4 fail, the MC-LAG protocol will activate the redundant link and MTU-5 will immediately issue a CFM message with the shared sourced SAP B-MAC that will be flooded in the B-VPLS domain.

PBB and IGMP Snooping

IGMP snooping can be enabled on I-VPLS SAPs and SDPs (it cannot be enabled on B-VPLS). SR OS can keep track of IGMP joins received over individual B-SDPs or B-SAPs, and it starts flooding the multicast group (and only the multicast group) to all B-components (using the group B-MAC for I-SID) as soon as the first IGMP join for that multicast group is received in one of the B-SAP/SDP components.

The first IGMP join message received over the local B-VPLS will add all the B-VPLS SAP/SDP components into the related multicast table associated with the I-VPLS context. When the querier is connected to a remote I-VPLS instance, over the B-VPLS infrastructure, its location is identified by the B-VPLS SDP/SAP on which the query was received and also by the source B-MAC address used in the PBB header for the query message, the B-MAC associated with the B-VPLS instance on the remote PBB PE.

The following configuration on MTU-4 enables IGMP snooping in I-VPLS 1 and adds some static groups on a SAP. The location of the querier is configured by adding the B-MAC where the querier is connected to (in this example, MTU-6) and adding the two B-VPLS spoke SDPs as mrouter ports (B-VPLS mrouter ports are added in the I-VPLS backbone-vpls context).

The **mac-name** command translates MAC address into strings so that the names can be used instead of typing the entire MAC address every time we need to.

```
configure
    service
       pbb
            source-bmac 00:04:04:04:04:04
            mac-name "MTU-4" 00:04:04:04:04:04
            mac-name "MTU-5" 00:05:05:05:05:05
            mac-name "MTU-6" 00:06:06:06:06:06
        exit
        vpls 1 customer 1 i-vpls create
            pbb
                backbone-vpls 100
                    igmp-snooping
                        mrouter-dest "MTU-6"
                    exit
                    sdp 41:100
                        igmp-snooping
                            mrouter-port
                        exit
                    exit
                    sdp 42:100
                        igmp-snooping
                            mrouter-port
                        exit
                    exit
                exit
            exit
            igmp-snooping
                no shutdown
            exit
            sap 1/1/1:7 create
                igmp-snooping
                    static
                        group 228.0.0.1
                            starg
                        exit
                        group 228.0.0.2
                            starg
                        exit
                        group 239.0.0.1
                            source 172.16.99.99
                        exit
                    exit
                exit
            exit
            no shutdown
```

As in regular VPLS instances, mrouter ports are added to all the multicast groups:

```
*A:MTU-4# show service id 1 mfib

===============================================================================
Multicast FIB, Service 1
===============================================================================
Source Address  Group Address         Sap/Sdp Id                    Svc Id   Fwd
                                                                             Blk
```

```
--------------------------------------------------------------------------------
*               *                     b-sdp:41:100                 100      Fwd
                                      b-sdp:42:100                 100      Fwd
*               228.0.0.1             sap:1/1/1:7                  Local    Fwd
                                      b-sdp:41:100                 100      Fwd
                                      b-sdp:42:100                 100      Fwd
*               228.0.0.2             sap:1/1/1:7                  Local    Fwd
                                      b-sdp:41:100                 100      Fwd
                                      b-sdp:42:100                 100      Fwd
172.16.99.99    239.0.0.1             sap:1/1/1:7                  Local    Fwd
                                      b-sdp:41:100                 100      Fwd
                                      b-sdp:42:100                 100      Fwd
--------------------------------------------------------------------------------
Number of entries: 4
================================================================================
*A:MTU-4#
```

When the **show service id x mfib** command is issued in an I-VPLS as in the preceding output, the IGMP (S,G) and (*,G) entries for the I and B components are shown if IGMP snooping is enabled. However, when the same command is launched in a B-VPLS as in the following output, the group B-MAC entries are shown.

```
*A:MTU-4# show service id 100 mfib

================================================================================
Multicast FIB, Service 100
================================================================================
Source Address  Group Address         Sap/Sdp Id                   Svc Id   Fwd
                                                                            Blk
--------------------------------------------------------------------------------
*               01:1e:83:00:00:01     b-sdp:41:100                 Local    Fwd
*               01:1e:83:00:00:02     b-sdp:41:100                 Local    Fwd
--------------------------------------------------------------------------------
Number of entries: 2
================================================================================
*A:MTU-4#
```

MMRP Policies and ISID-Based Filtering for PBB Inter-Domain Expansion

As described in the MMRP for Flooding Optimization section, MMRP is used in the backbone VPLS instances to build per I-VPLS flooding trees. Each I-VPLS has an associated group B-MAC in the B-VPLS, which is derived from the ISID, and is advertised by MMRP throughout the whole B-VPLS context, regardless of whether a certain I-VPLS is present in one or all the B-VPLS PEs.

In an inter-domain environment, the same B-VPLS can be defined in different domains and as such MMRP will advertise all the group B-MACs in every domain. The group B-MACs are consuming resources in all the PEs no matter if a particular ISID—and therefore its group B-MAC—is required in one of the domains or not. When MMRP is enabled in a particular PE, data plane and control plane resources are consumed and they must be taken into consideration when designing PBB-VPLS networks:

- Control plane – MRRP processing takes CPU cycles and the number of attributes that can be advertised is not unlimited
- Data plane – each group B-MAC registration takes one MFIB entry (the MFIB is shared between MMRP and IGMP/PIM snooping)

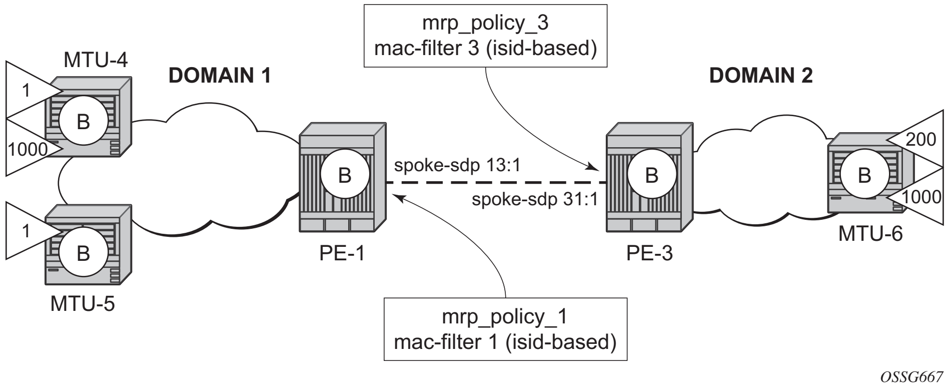
SR OS routers support MMRP policies and ISID-based filters so that control plane and data plane resources can be saved when I-VPLS instances are not defined in all the domains.

Figure 127 illustrates an example of usage for MMRP policies and ISID-based filters that will be configured in this section. “Domain 1” and “domain 2” will have a range of local ISIDs each and a range of “inter-domain” ISIDs:

- Domain 1 local ISIDs: from 1 to 100
- Domain 2 local ISIDs: from 101 to 200
- Inter-domain ISIDs: from 1000 to 2000

By applying the MMRP policies indicated in Figure 127, domain 1 attributes will be prevented from being declared and registered in domain 2 and vice versa, domain 2 attributes from being declared and registered in domain 1. The egress mac-filters will drop any traffic sourced from a local ISID preventing it to be transmitted to the remote domain.

Figure 127 **Inter-Domain B-VPLS and MMRP Policies/ISID-Based Filters Example**



MMRP Policies

The following shows the MMRP policy configuration on node PE-1. This policy will block any registration/declaration except those for ISIDs 1000-2000. Packets will be compared against the configured matching ISIDs as long as the pbb-etype matches the one configured on the port or SDP.

On PE-1:

```
configure
    service
        mrp
            mrp-policy "mrp_policy_1" create
                description "allow-inter-domain-isids"
                default-action block
                entry 10 create
                    action allow
                    match
                        isid 1000 to 2000
                    exit
                exit
            exit
        exit
    exit
```

Once the MMRP policy is configured, it must be applied on the corresponding SAP or sdp-binding. An mrp-policy can be applied to a B-VPLS SAP, B-VPLS spoke-sdp or B-VPLS mesh-sdp:

On PE-1:

```
configure
    service
        vpls 100
            spoke-sdp 14:100 create
                mrp
                    mrp-policy "mrp_policy_1"
                exit
            exit
            spoke-sdp 15:100 create
                mrp
                    mrp-policy "mrp_policy_1"
                exit
            exit
            no shutdown
        exit
```

In the same way, mrp_policy_3 will be configured in PE-3.

Some additional considerations about the MMRP policies:

- Different entries within the same mrp-policy can have overlapping ISID ranges. The entries will be evaluated in the order of their IDs and the first match will cause the implementation to execute the associated action for that entry and then to exit the mrp-policy.
- If no ISID is specified in the match condition then:
  - If the action is “end-station”, no entry is added and the action is block.
  - If the action is different from “end-station”, every ISID is considered for that action.
- The mrp-policy specifies either a forward or a drop action for the group B-MAC attributes associated with the ISIDs specified in the match criteria.

```
*A:PE-1>config>serv>mrp>mrp-policy>entry# action
 - action <action>
 - no action

 <action>             : none|block|allow|end-station
```

- There is an additional action called **end-station**. This action specifies that an end-station emulation is present on the SAP/SDP-binding where the policy has been applied. The matching ISIDs will not get declared/registered in the SAP/SDP-binding (just like the **block** action). However, those attributes will get mapped as static MMRP entries on the SAP/SDP-binding, which implicitly get instantiated in the data plane as MFIB entries associated with that SAP/SDP-binding for the related group B-MAC. When the action is “end-station”, the default-action must be block:

```
*A:PE-3>config>serv>mrp>mrp-policy# default-action allow
MINOR: SVCMGR #5904 Mrp-policy default-action must be block when end-station action
exists
```

- The **end-station** action can be used in the inter-domain gateways when, for instance, we do not want MMRP control plane exchanges between domains. The following output shows how to define the static MMRP entries 1000-2000 in PE-3 without receiving any declaration for any of those attributes or having any of those locally configured.

  On PE-3:

```
configure
    service
        mrp
            mrp-policy "mrp_policy_3" create
                default-action block
                entry 10 create
                    action end-station
                    match
                        isid 1000 to 2000
                    exit
                exit
            exit
        exit
    exit

*A:PE-3# show service id 100 mfib

===============================================================================
Multicast FIB, Service 100
===============================================================================
Source Address  Group Address         Port Id                      Svc Id   Fwd
                                                                            Blk
-------------------------------------------------------------------------------
*               01:1e:83:00:00:01     b-sdp:36:100                 Local    Fwd
*               01:1e:83:00:03:e8     b-sdp:36:100                 Local    Fwd
*               01:1e:83:00:03:e9     b-sdp:31:4294967294          Local    Fwd
                                      b-sdp:36:100                 Local    Fwd
---snip---
*               01:1e:83:00:07:ce     b-sdp:36:100                 Local    Fwd
*               01:1e:83:00:07:cf     b-sdp:36:100                 Local    Fwd
*               01:1e:83:00:07:d0     b-sdp:36:100                 Local    Fwd
-------------------------------------------------------------------------------
Number of entries: 1002
===============================================================================
*A:PE-3#
```

- The mrp-policy can be applied to multiple B-VPLS services as long as the scope of the policy is **template** (the scope can also be **exclusive**).
- Any changes made to the existing policy will be applied immediately to all services where this policy is applied. For this reason, when many changes are required on a mrp-policy, it is recommended that the policy be copied to a work-in-progress policy. That work-in-progress policy can be modified until complete and then written over the original mrp-policy. You can use the **config mrp-policy copy** command to work with the policies in this manner. The **renum** command can also help to change the entries sequence order.

```
*A:PE-3# configure service mrp copy
  - copy <src-mrp-policy> to <dst-mrp-policy>

 <src-mrp-policy>      : [32 chars max]
 <dst-mrp-policy>      : [32 chars max]



*A:PE-3# configure service mrp mrp-policy "mrp_policy_3" renum
  - renum <src-entry-id> to <dst-entry-id>

 <src-entry-id>        : [1..65535]
 <dst-entry-id>        : [1..65535]
```

- The **no** form of the **mrp-policy** command deletes the mrp-policy. An mrp policy cannot be deleted until it is removed from all the SAPs/SDP-bindings where it is applied.

ISID-Based Filters

The MMRP policies help to control the exchange of group B-MAC attributes across domains. Based on the registration state of a specific group B-MAC on a SAP/SDP-binding, the broadcast/ unknown-unicast/multicast traffic for a particular I-VPLS will be allowed or dropped. However, to avoid that any local ISID packet is flooded to the remote B-VPLS domain, all the packets tagged with the local ISIDs at the gateway PEs need to be filtered at the data plane. ISID- based filters will prevent the local ISIDs from sending any packet with unicast B-MAC to the remote domain. This is particularly useful for PBB-Epipe services across domains, where all the frames use unicast B-MACs and MMRP policies cannot help because they only act on group B-MAC packets.

The following CLI output shows how to configure an ISID-based filter that drops all the traffic sourced from the local ISIDs on PE-1 (the default action is drop and it does not show up in the configuration).

```
*A:PE-1# configure
    filter
        mac-filter 1 create
            description "drop_local_isids"
            type isid
            entry 10 create
                match frame-type 802dot3
                    isid 1000 to 2000
                exit
                log 101
                action
                    forward
                exit
            exit
```

Once the filter is configured, it must be applied on a B-VPLS SAP or SDP-binding and always at egress.

```
*A:PE-1# configure
    service
        vpls 100
            spoke-sdp 14:100 create
                egress
                    filter mac 1
                exit
            exit
            no shutdown
            spoke-sdp 15:100 create
                egress
                    filter mac 1
                exit
            exit
```

Some additional comments about ISID-based filters:

- The **type isid** statement must be added before introducing any ISID in the match command, otherwise the system will show an error:

```
*A:PE-1>config>filter>mac-filter>entry>match$ isid 1000 to 2000
MINOR: FILTER #1533 The match criteria entered are not compatible with the Mac
filter type - On a normal filter no ISID or VID match criteria are allowed


*A:PE-1>config>filter>mac-filter$ type isid
MINOR: FILTER #1561 Cannot change filter type when filter contains entries
```

- Once the operator sets the “type isid”, the filter cannot be applied at ingress. Only egress ISID-based filters are allowed:

```
*A:PE-1>config>service>vpls>mesh-sdp# ingress filter mac 1
MINOR: SVCMGR #2050 Can not apply filter of type 'isid' on ingress
```

- Like any filter or MMRP policy, the filter can be applied to multiple B-VPLS services as long as the scope of the policy is “template” (the scope can also be “exclusive”).
- The following command shows the filter configuration and packets that have matched the filter (field “Egr. Matches”):

```
*A:PE-1# show filter mac 1

===============================================================================
Mac Filter
===============================================================================
Filter Id           : 1                            Applied        : Yes
Scope               : Template                     Def. Action    : Drop
Entries             : 1                            Type           : isid
Description         : drop_local_isids
-------------------------------------------------------------------------------
Filter Match Criteria : Mac
-------------------------------------------------------------------------------
Entry               : 10                           FrameType      : Ethernet
Description         : (Not Specified)
Log Id              : 101
ISID                : 1000..2000
Primary Action      : Forward
Ing. Matches        : 0 pkts
Egr. Matches        : 5 pkts (580 bytes)

===============================================================================
*A:PE-1#
```

- Like any other filter, the matching packets can be logged. An example follows (the Ethertype is 0x88e7, which is the default standard Ethertype for PBB):

```
*A:PE-1# show filter log 101

===============================================================================
Filter Log
===============================================================================
Admin state : Enabled
```

```
Description : Default filter log
Destination : Memory
Wrap        : Enabled
-------------------------------------------------------------------------------
Maximum entries configured : 1000
Number of entries logged   : 5
-------------------------------------------------------------------------------
2017/04/24 08:40:41  Mac Filter: 1:10  Desc:
Interface: int-PE-1-MTU-4  Direction: Egress  Action: Forward
VID match: 0
Src MAC: 00-06-06-06-06-06  Dst MAC: 00-aa-aa-aa-00-0f  EtherType: 88e7
Hex: 00 00 03 e9 00 08 00 00 00 00 00 10 00 00 00 00
     08 00 45 00 00 54 00 24 00 00 40 01 4a 61 ac 10
     6c 02 ac 10 6c 01 00 00 fe 88 c0 09 00 01 57 44*

2017/04/24 08:40:42  Mac Filter: 1:10  Desc:
Interface: int-PE-1-MTU-4  Direction: Egress  Action: Forward
VID match: 0
Src MAC: 00-06-06-06-06-06  Dst MAC: 00-aa-aa-aa-00-0f  EtherType: 88e7
Hex: 00 00 03 e9 00 08 00 00 00 00 00 10 00 00 00 00
     08 00 45 00 00 54 00 25 00 00 40 01 4a 60 ac 10
     6c 02 ac 10 6c 01 00 00 fd 9d c0 09 00 02 57 44*

---snip---
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-1#
```

B-VPLS and I-VPLS Show and Debug Commands

For the following output, the MRP policies and ISID-based MAC filters have been removed from the spoke SDPs on PE-1 and PE-3. The following commands can help to check the B-VPLS and I-VPLS configuration and their related parameters. The first is for the B-VPLS on MTU-4:

```
*A:MTU-4# show service id 100 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 100                 Vpn Id            : 0
Service Type      : b-VPLS
Name              : (Not Specified)
Description       : (Not Specified)
Customer Id       : 1                   Creation Origin   : manual
Last Status Change: 24/04/2017 05:27:03
Last Mgmt Change  : 24/04/2017 06:49:29
Etree Mode        : Disabled
Admin State       : Up                  Oper State        : Up
MTU               : 2000                Def. Mesh VC Id   : 100
SAP Count         : 0                   SDP Bind Count    : 2
Snd Flush on Fail : Disabled            Host Conn Verify  : Disabled
SHCV pol IPv4     : None
Propagate MacFlush: Disabled            Per Svc Hashing   : Disabled
```

```
Allow IP Intf Bind: Disabled
Fwd-IPv4-Mcast-To*: Disabled            Fwd-IPv6-Mcast-To*: Disabled
Mcast IPv6 scope  : mac-based
Temp Flood Time   : Disabled            Temp Flood        : Inactive
Temp Flood Chg Cnt: 0
SPI load-balance  : Disabled
TEID load-balance : Disabled
Src Tep IP        : N/A
VSD Domain        : <none>
Oper Backbone Src : 00:aa:aa:aa:aa:04
Use SAP B-MAC     : Enabled
i-Vpls Count      : 2
Epipe Count       : 0
Use ESI B-MAC     : Disabled

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sdp:41:100 S(192.0.2.1)                  Spok         8000    8000    Up   Up
sdp:42:100 S(192.0.2.2)                  Spok         8000    8000    Up   Up
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:MTU-4#
```

For the I-VPLS on MTU-4:

```
*A:MTU-4# show service id 1 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 1                   Vpn Id            : 0
Service Type      : i-VPLS
Name              : (Not Specified)
Description       : (Not Specified)
Customer Id       : 1                   Creation Origin   : manual
Last Status Change: 24/04/2017 05:38:08
Last Mgmt Change  : 24/04/2017 07:28:35
Etree Mode        : Disabled
Admin State       : Up                  Oper State        : Up
MTU               : 1514                Def. Mesh VC Id   : 1
SAP Count         : 1                   SDP Bind Count    : 0
Snd Flush on Fail : Disabled            Host Conn Verify  : Disabled
SHCV pol IPv4     : None
Propagate MacFlush: Disabled            Per Svc Hashing   : Disabled
Allow IP Intf Bind: Disabled
Fwd-IPv4-Mcast-To*: Disabled            Fwd-IPv6-Mcast-To*: Disabled
Mcast IPv6 scope  : mac-based
Temp Flood Time   : Disabled            Temp Flood        : Inactive
Temp Flood Chg Cnt: 0
SPI load-balance  : Disabled
TEID load-balance : Disabled
Src Tep IP        : N/A
VSD Domain        : <none>
b-Vpls Id         : 100                 Oper ISID         : 1
b-Vpls Status     : Up
```

```
Snd Flush in bVpls: None
Flsh On bVpls Fail: Disabled            Prop Flsh fr bVpls: Disabled
Force QTag Fwd    : Disabled
SendBvplsEvpnFlush: Disabled

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/1:7                              q-tag        1518    1518    Up   Up
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:MTU-4#
```

The following command shows all the I-VPLS instances multiplexed into a particular B-VPLS.

```
*A:MTU-4# show service id 100 i-vpls

===============================================================================
Related i-Vpls services for b-Vpls service 100
===============================================================================
i-Vpls SvcId        Oper ISID           Admin               Oper
-------------------------------------------------------------------------------
1                   1                   Up                  Up
2                   2                   Up                  Up
-------------------------------------------------------------------------------
Number of Entries : 2
-------------------------------------------------------------------------------
===============================================================================
*A:MTU-4#
```

Some useful commands to check the I and B VPLS FIBs correlating C-MACs and B-MACs:

```
*A:MTU-4# show service id 1 fdb pbb

==============================================================================
Forwarding Database, i-Vpls Service 1
==============================================================================
MAC               Source-Identifier     B-Svc     b-Vpls MAC        Type/Age
------------------------------------------------------------------------------
00:07:00:00:00:00 sap:1/1/1:7           100       N/A               L/0
00:09:00:00:00:00 b-sdp:41:100          100       00:06:06:06:06:06 L/0
==============================================================================
*A:MTU-4#

*A:MTU-4# show service id 100 fdb pbb

=======================================================================
Forwarding Database, b-Vpls Service 100
=======================================================================
MAC               Source-Identifier     iVplsMACs  Epipes     Type/Age
-----------------------------------------------------------------------
00:06:06:06:06:06 sdp:41:100            2          0          L/0
```

```
16:09:ff:00:00:00 sdp:41:100            0          0          L/0
=======================================================================
*A:MTU-4#
```

If **mac-names** are used in the configuration, the following commands can show the translations:

```
*A:MTU-4# show service pbb mac-name

======================================================================
MAC Name Table
======================================================================
MAC-Name                           MAC-Address
----------------------------------------------------------------------
MTU-4                              00:04:04:04:04:04
MTU-5                              00:05:05:05:05:05
MTU-6                              00:06:06:06:06:06
======================================================================
*A:MTU-4#
```

```
*A:MTU-4# show service pbb mac-name "MTU-6" detail

======================================================================
Services Using MAC name='MTU-6' addr='00:06:06:06:06:06'
======================================================================
Svc-Id                             ISID
----------------------------------------------------------------------
1                                  N/A
----------------------------------------------------------------------
Number of services: 1
======================================================================
*A:MTU-4#
```

The following command shows the base MAC notification parameters as well as the source B-MAC configured at the service PBB level. Those values are overridden by any potential MAC notification or source B-MAC values configured under the B-VPLS service context.

```
*A:MTU-4# show service pbb base

======================================================================
PBB MAC Information
======================================================================
MAC-Notif Count                   : 3
MAC-Notif Interval                : 1
Source BMAC                       : 00:04:04:04:04:04
======================================================================
*A:MTU-4#
```

If MAC notification is used in a particular B-VPLS, the configured least significant bits for the SAP B-MAC on a particular MC-LAG can be shown by using the detailed view of the **show lag** command:

```
*A:MTU-4# show lag 1 detail
```

```
===============================================================================
LAG Details
===============================================================================
Description        : N/A
-------------------------------------------------------------------------------
Details
-------------------------------------------------------------------------------
Lag-id              : 1                      Mode                 : access
Adm                 : up                     Opr                  : up

---snip---

MC Peer Address     : 192.0.2.5              MC Peer Lag-id       : 1
MC System Id        : 00:00:00:00:00:01      MC System Priority   : 65535
MC Admin Key        : 15                     MC Active/Standby    : active
MC Lacp ID in use   : true                   MC extended timeout  : false
MC Selection Logic  : local master decided
MC Config Mismatch  : no mismatch
Source BMAC LSB     : use-lacp-key           Oper Src BMAC LSB    : 00:0f

---snip---
===============================================================================
*A:MTU-4#
```

The following debug commands allow the operator to check the LDP label mapping, label withdrawal, messages and also the MAC-flush messages for regular VPLS, for I-VPLS and B-VPLS including the PBB extensions and TLVs.

```
*A:MTU-4# show debug
debug
    router "Base"
        ldp
            peer 192.0.2.1
                event
                exit
                packet
                    init detail
                    label detail
                exit
            exit
            peer 192.0.2.2
                event
                exit
                packet
                    init detail
                    label detail
                exit
            exit
        exit
    exit
exit
```

The following debug commands can help the operator to troubleshoot MMRP.

```
*A:MTU-4# debug service id 100 mrp
```

```
  - mrp
  - no mrp

      all-events      - Enable/disable MRP debugging for all events
 [no] applicant-sm    - Enable/disable MRP debugging for applicant state machine
                        changes
 [no] leave-all-sm    - Enable/disable MRP debugging for leave all state machine
                        changes
 [no] mmrp-mac        - Enable/disable MRP debugging for a particular MAC address
 [no] mrpdu           - Enable/disable MRP debugging for Rx/Tx MRP PDUs
 [no] mvrp-vlan       - Enable/disable debugging for a particular vlan
 [no] periodic-sm     - Enable/disable MRP debugging for periodic state machine
                        changes
 [no] registrant-sm   - Enable/disable MRP debugging for registrant state machine
                        changes
 [no] sap             - Enable/disable MRP debugging for a particular SAP
 [no] sdp             - Enable/disable MRP debugging for a particular SDP

*A:MTU-4#
```

Conclusion

PBB-VPLS allows the service providers to scale VPLS services by multiplexing customer I-VPLS instances into one or more B-VPLS instances. This multiplexing dramatically reduces the number of services, pseudowires and MAC addresses in the core and therefore allows the service provider to scale Layer 2 multi-point networks and provide services across international backbones.

The example used in this section shows the configuration of the customer and backbone VPLS instances as well as all the related features which are required for this environment. Show and debug commands have also been suggested so that the operator can verify and troubleshoot the service.

Preference-based and Non-revertive EVPN DF Election

This chapter provides information about Preference-based and Non-revertive EVPN DF Election.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

The information and configuration in this chapter are based on SR OS Release 15.0.R3. Preference-based and non-revertive EPVN Designated Forwarder (DF) election is supported in SR OS Release 15.0.R1, and later. This mechanism works for Ethernet Segments (ESs) and virtual ESs (vESs).

Overview

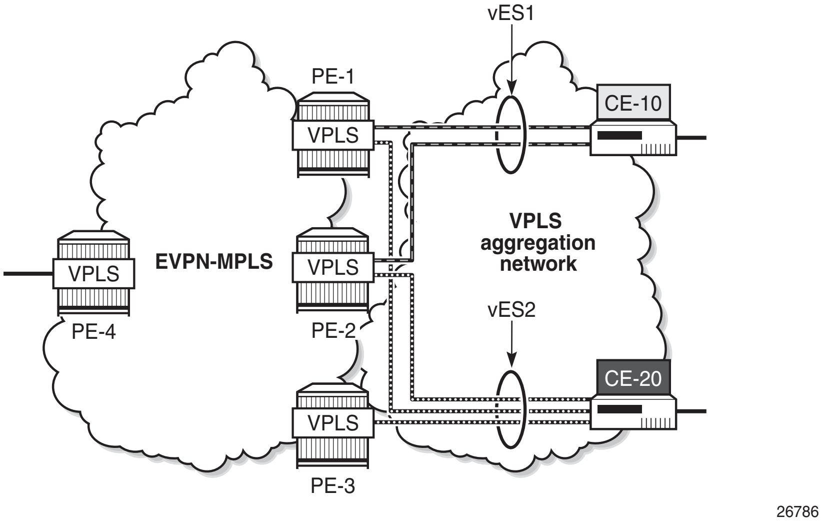
RFC7432 defines the Designated Forwarder (DF) in (PBB-)EVPN networks as the PE that will forward the following packets to a multi-homed node:

- Broadcast, Unknown unicast, and Multicast (BUM) traffic in an all-active multi-homing Ethernet Segment (ES)
- BUM and unicast in a single-active multi-homing ES

For more information about vESs, see chapter Virtual Ethernet Segments.

Figure 128 shows a topology with two vESs.

Figure 128 **Virtual Ethernet Segments**

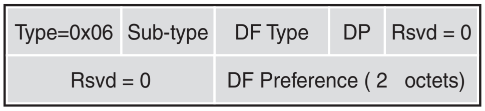


Taking the Ethernet VPN Identifier (EVI) or ISID and the number of PEs in the ES as input, the RFC7432 service-carving algorithm elects the DF from the list of candidate PEs that advertise the ES identifier (ESI). While this algorithm provides an automated and fair DF distribution across services in the ES, it does not allow the operator to control what PE is the DF for which service. In addition, in case of a DF failure, when the former DF comes back up, a new DF switchover will cause unnecessary packet loss (this mode of operation is called revertive). SR OS implements *draft-ietf-bess-evpn-pref-df* to give more control to the operator on the DF Election and avoid the revertive mode.

In SR OS, in addition to the automated service-carving, the DF election can also be controlled by configuring a preference manually. Also, it is possible to force an on-demand DF switchover without reconfiguring the PEs in the ES. Furthermore, the non-revertive option prevents an automatic switchover when a new active PE can preempt the existing DF PE. The non-revertive option avoids service impact when an ES comes back up.

Figure 129 shows the BGP-EVPN extended community defined for DF election and the different values described in draft-ietf-bess-evpn-pref-df.

Figure 129 **BGP-EVPN Extended Community for DF Election**



DP = Do not preempt (non-revertive)
DF = Designated forwarder
- Type 0 – Default, modulo-based DF election (RFC7432)
- Type 1 – Highest Random Weight (HRW) algorithm
- Type 2 – Preference algorithm

26787

The "Do not preempt" (DP) bit is set to enable the non-revertive option. When preference-based service carving is configured in the ES, DF type 2 is advertised along with a 2-byte preference value, which is 32767 by default.

Service carving can be configured in auto mode or manual mode. The preference can only be configured in manual mode.

```
*A:MTU-1>config>service>system>bgp-evpn>eth-seg>service-carving# mode
  - mode {manual|auto}

 <manual|auto>        : auto|manual|off
```

When manual mode is enabled, the following parameters can be configured to control which PE will be elected as DF:

```
*A:PE-2>config>service>system>bgp-evpn>eth-seg>service-carving# manual
  - manual

 [no] evi             - Configure EVI range (primary for non-
preference based DF election and lowest-preference for preference based DF election)
 [no] isid            - Configure ISID range (primary for non-
preference based DF election and lowest-preference for preference based DF election)
 [no] preference      + Configure DF preference election information
```

The EVI and ISID ranges configured in the service-carving context do not need to be consistent with any ranges configured for virtual ESs.

When preference is configured manually, a preference value can be provided along with the non-revertive option:

```
*A:PE-2>config>service>system>bgp-evpn>eth-seg>service-carving>manual# preference
  - no preference
  - preference [create] [non-revertive]

 <create>             : keyword
 <non-revertive>      : keyword

      value           - Configure DF preference value
```

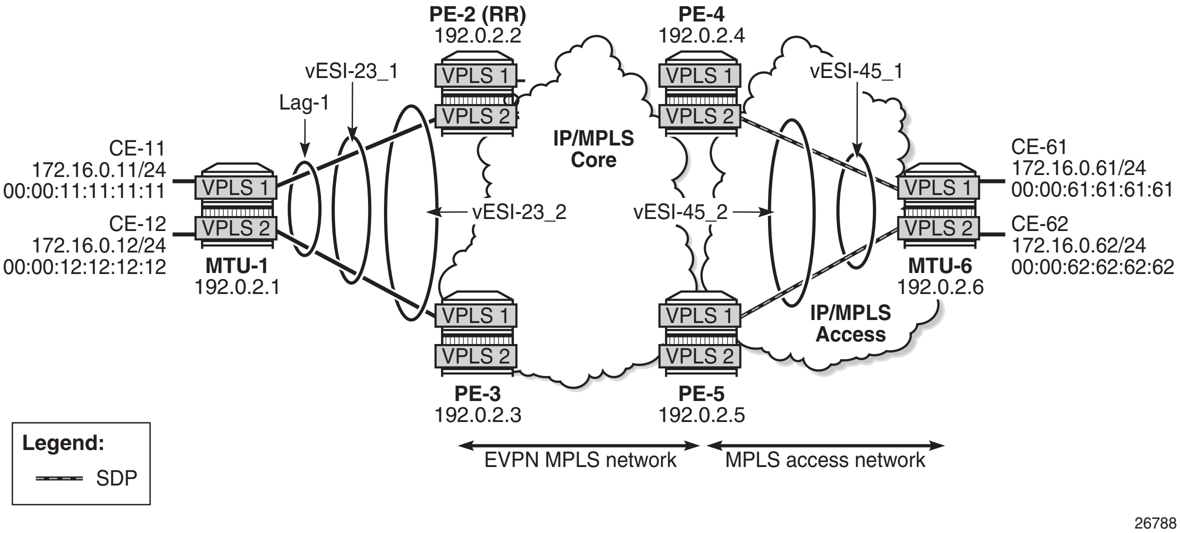
The preference-based EVPN DF election is as follows:

- By default, all SAPs and spoke-SDPs on the configured ES select the highest-preference PE as DF; however, when the EVI or ISID ranges are configured in the ES, the lowest-preference PE is selected.
- When the preference is equal, the DP bit is the tiebreaker: DP=1 wins over DP=0.
- For equal preference and DP, the PE IP address is the tiebreaker: the lowest IP address wins.

Configuration

Figure 130 shows the example topology with six nodes. EVPN-MPLS is configured between the core PE nodes. All-active vESs are configured between PE-2 and PE-3 and single-active vESs are configured between PE-4 and PE-5.

Figure 130 **Example Topology with All-active and Single-active vESs**



The initial configuration includes:

- Cards, MDAs, ports
- LAG 1 between MTU-1, PE-2, PE-3
- Router interfaces
- IS-IS (alternatively, OSPF could be used)

- LDP

BGP is configured on the four core PEs with PE-2 as Route Reflector (RR). The BGP configuration on RR PE-2 is as follows:

```
configure
    router
        autonomous-system 64500
        bgp
            vpn-apply-import
            vpn-apply-export
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            split-horizon
            rapid-update evpn
            group "internal"
                family evpn
                cluster 1.1.1.1
                peer-as 64500
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
                neighbor 192.0.2.5
                exit
            exit
```

VPLS 1 and VPLS 2 are configured on each node. The PEs have EVPN-MPLS enabled. The configuration on PE-2 is as follows:

```
configure
    service
        vpls 1 customer 1 create
            bgp
            exit
            bgp-evpn
                evi 1
                mpls
                    ingress-replication-bum-label
                    ecmp 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            sap lag-1:1.1 create
            exit
            no shutdown
        exit
        vpls 2 customer 1 create
            bgp
            exit
            bgp-evpn
                evi 2
                mpls
```

```
                    ingress-replication-bum-label
                    ecmp 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            sap lag-1:2.1 create
            exit
            no shutdown
        exit
```

The configuration on the other PEs is similar; PE-4 and PE-5 have a spoke-SDP configured instead of a SAP. For an explanation of the configuration, see chapter EVPN for MPLS Tunnels.

Service Carving: Auto Mode

On PE-2 and PE-3, the following all-active multi-homing vESs are configured:

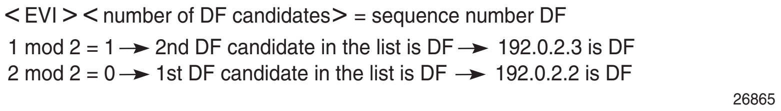
```
configure
    service
        system
            bgp-evpn
                ethernet-segment "vESI-23_1" virtual create
                    esi 01:00:00:00:00:23:01:00:00:01
                    es-activation-timer 3
                    service-carving
                        mode auto
                    exit
                    multi-homing all-active
                    lag 1
                    qinq
                        s-tag-range 1
                    exit
                    no shutdown
                exit
                ethernet-segment "vESI-23_2" virtual create
                    esi 01:00:00:00:00:23:02:00:00:01
                    es-activation-timer 3
                    service-carving
                        mode auto
                    exit
                    multi-homing all-active
                    lag 1
                    qinq
                        s-tag-range 2
                    exit
                    no shutdown
                exit
```

The service carving is set to **auto**, so the DF election is based on a modulo function of the EVI and the number of DF candidates. In the vES "vESI-23_1", there are two DF candidates, PE-2 and PE-3, listed in that order because PE-2 has the lower system IP address, as follows:

```
*A:PE-3# show service system bgp-evpn ethernet-segment name "vESI-
23_1" all | match "EVI Information" post-lines 20
EVI Information
===============================================================================
EVI                 SvcId               Actv Timer Rem      DF
-------------------------------------------------------------------------------
1                   1                   0                   yes
-------------------------------------------------------------------------------
Number of entries: 1
===============================================================================
-------------------------------------------------------------------------------
DF Candidate list
-------------------------------------------------------------------------------
EVI                                     DF Address
-------------------------------------------------------------------------------
1                                       192.0.2.2
1                                       192.0.2.3
-------------------------------------------------------------------------------
Number of entries: 2
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
```

The first DF candidate from the list will be selected when the result of the modulo function equals 0; the second DF candidate when the result equals 1. The calculation is as follows:

Figure 131 **Calculation**



< EVI > < number of DF candidates> = sequence number DF
1 mod 2 = 1 → 2nd DF candidate in the list is DF → 192.0.2.3 is DF
2 mod 2 = 0 → 1st DF candidate in the list is DF → 192.0.2.2 is DF
26865

The following shows that PE-2 is not the DF for VPLS 1, but it is the DF for VPLS 2:

```
*A:PE-2# show service id 1 ethernet-segment

===============================================================================
SAP Ethernet-Segment Information
===============================================================================
SAP                   Eth-Seg                            Status
-------------------------------------------------------------------------------
lag-1:1.1             vESI-23_1                          NDF
===============================================================================
No sdp entries


*A:PE-2# show service id 2 ethernet-segment
```

```
===============================================================================
SAP Ethernet-Segment Information
===============================================================================
SAP                   Eth-Seg                          Status
-------------------------------------------------------------------------------
lag-1:2.1             vESI-23_2                        DF
===============================================================================
No sdp entries
```

Instead of the preceding show commands, the following tools commands can be used:

```
*A:PE-2# tools dump service system bgp-evpn ethernet-segment "vESI-23_1" evi 1 df

[07/11/2017 07:31:27] Computed DF: 192.0.2.3 (Remote) (Boot Timer Expired: Yes)

*A:PE-2# tools dump service system bgp-evpn ethernet-segment "vESI-23_2" evi 2 df

[07/11/2017 07:31:27] Computed DF: 192.0.2.2 (This Node) (Boot Timer Expired: Yes)
```

Service Carving: Preference-based Manual Mode

To have more control, the vES can be configured in manual mode. The following reconfigures the vES "vESI-23_1" in manual mode, preference-based and revertive with preference 32767 (default) on PE-2 and 5000 on PE-3, whereas vES "vESI-23_2" is preference-based and non-revertive with preference 15000 on PE-2 and 20000 on PE-3.

An EVI range is configured for ES "vESI-23_2", but not for ES "vESI-23_1". When no EVI range is configured, the highest preference wins; for configured EVI ranges, the lowest preference wins. When there are no failures, PE-2 will be the DF for "vESI-23_1" (highest preference) and for "vESI-23_2" (lowest preference for configured EVI 2).

To modify the service-carving mode from auto to manual, the ES must be disabled first (shutdown). The following is configured on PE-2:

```
configure
    service
        system
            bgp-evpn
                ethernet-segment "vESI-23_1" virtual create
                    shutdown
                    service-carving
                        mode manual
                        manual
                            preference create
                            exit
                        exit
                    exit
                    no shutdown
                exit
                ethernet-segment "vESI-23_2" virtual create
                    shutdown
                    service-carving
                        mode manual
                        manual
                            preference non-revertive create
                                value 15000
                            exit
                            evi 2
                        exit
                    exit
                    no shutdown
                exit
```

The keyword **non-revertive** is added for vES "vESI-23_2", but not for vES "vESI-23_1".

The following is configured on PE-3:

```
configure
    service
        system
            bgp-evpn
                ethernet-segment "vESI-23_1" virtual create
                    shutdown
                    service-carving
                        mode manual
                        manual
                            preference create
                                value 5000
                            exit
                        exit
                    exit
                    no shutdown
                exit
                ethernet-segment "vESI-23_2" virtual create
                    shutdown
                    service-carving
                        mode manual
                        manual
                            preference non-revertive create
                                value 20000
                            exit
                            evi 2
                        exit
                    exit
                    no shutdown
                exit
```

For the single-active multi-homing vESs on PE-4 and PE-5, the same preferences are configured manually. The ES configuration on PE-4 is as follows:

```
configure
    service
```

```
        system
            bgp-evpn
                ethernet-segment "vESI-45_1" virtual create
                    esi 01:00:00:00:00:45:01:00:00:01
                    es-activation-timer 3
                    service-carving
                        mode manual
                        manual
                            preference create
                            exit
                        exit
                    exit
                    multi-homing single-active
                    sdp 46
                    vc-id-range 1
                    vc-id-range 500 to 501
                    no shutdown
                exit
                ethernet-segment "vESI-45_2" virtual create
                    esi 01:00:00:00:00:45:02:00:00:01
                    es-activation-timer 3
                    service-carving
                        mode manual
                        manual
                            preference non-revertive create
                                value 15000
                            exit
                            evi 2
                        exit
                    exit
                    multi-homing single-active
                    sdp 46
                    vc-id-range 2
                    no shutdown
                exit
```

The ES configuration on PE-5 is as follows:

```
configure
    service
        system
            bgp-evpn
                ethernet-segment "vESI-45_1" virtual create
                    esi 01:00:00:00:00:45:01:00:00:01
                    es-activation-timer 3
                    service-carving
                        mode manual
                        manual
                            preference create
                                value 5000
                            exit
                        exit
                    exit
                    multi-homing single-active
                    sdp 56
                    vc-id-range 1
                    vc-id-range 500 to 501
                    no shutdown
                exit
                ethernet-segment "vESI-45_2" virtual create
                    esi 01:00:00:00:00:45:02:00:00:01
                    es-activation-timer 3
                    service-carving
                        mode manual
                        manual
                            preference non-revertive create
                                value 20000
                            exit
                            evi 2
                        exit
                    exit
                    multi-homing single-active
                    sdp 56
                    vc-id-range 2
                    no shutdown
                exit
```

The preference configuration must be consistent across the PEs in the ES (manual or auto), otherwise the system reverts to the modulo-based DF election.

With preference-based DF election configured with default preference value 32767 and revertive, PE-4 sends the following BGP-EVPN update to the RR PE-2. The **df-election** extended community shows the DP=0 (revertive) and DF preference 32767.

```
50 2017/07/11 11:29:51.42 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 78
    Flag: 0x90 Type: 14 Len: 34 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.4
        Type: EVPN-Eth-Seg Len: 23 RD: 192.0.2.4:0
              ESI: 01:00:00:00:00:45:01:00:00:01, IP-Len: 4 Orig-IP-Addr: 192.0.2.4
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        df-election::DF-Type:Preference/DP:0/DF-Preference:32767
        target:00:00:00:00:45:01
"
```

The following command shows the information in the preceding BGP-EVPN Ethernet-segment route for "vESI-45_1" sent by PE-4 to the RR PE-2:

```
*A:PE-4# show router bgp routes evpn eth-seg hunt
===============================================================================
 BGP Router ID:192.0.2.4        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
```

```
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP EVPN Eth-Seg Routes
===============================================================================
---snip---
-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
Network        : N/A
Nexthop        : 192.0.2.4
To             : 192.0.2.2
Res. Nexthop   : n/a
Local Pref.    : 100                     Interface Name : NotAvailable
Aggregator AS  : None                    Aggregator     : None
Atomic Aggr.   : Not Atomic              MED            : 0
AIGP Metric    : None
Connector      : None
Community      :
                 df-election::DF-Type:Preference/DP:0/DF-Preference:32767
                 target:00:00:00:00:45:01
Cluster        : No Cluster Members
Originator Id  : None                    Peer Router Id : 192.0.2.2
Origin         : IGP
AS-Path        : No As-Path
EVPN type      : ETH-SEG
ESI            : 01:00:00:00:00:45:01:00:00:01
Originator IP  : 192.0.2.4
Route Dist.    : 192.0.2.4:0
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                       Dest Class     : 0
---snip---
```

The following command shows the DF preference election information for ES "vESI-45_1" with the preference mode revertive, the configured preference value on PE-4 (default 32767), and the operational preference value. No EVI ranges or ISID ranges are configured in this ES.

```
*A:PE-4# show service system bgp-evpn ethernet-segment name "vESI-45_1"
===============================================================================
Service Ethernet Segment
===============================================================================
Name                    : vESI-45_1
Eth Seg Type            : Virtual
Admin State             : Enabled            Oper State         : Up
ESI                     : 01:00:00:00:00:45:01:00:00:01
Multi-homing            : singleActive       Oper Multi-homing  : singleActive
ES SHG Label            : 262132
Source BMAC LSB         : <none>
Sdp Id                  : 46
ES Activation Timer     : 3 secs
Svc Carving             : manual             Oper Svc Carving   : manual
Cfg Range Type          : lowest-pref
-------------------------------------------------------------------------------
```

```
DF Pref Election Information
-------------------------------------------------------------------------------
Preference     Preference     Last Admin Change         Oper Pref      Do No
Mode           Value                                    Value          Preempt
-------------------------------------------------------------------------------
revertive      32767          07/11/2017 11:29:51       32767          Disabled
-------------------------------------------------------------------------------
EVI Ranges: <none>
ISID Ranges: <none>
===============================================================================
```

The following command shows the DF preference election information for ES "vESI-45_2" with the preference mode non-revertive, the configured preference value on PE-4 (15000), and the operational preference value. The only configured EVI range is from 2 to 2. No ISID ranges are configured. For the configured EVI or ISID values, the lowest preference wins, as shown by the **Cfg Range Type : lowest-pref** parameter.

```
*A:PE-4# show service system bgp-evpn ethernet-segment name "vESI-45_2"

===============================================================================
Service Ethernet Segment
===============================================================================
Name                    : vESI-45_2
Eth Seg Type            : Virtual
Admin State             : Enabled            Oper State         : Up
ESI                     : 01:00:00:00:00:45:02:00:00:01
Multi-homing            : singleActive       Oper Multi-homing  : singleActive
ES SHG Label            : 262131
Source BMAC LSB         : <none>
Sdp Id                  : 46
ES Activation Timer     : 3 secs
Svc Carving             : manual             Oper Svc Carving   : manual
Cfg Range Type          : lowest-pref

-------------------------------------------------------------------------------
DF Pref Election Information
-------------------------------------------------------------------------------
Preference     Preference     Last Admin Change         Oper Pref      Do No
Mode           Value                                    Value          Preempt
-------------------------------------------------------------------------------
non-revertive  15000          07/11/2017 11:29:52       15000          Enabled
-------------------------------------------------------------------------------


-------------------------------------------------------------------------------
EVI Ranges
-------------------------------------------------------------------------------
From                                    To
-------------------------------------------------------------------------------
2                                       2
-------------------------------------------------------------------------------
ISID Ranges: <none>
===============================================================================
*A:PE-4#
```

It is important to note that a router will prune a remote PE from the DF candidate list for an ES if it does not receive the corresponding Auto Discovery (AD) per-EVI and AD per-ES routes for that PE. A remote PE will not be shown in the DF Candidate list if its AD per-ES route is withdrawn. This is only true for EVPN. In PBB-EVPN, there are no AD routes, therefore the DF Candidate list is built out of the ES routes only.

DF Election: Higher Preference Prevails for Non-configured EVI Ranges

The PEs run the DF election per PE per EVI, and the elected DF for a service will activate the SAP/Spoke-SDP when the es-activation-timer expires. PE-4 is the DF in "vESI-45_1" used in VPLS 1, as follows. The EVI is not configured in ES "vESI-45_1", so the higher preference prevails. The ES "vESI-45_1" has (default) preference 32767 on PE-4 (DF) and preference 5000 on PE-5 (Non-Designated Forwarder (NDF)).

```
*A:PE-4# show service id 1 ethernet-segment
No sap entries

===============================================================================
SDP Ethernet-Segment Information
===============================================================================
SDP                    Eth-Seg                            Status
-------------------------------------------------------------------------------
46:1                   vESI-45_1                          DF
===============================================================================
*A:PE-4#

*A:PE-5# show service id 1 ethernet-segment
No sap entries

===============================================================================
SDP Ethernet-Segment Information
===============================================================================
SDP                    Eth-Seg                            Status
-------------------------------------------------------------------------------
56:1                   vESI-45_1                          NDF
===============================================================================
*A:PE-5#
```

The preference value can be modified on the fly on an active ES without the need to shut down the ES. This allows the user to force a new DF for the ES for maintenance operations on the former DF or other reasons.

DF Election: Lowest Preference Prevails for Configured EVI Ranges

ES "vESI-45_2" is configured with EVI 2, so the lowest preference prevails. The admin preference value is 15000 on PE-4 and 20000 on PE-5. Both PE-4 and PE-5 are DF candidates, but PE-4 has the lowest preference, so it will be the DF, as follows:

```
*A:PE-4# show service system bgp-evpn ethernet-segment name "vESI-
45_2" all | match "EVI Ranges" post-lines 30
EVI Ranges
-------------------------------------------------------------------------------
From                                    To
-------------------------------------------------------------------------------
2                                       2
-------------------------------------------------------------------------------
ISID Ranges: <none>
===============================================================================

===============================================================================
EVI Information
===============================================================================
EVI                 SvcId               Actv Timer Rem      DF
-------------------------------------------------------------------------------
2                   2                   0                   yes
-------------------------------------------------------------------------------
Number of entries: 1
===============================================================================

-------------------------------------------------------------------------------
DF Candidate list
-------------------------------------------------------------------------------
EVI                                     DF Address
-------------------------------------------------------------------------------
2                                       192.0.2.4
2                                       192.0.2.5
-------------------------------------------------------------------------------
Number of entries: 2
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
```

DF Election: DP Prevails when Preferences are Equal

In the preceding example, PE-4 was the DF in ES "vESI-45_1" because of the higher preference. The ES configuration is modified on PE-5 as follows: the preference is set to the default, which is equal to the preference on PE-4, and the non-revertive (do not preempt - DP) option enabled. The **non-revertive** keyword can only be configured at creation time. An attempt to modify this behavior afterward results in the following error message:

```
*A:PE-5>config>service>system>bgp-evpn>eth-seg>service-carving>manual#
```

```
preference non-revertive create
MINOR: CLI revertive mode can be specified only at creation time.
```

The existing preference first needs to be removed, which can only be done when the ES is disabled (shutdown); if not, the following error is raised:

```
*A:PE-5>config>service>system>bgp-evpn>eth-seg>service-carving>manual#
 no preference
MINOR: SVCMGR #8068 Cannot delete preference - ethernet-segment not shut
```

The service carving in the ES is configured with default preference and non-revertive option, as follows:

```
configure
    service
        system
            bgp-evpn
                ethernet-segment "vESI-45_1" virtual create
                    ahutdown
                    service-carving
                        mode manual
                        manual
                            no preference
                            preference non-revertive create
                            exit
                        exit
                    exit
                    no shutdown
```

The ES configuration on PE-4 remains unchanged, so the behavior is revertive. PE-4 and PE-5 have the same preference, but PE-5 is non-revertive and becomes the DF, as follows:

```
*A:PE-5# show service id 1 ethernet-segment
No sap entries

===============================================================================
SDP Ethernet-Segment Information
===============================================================================
SDP                   Eth-Seg                          Status
-------------------------------------------------------------------------------
56:1                  vESI-45_1                        DF
===============================================================================
*A:PE-5#
```

DF Election: Lowest IP Address Prevails when Preferences and DP are Equal

The vES configuration on PE-4 is modified by enabling the non-revertive option (after deleting the existing preference configuration), as follows:

```
configure
    service
        system
            bgp-evpn
                ethernet-segment "vESI-45_1" virtual create
                    shutdown
                    service-carving
                        mode manual
                        manual
                            no preference
                            preference non-revertive create
                            exit
                        exit
                    exit
                    no shutdown
```

PE-4 and PE-5 have an equal preference and non-revertive behavior. The tiebreaker for the DF selection is the IP address. PE-4 has the lower IP address and becomes the DF, as follows:

```
*A:PE-4# show service id 1 ethernet-segment
No sap entries

===============================================================================
SDP Ethernet-Segment Information
===============================================================================
SDP                     Eth-Seg                           Status
-------------------------------------------------------------------------------
46:1                    vESI-45_1                         DF
===============================================================================
*A:PE-4#
```

Service-carving Configuration Must Be Consistent

When the service carving on one of the PEs in the ES is configured in auto mode while one of the other PEs in the ES is configured in manual mode, the system reverts to modulo-based auto mode. The configuration of ES "vESI-45_1" remains unchanged on PE-4, but is modified on PE-5, as follows:

```
configure
    service
        system
            bgp-evpn
                ethernet-segment "vESI-45_1" virtual create
                    shutdown
                    service-carving
                        manual no preference
                        mode auto
                    exit
                    no shutdown
```

ES "vESI-45_1" will operate in auto mode on PE-4 and on PE-5. The following **show** command on PE-4 shows that the ES is configured in manual mode, but operates in auto mode:

```
*A:PE-4# show service system bgp-evpn ethernet-segment name "vESI-45_1"

===============================================================================
Service Ethernet Segment
===============================================================================
Name                    : vESI-45_1
Eth Seg Type            : Virtual
Admin State             : Enabled            Oper State         : Up
ESI                     : 01:00:00:00:00:45:01:00:00:01
Multi-homing            : singleActive       Oper Multi-homing  : singleActive
ES SHG Label            : 262130
Source BMAC LSB         : <none>
Sdp Id                  : 46
ES Activation Timer     : 3 secs
Svc Carving             : manual             Oper Svc Carving   : auto
Cfg Range Type          : lowest-pref

-------------------------------------------------------------------------------
DF Pref Election Information
-------------------------------------------------------------------------------
Preference     Preference     Last Admin Change        Oper Pref      Do No
Mode           Value                                   Value          Preempt
-------------------------------------------------------------------------------
non-revertive  32767          07/11/2017 12:17:25      32767          Enabled
-------------------------------------------------------------------------------
EVI Ranges: <none>
ISID Ranges: <none>
===============================================================================
*A:PE-4#
```

The following command on PE-5 shows that the ES is configured in auto mode and operates in auto mode:

```
*A:PE-5# show service system bgp-evpn ethernet-segment name "vESI-45_1"
===============================================================================
Service Ethernet Segment
===============================================================================
Name                    : vESI-45_1
Eth Seg Type            : Virtual
Admin State             : Enabled            Oper State         : Up
ESI                     : 01:00:00:00:00:45:01:00:00:01
Multi-homing            : singleActive       Oper Multi-homing  : singleActive
ES SHG Label            : 262132
Source BMAC LSB         : <none>
Sdp Id                  : 56
ES Activation Timer     : 3 secs
Svc Carving             : auto               Oper Svc Carving   : auto
Cfg Range Type          : primary
===============================================================================
*A:PE-5#
```

For the remainder of the chapter, the vES configuration for "vESI-45_1" on PE-4 and PE-5 is restored to the initial settings: the behavior is revertive; PE-4 has the default preference, and PE-5 has preference 5000. When there are no failures, PE-4 is the DF, because it has a higher preference.

Revertive Behavior

When SDP 64 fails on MTU-6, PE-4 becomes the NDF for ES "vESI-45_1" and PE-5 will be the DF instead, as follows. The failure is emulated by disabling the SDP on MTU-6.

```
*A:MTU-6# configure service sdp 64 shutdown
```

When the PE is not a candidate DF because it cannot be used, the operational preference value equals 0, as follows:

```
*A:PE-4# show service system bgp-evpn ethernet-segment name "vESI-
45_1" | match "DF Pref Election" post-lines 6
DF Pref Election Information
-------------------------------------------------------------------------------
Preference      Preference      Last Admin Change          Oper Pref      Do No
Mode            Value                                      Value          Preempt
-------------------------------------------------------------------------------
revertive       32767           07/11/2017 12:31:10        0              Disabled
-------------------------------------------------------------------------------
```

PE-5 is the only DF candidate in the ES and becomes the DF:

```
*A:PE-4# show service system bgp-evpn ethernet-segment name "vESI-45_1" evi 1

===============================================================================
EVI DF and Candidate List
===============================================================================
EVI           SvcId         Actv Timer Rem      DF  DF Last Change
-------------------------------------------------------------------------------
1             1             0                   no  07/11/2017 12:42:07
===============================================================================

===============================================================================
DF Candidates                           Time Added
-------------------------------------------------------------------------------
192.0.2.5                               07/11/2017 12:31:14
-------------------------------------------------------------------------------
Number of entries: 1
===============================================================================
*A:PE-4#


*A:PE-5# show service id 1 ethernet-segment
No sap entries

===============================================================================
SDP Ethernet-Segment Information
===============================================================================
SDP                   Eth-Seg                          Status
-------------------------------------------------------------------------------
56:1                  vESI-45_1                        DF
===============================================================================
*A:PE-5#
```

The preference mode for this vES is revertive and the DF preference for PE-5 is 5000, as follows:

```
*A:PE-5# show service system bgp-evpn ethernet-segment name "vESI-45_1"
| match "DF Pref Election" post-lines 6
DF Pref Election Information
-------------------------------------------------------------------------------
Preference     Preference     Last Admin Change        Oper Pref      Do No
Mode           Value                                   Value          Preempt
-------------------------------------------------------------------------------
revertive      5000           07/11/2017 12:33:46      5000           Disabled
-------------------------------------------------------------------------------
```

When the failure is restored, the system reverts and PE-4 will again be the DF for "vESI-45_1" in VPLS 1.

```
*A:MTU-6# configure service sdp 64 no shutdown

*A:PE-4# show service id 1 ethernet-segment
No sap entries
===============================================================================
SDP Ethernet-Segment Information
===============================================================================
SDP                   Eth-Seg                          Status
-------------------------------------------------------------------------------
46:1                  vESI-45_1                        DF
===============================================================================
*A:PE-4#
```

Non-revertive Behavior

When no failures have occurred, PE-4 is the DF for "vESI-45_2" because the lowest preference prevails for the configured EVI 2. The preference of PE-4 is 15000, which is lower than PE-5's preference of 20000.

```
*A:PE-4# show service id 2 ethernet-segment
No sap entries
===============================================================================
SDP Ethernet-Segment Information
===============================================================================
SDP                   Eth-Seg                          Status
-------------------------------------------------------------------------------
46:2                  vESI-45_2                        DF
===============================================================================
```

When the DF goes down in VPLS 2 on PE-4, PE-5 becomes the DF for "vESI-45_2", as follows:

```
*A:MTU-6# configure service sdp 64 shutdown
```

PE-4 is no longer the DF for "vESI-45_2" and not even a DF candidate anymore. The operational preference value is 0.

```
*A:PE-4# show service system bgp-evpn ethernet-segment name "vESI-
45_2" | match "DF Pref Election" post-lines 6
DF Pref Election Information
-------------------------------------------------------------------------------
Preference     Preference     Last Admin Change         Oper Pref      Do No
Mode           Value                                    Value          Preempt
-------------------------------------------------------------------------------
non-revertive  15000          07/11/2017 11:29:52       0              Disabled
-------------------------------------------------------------------------------
```

PE-5 becomes the DF for "vESI-45_2" in VPLS 2, as follows:

```
*A:PE-5# show service id 2 ethernet-segment
No sap entries
===============================================================================
SDP Ethernet-Segment Information
===============================================================================
SDP                    Eth-Seg                         Status
-------------------------------------------------------------------------------
56:2                   vESI-45_2                       DF
===============================================================================
*A:PE-5#
```

```
*A:PE-5# show service system bgp-evpn ethernet-segment name "vESI-45_2"
| match "DF Pref Election" post-lines 6
DF Pref Election Information
-------------------------------------------------------------------------------
Preference     Preference     Last Admin Change         Oper Pref      Do No
Mode           Value                                    Value          Preempt
-------------------------------------------------------------------------------
non-revertive  20000          07/11/2017 11:32:30       20000          Enabled
```

When the SDP is restored, the DF does not revert even though the list of DF candidates contains both PE-4 and PE-5. The preference mode is non-revertive; therefore, the DP bit has been set. PE-4 will not become the DF, as follows:

```
*A:PE-4# show service system bgp-evpn ethernet-segment name "vESI-45_2" evi 2

===============================================================================
EVI DF and Candidate List
===============================================================================
EVI            SvcId          Actv Timer Rem       DF  DF Last Change
-------------------------------------------------------------------------------
2              2              0                    no  07/11/2017 12:42:07
===============================================================================
===============================================================================
DF Candidates                           Time Added
-------------------------------------------------------------------------------
192.0.2.4                               07/11/2017 12:42:34
192.0.2.5                               07/11/2017 12:38:53
-------------------------------------------------------------------------------
Number of entries: 2
===============================================================================
*A:PE-4#
```

The operational preference value on NDF PE-4 equals the preference value on DF PE-5, as follows. In this example, EVI 2 is included in the configured EVI range, so the lowest preference wins. To avoid the system reverting to the lower preference of 15000, the operational preference is raised to the value of 20000, which equals the preference of the current DF PE-5.

```
*A:PE-4# show service system bgp-evpn ethernet-segment name "vESI-45_2"
| match "DF Pref Election" post-lines 6
DF Pref Election Information
-------------------------------------------------------------------------------
Preference     Preference     Last Admin Change        Oper Pref      Do No
Mode           Value                                   Value          Preempt
-------------------------------------------------------------------------------
non-revertive  15000          07/11/2017 11:29:52      20000          Disabled
```

PE-4 checks its own administrative preference and compares it with the one of the Highest-PE and Lowest-PE that have DP=1 in their ES routes.

- The Highest-PE is the PE with higher preference, using the DP bit (with DP=1 being better) and, after that, the lower PE-IP address as tie-breakers.
- The Lowest-PE is the PE with lower preference, using the DP bit (with DP=1 being better) and, after that, the lower PE-IP address as tie-breakers.

Depending on this comparison, PE-4 will send the ES route with a preference and DP that may be different from its administrative values.

- If PE-4's preference value is higher than the Highest-PE's, PE-4 will send the ES route with an 'in-use' operational preference equal to the Highest-PE's and DP=0.
- If PE-4's preference value is lower than the Lowest-PE's, PE-4 will send the ES route with an 'in-use' operational preference equal to the Lowest-PE's and DP=0.
- If PE-4's preference value is neither higher nor lower than the Highest-PE's or the Lowest-PE's respectively, PE-4 will send the ES route with its administrative [preference,DP]=[15000,1].

In this example, NDF PE-4 sends operational preference 20000 and DP=0, because its admin preference value was lower than the Lowest-PE's (PE-5), as follows:

```
224 2017/07/11 12:51:38.77 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 78
    Flag: 0x90 Type: 14 Len: 34 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.4
        Type: EVPN-Eth-Seg Len: 23 RD: 192.0.2.4:0
              ESI: 01:00:00:00:00:45:02:00:00:01, IP-Len: 4 Orig-IP-Addr: 192.0.2.4
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        df-election::DF-Type:Preference/DP:0/DF-Preference:20000
        target:00:00:00:00:45:02
"
```

With equal preference, the current DF PE-5 sends DP=1, which is preferred over DP=0, so the DF remains unchanged, as follows:

```
271 2017/07/11 12:41:22.90 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 78
    Flag: 0x90 Type: 14 Len: 34 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.5
        Type: EVPN-Eth-Seg Len: 23 RD: 192.0.2.5:0
              ESI: 01:00:00:00:00:45:02:00:00:01, IP-Len: 4 Orig-IP-Addr: 192.0.2.5
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        df-election::DF-Type:Preference/DP:1/DF-Preference:20000
        target:00:00:00:00:45:02
"
```

Either of the following events cause PE-4 to re-advertise its admin preference 15000 and DP=1:

- DF PE-5 withdraws its ES route.
- The admin preference for ES "vESI-45_2" on DF PE-5 is modified by configuration to a value preferred over PE-4's admin preference; in this case, to a value lower than 15000.

The admin preference value can be modified on ES "vESI-45_2" on DF PE-5 on the fly, as follows:

```
configure
    service
```

```
system
    bgp-evpn
        ethernet-segment "vESI-45_2" virtual creat
            service-carving
                manual
                    preference non-revertive create
                        value 10000
                    exit
```

The preference value 10000 is lower than 15000 and, therefore, preferred when the lowest preference wins. PE-5 remains DF, but now there is no need to modify the preference of PE-4, because the system does not need to revert. Therefore, PE-4 can send the admin preference 15000 and configured DP=1, as follows:

```
239 2017/07/11 13:14:30.88 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 78
    Flag: 0x90 Type: 14 Len: 34 Multiprotocol Reachable NLRI:
        Address Family EVPN
        NextHop len 4 NextHop 192.0.2.4
        Type: EVPN-Eth-Seg Len: 23 RD: 192.0.2.4:0
              ESI: 01:00:00:00:00:45:02:00:00:01, IP-Len: 4 Orig-IP-Addr: 192.0.2.4
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        df-election::DF-Type:Preference/DP:1/DF-Preference:15000
        target:00:00:00:00:45:02
"
```

Conclusion

Preference-based DF election offers more control over the DF Election and applies to regular ESs and vESs, either in single-active or in all-active multi-homing mode, in VPLS, I-VPLS, or Epipe services. The DF election is by default revertive, but when preference mode is chosen, it can be configured as non-revertive to reduce service impact.

Shortest Path Bridging for MAC

This chapter describes advanced shortest path bridging for MAC configurations.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

Shortest Path Bridging for MAC (SPBM) is supported in SR OS release 10.0.R4, or later. SPB Static MAC, static backbone-service instance identifiers (ISIDs) and ISID-policies for SPB are supported in SR OS Release 11.0.R4, or later. This chapter was initially written for SR OS release 11.0.R4, but the CLI in the current edition is based on SR OS Release 15.0.R2.

Overview

SPB enables a next generation control plane for Provider Backbone Bridges (PBB) and PBB-VPLS that adds the stability and efficiency of link state to unicast and multicast services (Epipes and I-VPLSs). In addition, SPBM provides resiliency, load-balancing and multicast optimization without the need for any other control plane in the B-VPLS (for example, there is no need for spanning tree, or G.8032, or Multiple MAC Registration Protocol (MMRP)).

SPBM exploits the complete knowledge of backbone addressing, which is a key consequence of the PBB hierarchy, by advertising and distributing the backbone MAC addresses (BMACs) through a link-state protocol, namely IS-IS. An immediate effect of this is that the old “flood-and-learn” can at last be turned off in the backbone and every B-VPLS node in the network will know what destination BMAC addresses are expected and valid. As a result of that, receiving an unknown unicast BMAC on a B-VPLS SAP/PW is indicative of an error, whereupon the frame is discarded (due to the Reverse Path Forwarding Check – RPFC – performed in SPBM) instead of flooded. Furthermore, SPBM allows condensing all the relevant information distribution (unicast and multicast) into a single control protocol: IS-IS.

SPBM can be easily enabled on the existing B-VPLS instances being used for multiplexing I-VPLS/Epipe services, providing the following benefits:

- Per-service flood containment (for I-VPLS services) without the need for an additional protocol such as MMRP.
- Loop avoidance in the B-VPLS domain without the need for MSTP or other technologies.
- No unknown BMAC flooding in the B-VPLS domain.
- No need for MAC notification mechanisms or vMEPs in the B-VPLS to update the B-VPLS forwarding databases (FDBs) (vMEPs can still be configured though for OAM purposes).

Some other characteristics of the SPB implementation in the SR OS are:

- The SR OS SPB implementation always uses Multi-Topology (MT) topology instance zero. However, up to four logical instances (that is, SPB instances in different B-VPLS services) are supported if different topologies are required for different services.
- Area addresses are not used and SPB is assumed to be a single area. SPB must be consistently configured on nodes in the system. SPB Regions information and IS-IS hello logic that detect mismatched configuration are not supported. IS-IS area is always zero.
- SPB uses all-intermediate-systems 09-00-2B-00-00-05 destination MAC to communicate.
- SPB Source ID is always zero.
- SPB uses a separate instance of IS-IS from the base IP IS-IS. IS-IS for SPB is configured in the SPB context under the B-VPLS component. Up to four ISIS-SPB instances are supported, where the instance identifier can be any number between 1024 and 2047. The instance number is not in TLVs.
- Two Equal Cost Tree (ECT) algorithms (IEEE 802.1aq) per SPB instance are supported: low-path-id and high-path-id algorithms.
- SPB link state protocol data units (link state packets) contain BMACs, ISIDs (for multicast services) and link and metric information for an IS-IS database.
  - Epipe ISIDs are not distributed in SR OS SPB allowing high scalability of PBB Epipes.
  - I-VPLS ISIDs are distributed in SR OS SPB and the respective multicast group addresses (composed of PBB-OUI plus ISID) are automatically populated in a manner that provides automatic pruning of multicast to the subset of the multicast tree that supports an I-VPLS with a common ISID. This replaces the function of MMRP and is more efficient than MMRP.

- Multiple ISIS-SPB adjacencies between two nodes are not supported as per the IEEE 802.1aq standard specification. If multiple links between two nodes exist, LAG must be used.

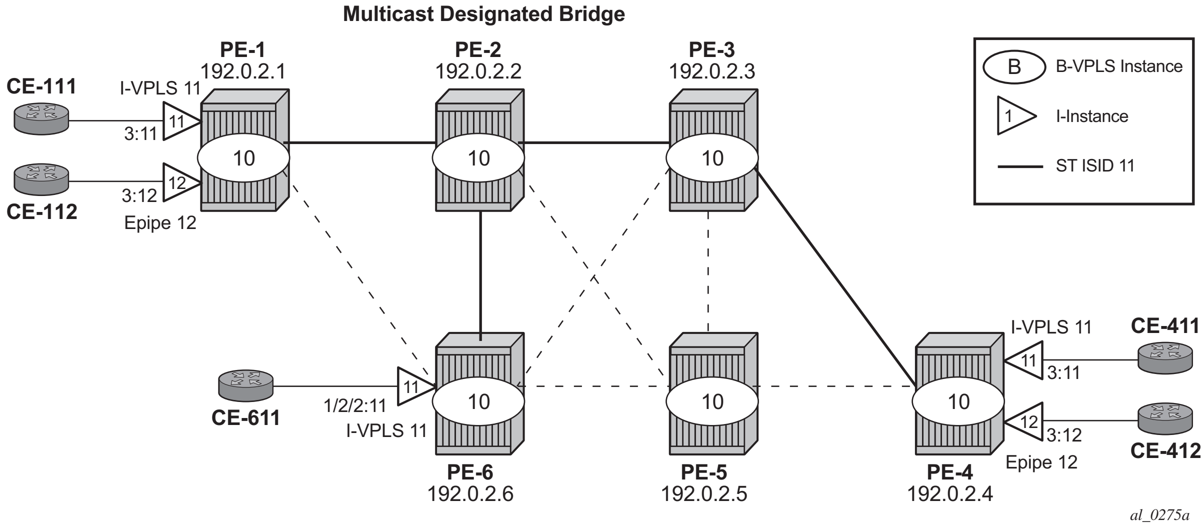
Configuration

This section describes the configuration of SPBM on SR OS as well as the available troubleshooting commands.

Basic SPBM Configuration

Figure 132 shows the topology used as an example of a basic SPBM configuration.

Figure 132 **Basic SPBM Topology**



Assume the following protocols and objects are configured beforehand:

- The six PEs shown in Figure 132 are running IS-IS for the global routing table with all the interfaces being Level-2.
- LDP is used as the MPLS protocol to signal transport tunnel labels.
- LDP SDPs are configured among the six PEs, as shown in Figure 132 (dashed lines and bold lines among PEs).

Once the network infrastructure is properly running, the actual service configuration can be carried out. In the example, B-VPLS 10 will provide backbone connectivity for the services I-VPLS 11 and Epipe 12.

The SPBM configuration is only relevant to the B-VPLS instance and can be added to an existing B-VPLS, assuming that such a B-VPLS does not contain any non-SPB-compatible configuration parameters. The following parameters are not supported in SPB-enabled B-VPLS instances:

- Mesh SDPs (only SAPs or spoke-SDPs are supported in SPB-enabled B-VPLS)
- Spanning Tree Protocol (STP)
- Split-Horizon Groups
- Non-conditional Static-MACs (configured under SAP/spoke-SDPs, see section about static BMAC configuration)
- G.8032
- Propagate-mac-flush and send-flush-on-failure
- Maximum number of MACs (max-nbr-mac-addr)
- Bridge Protocol Data Unit (BPDU) translation
- Layer 2 Protocol Termination (L2PT)
- MAC-pinning
- Oper-groups
- MAC-move
- Any BGP, BGP-AD (BGP auto-discovery), or BGP-VPLS (BGP virtual private LAN services) parameters
- Endpoints
- Local/remote age
- MAC-notification
- MAC-protect
- Multiple MAC Registration Protocol (MMRP)
- Provider-tunnel
- Temporary flooding

Assuming all the parameters mentioned are not configured in the B-VPLS (B-VPLS 10 in the example), SPBM can be enabled. The SPBM parameters are all configured in the **config>service>vpls(b-vpls)>spb** and **config>service>vpls(b-vpls)>spoke-sdp/sap>spb** contexts:

```
*A:PE-1# configure service vpls 10 spb ?
  - no spb
  - spb [<isis-instance>] [fid <fid>] [create]
```

```
 <isis-instance>      : [1024..2047]
 <fid>                : [1..4095]

      level           + Configure SPB level information
 [no] lsp-lifetime    - Configure LSP lifetime
 [no] lsp-refresh-in* - Configure LSP refresh interval
 [no] overload      - Configure the local router so that it appears to be overloaded
 [no] overload-on-bo* - Configure the local router so that it appears to be
                        overloaded at boot up
 [no] shutdown        - Administratively enable or disable the operation of ISIS
      timers          + Configure ISIS timers


*A:PE-1# configure service vpls 10 spb timers ?
  - timers

 [no] lsp-wait        -  [no] spf-wait        - ?


*A:PE-1# configure service vpls 10 spoke-sdp 35:10 spb ?
  - no spb
  - spb [create]

 <create>             : keyword

      level           + Configure SPB level information
 [no] lsp-pacing-int* - Configure the interval between LSP packets are sent from the
                        interface
 [no] retransmit-int* - Configure the minimum interval between LSP packets
                        retransmission for the given interface
 [no] shutdown        - Administratively Enable/disable the interface


*A:PE-1# configure service vpls 10 spoke-sdp 35:10 spb level 1 ?
  - level <[1..1]>

 [no] hello-interval  - Configure hello-interval for this interface
 [no] hello-multipli* - Configure hello-multiplier for this level
 [no] metric          - Configure IS-IS interface metric for IPv4 unicast
```

The parameters configured in the **spb** context refer to the SPB IS-IS and they should be configured following the same considerations as for the IS-IS base instance:

- spb [<isis-instance>] [fid <fid>] [create]
  - <isis-instance> identifies the SPB IS-IS process. Up to four different IS-IS SPB processes can be run in a system (range 1024 to 2047).
  - <fid> or *forwarding identifier* identifies the standard SPBM B-VID which is signaled in IS-IS with each advertised BMAC. Each B-VPLS has a single configurable FID.
- spb>lsp-lifetime <seconds>   : [350..65535]
- spb>lsp-refresh-interval <seconds> : [150..65535]
- spb>overload [timeout <seconds>]   : [60..1800]
- spb>overload-on-boot [timeout <seconds>]   : [60..1800]

- spb>timers>lsp-wait <lsp-wait> [<lsp-initial-wait> [<lsp-second-wait>]]
  - <lsp-wait> : [10..120000] – milliseconds
  - <lsp-initial-wait> : [10..100000] – milliseconds
  - <lsp-second-wait> : [10..100000] – milliseconds
- spb>timers>spf-wait <spf-wait> [<spf-initial-wait> [<second-wait>]]
  - <spf-wait> : [10..120000] – milliseconds
  - <spf-initial-wait> : [10..100000] – milliseconds
  - <second-wait> : [10..100000] – milliseconds
- spoke-sdp/sap>spb>lsp-pacing-interval <milli-seconds> : [0..65535]
  - spoke-sdp/sap>spb>retransmit-interval <seconds> : [1..65535]
  - spoke-sdp/sap>spb>level 1>hello-interval <seconds> : [1..20000]
  - spoke-sdp/sap>spb>level 1>hello-multiplier <multiplier> : [2..100]

In the same way lsp-wait (initial-wait) and spf-wait (initial wait) can be tuned in the base router IS-IS instance to minimize the convergence time (to 0 and 10 respectively), the equivalent SPB IS-IS parameters should also be adjusted so that failover time is minimized at the service level.

The following parameters are specific to SPBM (note that only IS-IS level 1 is supported for SPB):

- spb>level 1>bridge-priority <bridge-priority> : [0..15]
  - This parameter will influence the election of the multicast designated bridge through which all the Single Trees (STs) for the multicast traffic will be established. The default value will be lowered on that node where the multicast designated bridge function is desired, normally because that node is the best connected node. In the example, PE-2 is the multicast designated bridge for B-VPLS 10 and therefore, PE-2 will be the root of the STs for the I-VPLS instances in that B-VPLS. Default value = 8.
- spb>level 1>ect-algorithm fid-range <fid-range> {low-path-id|high-path-id}
  - This command defines the ect-algorithm used and the FIDs assigned. Two algorithms are supported: low-path-id and high-path-id. They can provide the required path diversity for an efficient load balancing in the B-VPLS. Default = fid-range 1-4095 low-path-id
- spb>level 1>forwarding-tree-topology unicast {spf|st}

  - This command configures the type of tree that will be used for unicast traffic: shortest path tree or single tree. The multicast traffic (that encapsulated I-VPLS Broadcast, Unknown unicast and Multicast (BUM) traffic always uses the ST path. Using SPF for unicast traffic can produce some packet re-ordering for unicast traffic compared to BUM traffic because different trees are used, therefore, when the B-VPLS transports I-VPLS traffic and the unicast and multicast trees do not follow the same path, it is recommended to use ST paths for unicast and multicast. Default value = spf.

- spoke-sdp/sap>spb>level 1>metric <ipv4-metric> : [1..16777215]
  - This command configures the metric for each SPB interface (spoke-SDP or SAP). This value helps influence the SPF calculation in order to pick a certain path for the traffic to a remote system BMAC.When the SPB link metric advertised by two peers is different, the maximum value is chosen according to the RFC 6329. Default metric = 10.

As an example, the following CLI output shows the relevant configuration of PE-1 and PE-2 (the multicast designated bridge). SPB has to be created and enabled (**no shutdown**) at B-VPLS service level first and then created and enabled under each and every SAP/spoke-SDP in the B-VPLS. Non-SPB-enabled SAPs/spoke-SDPs can exist in the SPB B-VPLS only if conditional static-MACs are configured for them (see Static BMACs and Static ISIDs Configuration section). As for regular B-VPLS services, the service MTU has to be changed from the default value (1500) to a number 18-bytes greater than the I-VPLS service MTU in order to allow for the PBB encapsulation.

```
*A:PE-1# configure
    service
        pbb
            source-bmac 00-00-00-01-01-01
            mac-name "PE-1" 00-00-00-01-01-01
            mac-name "PE-2" 00-00-00-02-02-02
            mac-name "PE-3" 00-00-00-03-03-03
            mac-name "PE-4" 00-00-00-04-04-04
            mac-name "PE-5" 00-00-00-05-05-05
            mac-name "PE-6" 00-00-00-06-06-06
        exit
        vpls 10 customer 1 b-vpls create
            service-mtu 2000
            spb 1024 fid 10 create
                overload-on-boot timeout 60
                timers
                    spf-wait 2000 spf-initial-wait 50000 spf-second-wait 100000
                    lsp-wait 8000 lsp-initial-wait 10 lsp-second-wait 1000
                exit
                no shutdown
            exit
            spoke-sdp 12:10 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            spoke-sdp 16:10 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            no shutdown
        exit
        vpls 11 customer 1 i-vpls create
            pbb
                backbone-vpls 10
                exit
            exit
            sap 1/1/3:11 create
            exit
            no shutdown
        exit
        epipe 12 customer 1 create
            pbb
                tunnel 10 backbone-dest-mac "PE-4" isid 12
            exit
            sap 1/1/3:12 create
            exit
            no shutdown
        exit
```

As discussed, the **bridge-priority** will influence the election of the multicast designated bridge. By making PE-2's bridge-priority zero, it ensures that PE-2 becomes the root of all the STs for B-VPLS 10 as long as the priority for the rest of the PEs is larger than zero. In case of a tie, the PE owning the lowest system BMAC will be elected as multicast designated bridge. Figure 132 shows the ST for I-VPLS 11 (see a thicker continuous line representing the ST). PE-2 is the root of the ST tree.

```
*A:PE-2# configure
    service
        pbb
            source-bmac 00:00:00:02:02:02
            mac-name "PE-1" 00:00:00:01:01:01
            mac-name "PE-2" 00:00:00:02:02:02
            mac-name "PE-3" 00:00:00:03:03:03
            mac-name "PE-4" 00:00:00:04:04:04
            mac-name "PE-5" 00:00:00:05:05:05
            mac-name "PE-6" 00:00:00:06:06:06
        exit
        vpls 10 customer 1 b-vpls create
            service-mtu 2000
            spb 1024 fid 10 create
                level 1
                    bridge-priority 0
                exit
                overload-on-boot timeout 60
                timers
                    spf-wait 2000 spf-initial-wait 50000 spf-second-wait 100000
                    lsp-wait 8000 lsp-initial-wait 10 lsp-second-wait 1000
                exit
                no shutdown
            exit
            spoke-sdp 21:10 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            spoke-sdp 23:10 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            spoke-sdp 25:10 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            spoke-sdp 26:10 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            no shutdown
        exit
```

The rest of the nodes will be configured accordingly. SPB instance 1024 will set up Shortest Path First (SPF) trees for unicast traffic and a Single Tree (ST) per ISID with PE-2 as the root-bridge (because it has the lowest bridge-priority 0 configured) for BUM traffic. The ECT algorithm chosen for the B-VPLS FID (10) is the low-path-id (default one).

Once SPBM is configured on all the six nodes, the six system BMACs and the ISID 11 will be advertised by SPB IS-IS.

The following show commands can help understand the IS-IS configuration for SPB 1024 and the BMACs populated by IS-IS:

- **show service id spb base**: provides the SPB configuration and parameters for a particular SPB B-VPLS.
- **show service id 10 spb fdb:** provides the B-VPLS FDB that has been populated by IS-IS, for the unicast and multicast entries.

```
A:PE-1# show service id 10 spb base

===============================================================================
Service SPB Information
===============================================================================
Admin State        : Up                 Oper State         : Up
ISIS Instance      : 1024               FID                : 10
Bridge Priority    : 8                  Fwd Tree Top Ucast : spf
Fwd Tree Top Mcast : st
```

```
Bridge Id          : 80:00.00:00:00:01:01:01
Mcast Desig Bridge : 00:00.00:00:00:02:02:02
===============================================================================
Router Base ISIS Instance 1024 Interfaces
===============================================================================
Interface                        Level CircID  Oper State   L1/L2 Metric
-------------------------------------------------------------------------------
sdp:12:10                        L1    65536   Up           10/-
sdp:16:10                        L1    65537   Up           10/-
-------------------------------------------------------------------------------
Interfaces : 2
===============================================================================
FID ranges using ECT Algorithm
-------------------------------------------------------------------------------
1-4095    low-path-id
===============================================================================
A:PE-1#
```

```
*A:PE-1# show service id 10 spb fdb

==============================================================================
User service FDB information
==============================================================================
MAC Addr          UCast Source          State   MCast Source          State
------------------------------------------------------------------------------
00:00:00:02:02:02 12:10                 ok      12:10                 ok
00:00:00:03:03:03 12:10                 ok      12:10                 ok
00:00:00:04:04:04 12:10                 ok      12:10                 ok
00:00:00:05:05:05 12:10                 ok      12:10                 ok
00:00:00:06:06:06 16:10                 ok      12:10                 ok
------------------------------------------------------------------------------
Entries found: 5
==============================================================================
*A:PE-1#
```

It can be seen that the unicast (SPF) tree and the multicast (ST) tree differ with respect to PE-6.

The following commands help check the unicast and multicast topology for B-VPLS 10:

- **show service id 10 spb routes** provides a detailed view of the unicast and multicast routes computed by SPF. As shown below, the SPB unicast and multicast routes match on PE-2 since this node is the multicast designated bridge. Unicast and multicast routes will differ on most other nodes.
- **show service id 10 spb mfib** and **show service id 10 mfib** show information of the MFIB entries generated in the B-VPLS as well as the outgoing interface (OIF) associated with those MFIB entries.

```
*A:PE-2# show service id 10 spb routes

================================================================
MAC Route Table
================================================================
FID  MAC Addr                                     Ver.   Metric
      NextHop If                    SysID
-----------------------------------------------------------------

Fwd Tree: unicast
-----------------------------------------------------------------
10   00:00:00:01:01:01                            4      10
      sdp:21:10                     PE-1
10   00:00:00:03:03:03                            6      10
      sdp:23:10                     PE-3
10   00:00:00:04:04:04                            8      20
      sdp:23:10                     PE-3
10   00:00:00:05:05:05                            12     10
      sdp:25:10                     PE-5
10   00:00:00:06:06:06                            14     10
      sdp:26:10                     PE-6

Fwd Tree: multicast
-----------------------------------------------------------------
10   00:00:00:01:01:01                            4      10
      sdp:21:10                     PE-1
10   00:00:00:03:03:03                            6      10
      sdp:23:10                     PE-3
10   00:00:00:04:04:04                            8      20
      sdp:23:10                     PE-3
10   00:00:00:05:05:05                            12     10
      sdp:25:10                     PE-5
10   00:00:00:06:06:06                            14     10
      sdp:26:10                     PE-6
-----------------------------------------------------------------
No. of MAC Routes: 10
=================================================================
=================================================================
ISID Route Table
=================================================================
FID  ISID                                         Ver.
      NextHop If                    SysID
-----------------------------------------------------------------
10   11                                           17
      sdp:21:10                     PE-1
      sdp:23:10                     PE-3
      sdp:26:10                     PE-6
-----------------------------------------------------------------
No. of ISID Routes: 1
=================================================================
*A:PE-2#


*A:PE-2# show service id 10 spb mfib

===============================================================================
User service MFIB information
===============================================================================
MAC Addr          ISID     Status
-------------------------------------------------------------------------------
01:1E:83:00:00:0B 11       Ok
-------------------------------------------------------------------------------
Entries found: 1
===============================================================================
*A:PE-2#
```

```
*A:PE-2# show service id 10 mfib

===============================================================================
Multicast FIB, Service 10
===============================================================================
Source Address  Group Address         Sap/Sdp Id               Svc Id   Fwd/Blk
-------------------------------------------------------------------------------
*               01:1E:83:00:00:0B     b-sdp:21:10              Local    Fwd
                                      b-sdp:23:10              Local    Fwd
                                      b-sdp:26:10              Local    Fwd
-------------------------------------------------------------------------------
Number of entries: 1
===============================================================================
*A:PE-2#
```

SPB multicast trees (STs) are pruned for each particular I-VPLS ISID, based on the advertisement of I-VPLS ISIDs in SPB IS-IS by each individual PE. Multicast B-VPLS traffic not belonging to any particular I-VPLS follows the *default tree*. The default tree is an ST for the B-VPLS which is not pruned and therefore reaches all the PE nodes in the B-VPLS. For instance, Ethernet-CFM CCM messages sent from vMEPs configured on the SPB B-VPLS will use the default tree. The default tree does not consume MFIB entries and can be checked in each node through the use of the following command:

```
*A:PE-5# tools dump service id 10 spb default-multicast-list
saps : { }
spoke-sdps : { 52:10 }
```

PE-5 is not part of the tree for I-VPLS 11. However, as with any SPB node part of B-VPLS 10, PE-5 is part of the default tree. Refer to Configuration of ISID-Policies in SPB B-VPLS to see more use-cases for the default tree.

The following tools commands allow the operator to easily see the forwarding path (unicast and multicast) followed by the traffic to a remote node, with the aggregate metric from the source.

```
*A:PE-1# tools dump service id 10 spb fid 10 forwarding-path destination PE-4
forwarding-tree unicast
Hop  BridgeId                 Metric From Src
0    PE-1                     0
1    PE-2                     10
2    PE-3                     20
3    PE-4                     30

*A:PE-1# tools dump service id 10 spb fid 10 forwarding-path destination PE-4
forwarding-tree multicast
Hop  BridgeId                 Metric From Src
0    PE-1                     0
1    PE-2                     10
2    PE-3                     20
3    PE-4                     30
```

In large networks or networks where IP multicast, PBB and PBB-SPB services coexist, the data plane MFIB entries is a hardware resource that should be periodically checked. The tools dump service vpls-mfib-stats shows the total number of hardware MFIB entries and the entries being used by IP multicast or PBB (MMRP or SPB). The tools dump service vpls-pbb-mfib-stats shows the breakdown between MFIB entries populated by MMRP, SPB, or by EVPN, and the individual limits, system-wide and per service:

```
*A:PE-2# tools dump service vpls-mfib-stats
Service Manager VPLS MFIB info at 000 00:45:28.450:

Statistics last cleared at 000 00:00:00.000

            Statistic            |    Count
---------------------------------+-------------
             HW limit SG entries |      40959   # total number of MFIB entries
              Current SG entries |          1
        Limit Non PBB SG entries |      16383   # IP Multicast MFIB limit
      Current Non PBB SG entries |          0
---snip---

*A:PE-2# tools dump service vpls-pbb-mfib-stats detail

Service Manager VPLS PBB MFIB statistics at 000 00:45:28.540:

Usage per Service
   ServiceId    MFIB User      Count
  ------------+-------------+-------
   10           spb              1
  ------------+-------------+-------
                       Total       1

MMRP
  Current Usage     :     0
  System Limit      :  8191 Full, 40959 ESOnly
  Per Service Limit :  2048 Full,  8192 ESOnly

SPB
  Current Usage     :     1
  System Limit      :  8191
  Per Service Limit :  8191

Evpn
  Current Usage     :     0
  System Limit      : 40959
  Per Service Limit :  8191
```

Finally, the following debug commands can help monitor the SPB IS-IS process and the protocol PDU exchanges:

- debug service id <svcId> spb
- debug service id <svcId> spb adjacency
- debug service id <svcId> spb interface

- debug service id <svcId> spb l2db
- debug service id <svcId> spb lsdb
- debug service id <svcId> spb packet <detail>
- debug service id <svcId> spb spf

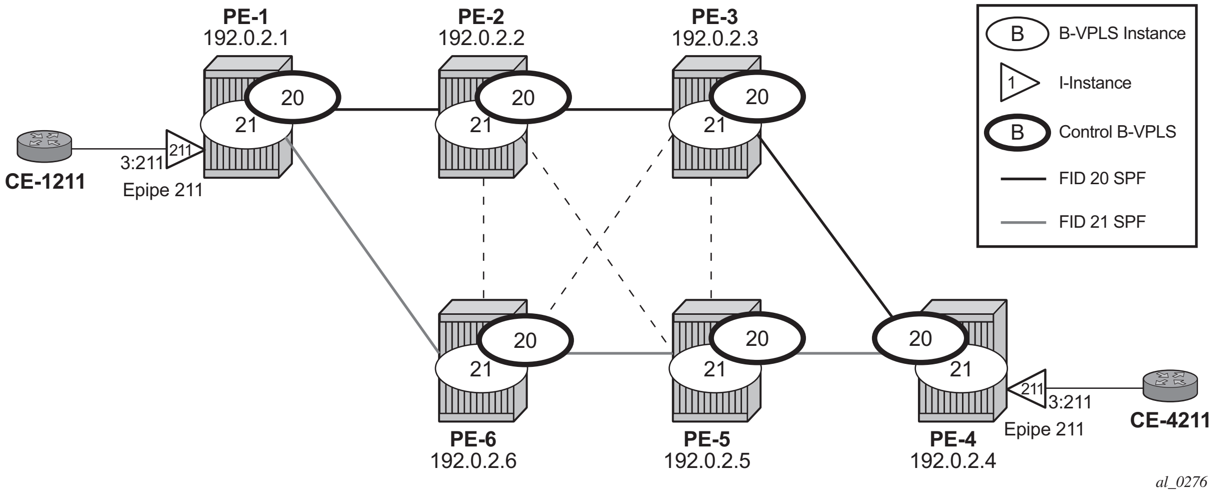
Control and User B-VPLS Configuration

The SR OS implementation of SPB allows a single SPB IS-IS instance to control the paths and FDBs of many B-VPLS instances. This is done by using the control B-VPLS, user B-VPLS, and fate-sharing concepts.

The control B-VPLS will be SPB-enabled and configured with all the related SPB IS-IS parameters. Although the control B-VPLS might or might not have I-VPLS/Epipes directly attached, it must be configured on all the nodes where SPB forwarding is expected to be active. SPB uses the logical instance and a Forwarding ID (FID) to identify SPB locally on the node. That FID must be consistently configured on all the nodes where the B-VPLS exists. User B-VPLS are other instances of B-VPLS that are usually configured to separate the traffic for manageability reasons, QoS, or ECT different treatment.

Figure 133 illustrates the control B-VPLS (B-VPLS 20) and user B-VPLS (B-VPLS 21) concept (in this case, there is only one user B-VPLS, but there might be many B-VPLSs sharing fate with the same control B-VPLS). Both B-VPLSs must share the same topology and both B-VPLSs must share exactly the same interfaces. The user B-VPLS, which is linked to the control B-VPLS by its FID, follows (that is, inherits the state of) the control B-VPLS, but may use a different ECT path in case of equal metric paths, like in this example: FID 20, that is, the control B-VPLS, follows the low-path-id ECT, whereas FID 21, for example, the user B-VPLS, follows the high-path-id ECT.

***Figure 133* Control and User B-VPLS Test Topology**



The configurations of B-VPLSs 20 and 21, on PE-1 and PE-2, are as follows. The **spbm-control-vpls 20 fid 21** command in B-VPLS 21 associates FID 21 to the user B-VPLS and links the B-VPLS to its control B-VPLS 20.

```
*A:PE-1# configure
    service
        vpls 20 customer 1 b-vpls create
            service-mtu 2000
            spb 1025 fid 20 create
                level 1
                    ect-algorithm fid-range 21-4095 high-path-id
                exit
                no shutdown
            exit
            spoke-sdp 12:20 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            spoke-sdp 16:20 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            no shutdown
        exit
        vpls 21 customer 1 b-vpls create
            service-mtu 2000
            spbm-control-vpls 20 fid 21
            spoke-sdp 12:21 create
                no shutdown
            exit
            spoke-sdp 16:21 create
                no shutdown
            exit
            no shutdown
        exit
        epipe 211 customer 1 create
            pbb
                tunnel 21 backbone-dest-mac "PE-4" isid 211
            exit
            sap 1/1/3:211 create
            exit
            no shutdown
        exit
```

```
*A:PE-2# configure
    service
        vpls 20 customer 1 b-vpls create
            service-mtu 2000
            spb 1025 fid 20 create
                level 1
                    ect-algorithm fid-range 21-4095 high-path-id
                exit
                no shutdown
            exit
            spoke-sdp 21:20 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            spoke-sdp 23:20 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            spoke-sdp 25:20 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            spoke-sdp 26:20 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            no shutdown
        exit
        vpls 21 customer 1 b-vpls create
            service-mtu 2000
            spbm-control-vpls 20 fid 21
            spoke-sdp 21:21 create
                no shutdown
            exit
            spoke-sdp 23:21 create
                no shutdown
            exit
            spoke-sdp 25:21 create
                no shutdown
            exit
            spoke-sdp 26:21 create
                no shutdown
            exit
            no shutdown
        exit
```

If there is a mismatch between the topology of a user B-VPLS and its control B-VPLS, only the user B-VPLS links and nodes that are in common with the control B-VPLS will function.

User B-VPLS instances supporting only unicast services (PBB-Epipes) may share the FID with the other B-VPLS (control or user). This is a configuration shortcut that reduces the LSP advertisement size for B-VPLS services but results in the same separation for forwarding between the B-VPLS services. In the case of PBB-Epipes, only BMACs are advertised per FID, but BMACs are populated per B-VPLS in the FIB. If I-VPLS services are to be supported on a B-VPLS that B-VPLS must have an independent FID.

Although user B-VPLS 21 does not have any SPB setting (other than the spbm-control-vpls), the spoke-SDPs use the same SDPs as the parent control B-VPLS 20. The **show service id <user b-vpls> spb fate-sharing** command shows the control spoke-SDP/SAPs that control the user spoke-SDP/SAPs.

```
*A:PE-1# show service id 21 spb fate-sharing

===============================================================================
User service fate-shared sap/sdp-bind information
===============================================================================
Control   Control Sap/             FID       User      User Sap/
SvcId     SdpBind                            SvcId     SdpBind
-------------------------------------------------------------------------------
20        12:20                    21        21        12:21
20        16:20                    21        21        16:21
===============================================================================
*A:PE-1#
```

SPBM Access Resiliency Configuration

The following example shows how to configure an I-VPLS/Epipe attached to an SPB-enabled B-VPLS when access resiliency is used.

Multi-Chassis LAG (MC-LAG) is the only resiliency mechanism supported for PBB-Epipes. The MC-LAG active node will advertise the MC-LAG BMAC (or sap-bmac) in SPB IS-IS. In case of failure, when the standby node takes over, it will advertise the MC-LAG sap-bmac. Without SPB, the MC-LAG solution for PBB-Epipe required the use of MAC-notification and periodic MAC-notification. SPB provides a faster and more efficient solution without the need for any extra MAC-notification mechanism. In the example described in this section, Epipe 31 uses MC-LAG access resiliency to get connected to the B-VPLS 30 on nodes PE-2 and PE-6.

As far as I-VPLS access resiliency is concerned, the same mechanisms supported for regular B-VPLS are supported for SPB-enabled B-VPLS, except for G.8032. A very important aspect of the I-VPLS resiliency is a proper mac-flush propagation when there is a failure at the I-VPLS access links.

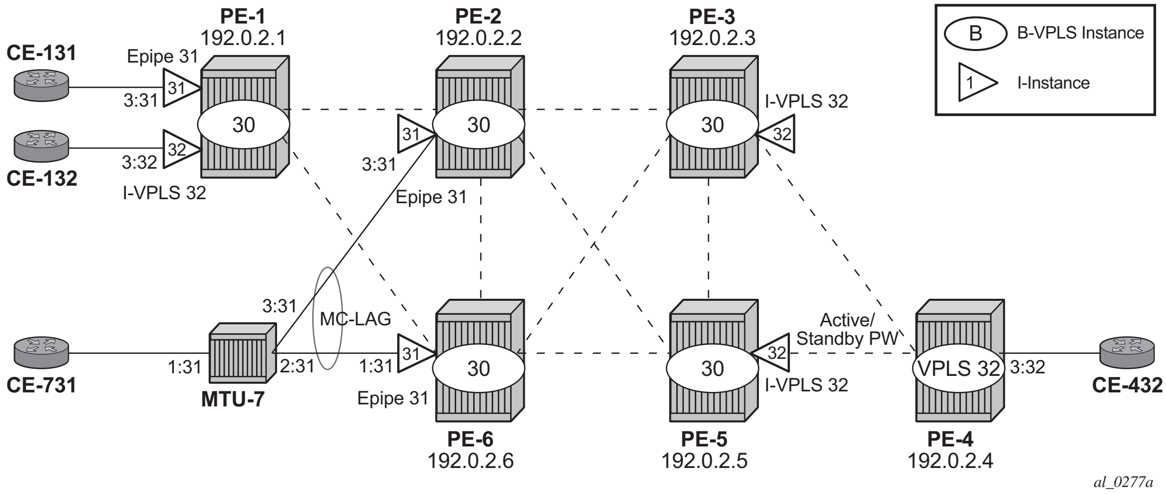
If the SPB-enabled B-VPLS uses B-SAPs for its connectivity to the backbone, there is no mac-flush propagation (because there is no TLDP). In this case, if MC-LAG is used and there is an MC-LAG switchover, the new active chassis will keep using the same source BMAC, such as the sap-bmac, and it will advertise it in the B-VPLS domain so that the remote FDBs can be properly updated. No mac-flush is required in this case.

When the B-VPLS uses spoke-SDPs for its backbone connectivity, the traditional LDP MAC flush propagation mechanisms and commands can be used as follows:

- **send-flush-on-failure** works as expected when SPB is used at the B-VPLS. When configured, a flush-all-from-me event is triggered upon a SAP or spoke-SDP failure in the I-VPLS.
- **send-bvpls-flush** works as expected when SPB is used at the B-VPLS. Two variants are configurable: all-from-me/all-but-mine. Any I-VPLS SAP/spoke-SDP failure is propagated to the I-VPLS on the peers to flush their respective customer MACs (CMACs). It works only in conjunction with send-flush-on-failure configuration on I-VPLS. The associated ISID list is passed along with the LDP mac-flush message, which is flushed/retained according to the **all-from-me/all-but-me** flag.
- **send-flush-on-bvpls-failure** works as expected when SPB is used at the B-VPLS. A local B-VPLS failure is propagated to the I-VPLS, which then triggers a LDP mac-flush if it has any spoke-sdp on it.
- **propagate-mac-flush-from-bvpls** does not work when SPB is used at the B-VPLS (because failures within the B-VPLS are handled by SPB) and its configuration is blocked.

In the example described later in this section, I-VPLS 32 uses active/standby spoke-sdp resiliency to get connected to the B-VPLS 30 on nodes PE-3 and PE-5.

Figure 134 **Access Resiliency Test Topology**



As an example of MC-LAG connectivity, the Epipe 31 configuration is shown. Just like for regular PBB-VPLS, a sap-bmac is used as source BMAC for the Epipe traffic from PE-2/PE-6 to PE-1. A sap-bmac is a virtual BMAC formed from the configured source-bmac plus the MC-LAG LACP-key (if configured this way) and owned by the MC-LAG active chassis. The following CLI output shows the configuration of MC-LAG as well as the generation of the sap-bmac. Once it is properly configured and the MC-LAG and Epipe are up and running, SPB IS-IS will distribute the sap-bmac throughout the B-VPLS, as it does for the system BMACs and OAM vMEP MACs. In this example, PE-2 is the MC-LAG active node, therefore the sap-bmac for Epipe 31 is generated from PE-2.

```
*A:PE-2# configure
    lag 1
        mode access
        encap-type dot1q
        port 1/1/3
        lacp active administrative-key 32768
        no shutdown
        exit


*A:PE-2# configure
    redundancy
        multi-chassis
            peer 192.0.2.6 create
                mc-lag
                    lag 1 lacp-key 1 system-id 00:00:00:00:02:06 system-priority 65535
                                     source-bmac-lsb use-lacp-key
                    no shutdown
                exit
                no shutdown
            exit
        exit
```

```
*A:PE-2# configure
    service
        vpls 30 customer 1 b-vpls create
            service-mtu 2000
            pbb
                use-sap-bmac
            exit
            spb 1026 fid 30 create
                level 1
                    bridge-priority 0
                exit
                no shutdown
            exit
            spoke-sdp 21:30 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            spoke-sdp 23:30 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            spoke-sdp 25:30 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            spoke-sdp 26:30 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            no shutdown
        exit
        epipe 31 customer 1 create
            pbb
                tunnel 30 backbone-dest-mac "PE-1" isid 31
            exit
            sap lag-1:31 create
            exit
            no shutdown
        exit
```

```
*A:PE-6# show service id 30 spb fdb

===============================================================================
User service FDB information
===============================================================================
MAC Addr          UCast Source          State   MCast Source          State
-------------------------------------------------------------------------------
00:00:00:01:01:01 61:30                 ok      62:30                 ok
00:00:00:02:00:01 62:30                 ok      62:30                 ok
00:00:00:02:02:02 62:30                 ok      62:30                 ok
00:00:00:03:03:03 63:30                 ok      62:30                 ok
```

```
00:00:00:05:05:05 65:30                  ok      62:30                 ok
-------------------------------------------------------------------------------
Entries found: 5
===============================================================================
```

The configuration for I-VPLS 32 on nodes PE-4 and PE-3 is as follows.

```
*A:PE-4# configure
    service
        vpls 32 customer 1 create     # Ordinary VPLS, no I-VPLS (no B-VPLS present)
            endpoint "CORE" create
                no suppress-standby-signaling
            exit
            sap 1/1/3:32 create
            exit
            spoke-sdp 43:32 endpoint "CORE" create
                precedence primary
                no shutdown
            exit
            spoke-sdp 45:32 endpoint "CORE" create
                no shutdown
            exit
            no shutdown
        exit
```

```
*A:PE-3# configure
    service
        vpls 30 customer 1 b-vpls create
            service-mtu 2000
            spb 1026 fid 30 create
                no shutdown
            exit
            spoke-sdp 32:30 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            spoke-sdp 35:30 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            spoke-sdp 36:30 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            no shutdown
        exit
        vpls 32 customer 1 i-vpls create
            send-flush-on-failure
            pbb
                backbone-vpls 30
                exit
                send-flush-on-bvpls-failure
                send-bvpls-flush all-from-me
            exit
            spoke-sdp 34:32 create
                no shutdown
            exit
            no shutdown
        exit
```

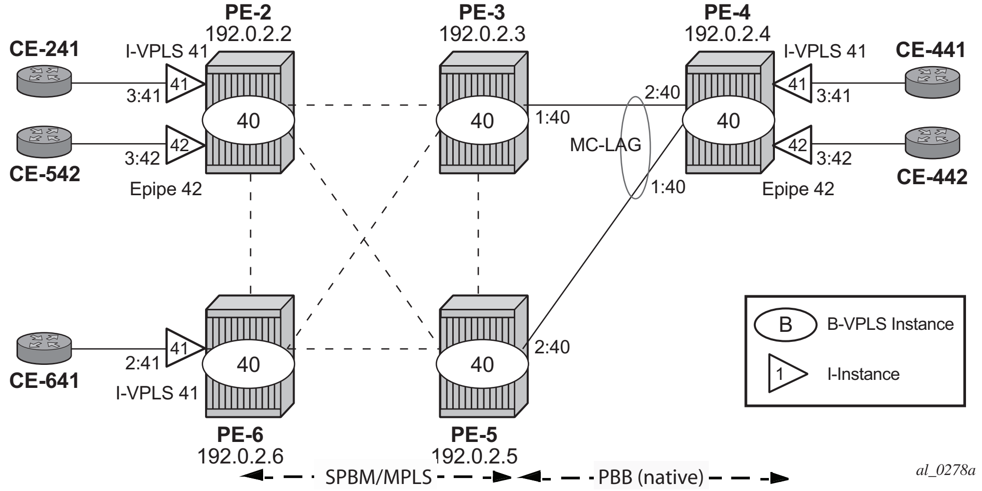
As discussed, **send-flush-on-failure** and **send-bvpls-flush all-from-me** are configured in the I-VPLS. When the active spoke-SDP goes down on PE-3, a flush-all-from-me message will be propagated through the backbone and will flush the corresponding CMACs associated to the I-VPLS 32 in node PE-1. MAC flush-all-from-me messages are automatically propagated in the core up to the remote I-VPLS 32 on node PE-1 (there is no need for any propagate-mac-flush in the intermediate nodes). The *send-flush-on-bvpls-failure* command works as expected. The command *propagate-mac-flush-from-bvpls* is never used when the B-VPLS is SPB-enabled (the command is not allowed).

Static BMACs and Static ISIDs Configuration

SR OS supports the interworking between SPB-enabled B-VPLS and non-SPB B-VPLS instances. SPB networks can be connected to non-SPB capable nodes, for example third party vendor PBB switches or 7210 SAS nodes. This is possible through the use of conditional static BMACs and static ISIDs on the nodes doing the interworking function. Conditional static BMACs and static ISIDs can be associated to non-SPB B-VPLS SAPs or spoke-SDPs.

The following example shows an SPB-enabled B-VPLS (40) on nodes PE-2, PE-6, PE-3 and PE-5. Node PE-4 supports PBB, but not SPB and it is connected by a MC-LAG to nodes PE-3 and PE-5. Services I-VPLS 41 and Epipe 42 have end-points on node PE-4. In this example, nodes PE-3 and PE-5 are acting as interworking nodes. They will be configured with the BMAC of PE-4 so that the MC-LAG active node advertises the non-SPB capable node BMAC into SPB IS-IS. The BMAC will be configured as a conditional static BMAC so that an SPB node, such as PE-3 or PE-5, will only advertise PE-4's BMAC if its connection to PE-4 is active. Besides the conditional static BMAC, nodes PE-3/PE-5 should advertise the I-VPLS ISIDs defined in PE-4. Epipe ISIDs are not advertised in SPB IS-IS, therefore it is not necessary to create a static ISID for Epipe 42.

Figure 135 **Access Resiliency Example Topology**



The commands to configure conditional static BMACs and static ISIDs are as follows.

```
*A:PE-3# configure service vpls 40 static-mac mac
  - mac <ieee-address> [create] black-hole
  - mac <ieee-address> [create] sap <sap-id> monitor {fwd-status}
  - no mac <ieee-address>
  - mac <ieee-address> [create] spoke-sdp <sdp-id:vc-id> monitor {fwd-status}
---snip---
```

```
*A:PE-3# configure service vpls 40 sap lag-1:40 static-isid range
  - no range <range-id>
  - range <range-id> isid <isid-value> [to <isid-value>] [create]

 <range-id>           : [1..8191]
 <isid-value>         : [1..16777215]
 <create>             : keyword
```

The **monitor fwd-status** attribute identifies this to be a conditional MAC and is mandatory for static BMACs. This parameter instructs SR OS to advertise the BMAC only if the corresponding SAP/spoke-sdp is in forwarding state.

The configuration of the conditional static BMAC and static ISID is as follows. The values for **spf-wait** are the default ones.

```
*A:PE-3# configure
    service
        vpls 40 customer 1 b-vpls create
            service-mtu 2000
            spb 1027 fid 40 create
                timers
                    spf-wait 10000 spf-initial-wait 10 spf-second-wait 1000
                exit
                no shutdown
            exit
            sap lag-1:40 create
                static-isid
                    range 1 create isid 41
                exit
            exit
            spoke-sdp 32:40 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            spoke-sdp 35:40 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            spoke-sdp 36:40 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            static-mac
                mac 00:00:00:04:04:04 create sap lag-1:40 monitor fwd-status
            exit
            no shutdown
        exit

*A:PE-5# configure
    service
        vpls 40 customer 1 b-vpls create
            service-mtu 2000
            spb 1027 fid 40 create
                timers
                    spf-wait 10000 spf-initial-wait 10 spf-second-wait 1000
                exit
                no shutdown
            exit
            sap lag-1:40 create
                static-isid
                    range 1 create isid 41
                exit
            exit
            spoke-sdp 52:40 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            spoke-sdp 53:40 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            spoke-sdp 56:40 create
                spb create
                    no shutdown
                exit
                no shutdown
            exit
            static-mac
                mac 00:00:00:04:04:04 create sap lag-1:40 monitor fwd-status
            exit
            no shutdown
        exit
```

The configuration of the conditional static BMAC is different from the legacy **static-mac** command, configured within the SAP/SDP-binding context. The latter static-MAC is not conditional and it is always added to the FDB. The conditional static BMAC is added to the FDB based on the SAP/SDP-binding state (the conditional static BMAC is tagged in the FDB as **CStatic**, for Conditional Static).

```
*A:PE-3# show lag 1

================================================================================
Lag Data
================================================================================
Lag-id         Adm     Opr     Weighted Threshold Up-Count MC Act/Stdby
--------------------------------------------------------------------------------
1              up      up      No       0         1        active
================================================================================
```

```
*A:PE-3# show service id 40 fdb pbb

========================================================================
Forwarding Database, b-Vpls Service 40
========================================================================
MAC               Source-Identifier     iVplsMACs  Epipes     Type/Age
------------------------------------------------------------------------
00:00:00:02:02:02 sdp:32:40             0          0          Spb
00:00:00:04:04:04 sap:lag-1:40          0          0          CStatic
00:00:00:05:05:05 sdp:35:40             0          0          Spb
00:00:00:06:06:06 sdp:36:40             0          0          Spb
========================================================================
```

On PE-5, LAG 1 is in standby, as follows:

```
*A:PE-5# show lag 1

================================================================================
Lag Data
================================================================================
Lag-id         Adm     Opr     Weighted Threshold Up-Count MC Act/Stdby
--------------------------------------------------------------------------------
1              up      down    No       0         0        standby
================================================================================
```

SAP LAG 1 in VPLS 40 is not forwarding any traffic. The FDB for VPLS 40 on PE-5 does not contain any conditional static MAC addresses, even though MAC 00:00:00:04:04:04 is configured on SAP LAG 1. In the FDB for VPLS 40 on PE-5, this MAC address is assigned to SDP 53:40 (type SPB), as follows:

```
*A:PE-5# show service id 40 fdb pbb

=======================================================================
Forwarding Database, b-Vpls Service 40
=======================================================================
MAC               Source-Identifier     iVplsMACs  Epipes     Type/Age
-----------------------------------------------------------------------
00:00:00:02:02:02 sdp:52:40             0          0          Spb
00:00:00:03:03:03 sdp:53:40             0          0          Spb
00:00:00:04:04:04 sdp:53:40             0          0          Spb
00:00:00:06:06:06 sdp:56:40             0          0          Spb
===============================================================================
```

The **static-isid** command identifies a set of ISIDs for I-VPLS services that are external to SPBM. These ISIDs are advertised as supported locally on this node unless altered by an ISID-policy. Although the preceding example shows the use of the static-isid associated to a MC-LAG SAP, regular SAPs or spoke-SDPs are also supported. ISIDs declared in this way become part of the ISID multicast and consume MFIBs. Multiple SPBM static-ISID ranges are allowed under a SAP/spoke SDP. ISIDs are advertised as if they were attached to the local BMAC. Only remote I-VPLS ISIDs need to be defined. In the MFIB, the backbone group MACs are then associated with the active SAP or spoke SDP.

Once the conditional static BMAC for PE-4 and the static-ISID 41 (for I-VPLS 41) are configured as discussed, the advertised BMAC and ISID can be checked in the remote SPB nodes:

```
*A:PE-6# show service id 40 spb fdb

==============================================================================
User service FDB information
==============================================================================
MAC Addr          UCast Source          State   MCast Source          State
------------------------------------------------------------------------------
00:00:00:02:02:02 62:40                 ok      62:40                 ok
00:00:00:03:03:03 63:40                 ok      62:40                 ok
00:00:00:04:04:04 63:40                 ok      62:40                 ok
00:00:00:05:05:05 65:40                 ok      62:40                 ok
------------------------------------------------------------------------------
Entries found: 4
==============================================================================


*A:PE-6# show service id 40 spb mfib
===============================================================================
User service MFIB information
===============================================================================
MAC Addr          ISID     Status
-------------------------------------------------------------------------------
01:1E:83:00:00:29 41       Ok
-------------------------------------------------------------------------------
Entries found: 1
===============================================================================


*A:PE-6# show service id 40 mfib
```

```
===============================================================================
Multicast FIB, Service 40
===============================================================================
Source Address  Group Address         Sap/Sdp Id               Svc Id   Fwd/Blk
-------------------------------------------------------------------------------
*               01:1E:83:00:00:29     b-sdp:62:40              Local    Fwd
-------------------------------------------------------------------------------
Number of entries: 1
===============================================================================
```

The group address terminates in hex 29, which corresponds to ISID 41.

The configured static-isids can be displayed with the following command (a range 41-100 has been added to the sap lag-1:40 to demonstrate this output):

```
*A:PE-5# configure
    service
        vpls 40
            sap lag-1:40 create
                static-isid
                    range 1 create isid 41 to 100
                exit
            exit
```

```
*A:PE-5# show service id 40 sap lag-1:40 static-isids
===================================
Static Isid Entries
===================================
Entry          Range
-----------------------------------
1              41-100
===================================
```

Configuration of ISID-Policies in SPB B-VPLS

ISID policies are an optional aspect of SPBM which allow additional control of the advertisement of ISIDs and creation of MFIB entries for I-VPLS (Epipe services do not trigger ISID advertisements or the creation of MFIB entries). By default, if no ISID-policies are used, SPBM automatically advertises and populates MFIB entries for I-VPLS and static-isids. ISID-policies can be used on any SPB-enabled node with locally defined I-VPLS instances or static-isids. The isid-policy parameters are shown below:

```
A:PE-3# configure service vpls 40 isid-policy entry ?
  - entry <range-entry-id> [create]
  - no entry <range-entry-id>

 <range-entry-id>     : [1..8191]
 <create>             : keyword

 [no] advertise-local - Configure local advertisement of the range
 [no] range           - Configure ISID range for the entry
 [no] use-def-mcast   - Use default multicast tree to propogate ISID range
```

Where:

- **advertise-local** defines whether the local ISIDs (I-VPLS ISIDs linked to the B-VPLS) or static ISIDs contained in the configured range are advertised in SPBM.
- **use-def-mcast** controls whether the ISIDs contained in the range use MFIB entries (if **no use-def-mcast** is used) or just the default tree which does not use any MFIB entry.

The **isid-policy** becomes active as soon as it is defined, as opposed to other policies in SR OS, which require the policy itself to be applied within the configuration.

The typical use of ISID-policies is to reduce the number of ISIDs being advertised and/or to save MFIB space (in deployments where MFIB space is shared with MMRP and IP Multicast). The use of ISID-policies is recommended for I-VPLS where most of the traffic is unicast or for I-VPLS where the ISID end-points are present in all the backbone edge bridges (BEBs) of the SPB network. In both cases, advertising ISIDs or consuming MFIB entries for those I-VPLSs has little value since no multicast (first case) or the default tree (second case) are as efficient as using MFIB entries.

The following configuration example will use the example topology in Figure 135. In this case, the objective of the isid-policy will be to use the default tree for all the I-VPLS services with ISIDs between 41 and 100, excluding the range 80-90. The following example shows the policy configuration in PE-3. The same policy will be configured in the rest of the SPB nodes, that is, PE-2, PE-6, and PE-5.

```
*A:PE-3# configure
    service
        vpls 40
            isid-policy
                entry 10 create
                    range 80 to 90
                exit
                entry 20 create
                    use-def-mcast
                    no advertise-local
                    range 41 to 79
                exit
                entry 30 create
                    use-def-mcast
                    no advertise-local
                    range 91 to 100
                exit
            exit
```

The **no advertise-local** option can only be configured if the **use-def-mcast** option is also configured.

```
*A:PE-3# configure service vpls 40 isid-policy entry 40 create no advertise-local
```

```
MINOR: SVCMGR #7855 Cannot set AdvLocal for entry - advertise-local or use-def-mcast
option must be specified
```

Overlapping ISID values can be configured as long as the actions are consistent for the same ISID. Conflicting actions are shown in the CLI.

```
*A:PE-3# configure service vpls 40 isid-policy entry 40 create
*A:PE-3>config>service>vpls>isid-policy>entry# range 82 to 85
*A:PE-3>config>service>vpls>isid-policy>entry# use-def-mcast
MINOR: SVCMGR #7854 Cannot set UseDefMctree for entry - Conflicting Actions with Entry-
10
```

The isid-policy configured for B-VPLS 40 in all the four nodes makes the SPB network to use the default tree for ISIDs 41-79 and 91-100 and not advertise those ISIDs in SPB ISIS even if the ISID is locally defined (as in the case for ISIDs 41-100 in PE-3). As discussed in Basic SPBM Configuration, the default tree path can be checked from each node by using the **tools dump service id 40 spb default-multicast-list** command.

Due to entry 10 in the policy, ISIDs 80-90 will be advertised by PE-3 (active MC-LAG node). However, nodes PE-2 and PE-6 will not create any MFIB entry for those ISIDs until the corresponding I-VPLS ISIDs are locally created (or configured through static-ISIDs). The following command executed on PE-2 proves that ISIDs 80-90 are indeed being advertised by PE-3:

```
*A:PE-2# show service id 40 spb database detail

================================================================================
Rtr Base ISIS Instance 1027 Database (detail)
================================================================================

Displaying Level 1 database
--------------------------------------------------------------------------------
LSP ID    : PE-2.00-00                                   Level     : L1
---snip---

--------------------------------------------------------------------------------
LSP ID    : PE-3.00-00                                   Level     : L1
---snip---

TLVs :
---snip---
  MT Capability :
    TLV Len         : 56
    MT ID           : 0
    SPBM Service ID:
    Sub TLV Len     : 52
      BMac Addr               : 00:00:00:03:03:03
      Base VID                : 40
      ISIDs                   :
        80       Flags:TR
        81       Flags:TR
        82       Flags:TR
        83       Flags:TR
        84       Flags:TR
        85       Flags:TR
        86       Flags:TR
        87       Flags:TR
        88       Flags:TR
        89       Flags:TR
        90       Flags:TR

-------------------------------------------------------------------------------
LSP ID    : PE-5.00-00                                  Level     : L1
---snip---
===============================================================================
```

The **mfib** parameter in the **show service id 40 sap static-isids mfib** command can help understand the state of the MFIB entries added (or not) by the configured static-isid. The following possible states can be shown:

- If the static-ISID is configured and programmed in the MFIB, the status is shown as:
  - ok
- If the static-ISID is not configured and not programmed in the MFIB, the reasons can be (order of priority):
  - useDefMCTree - ISID policy is applied on the service for the ISID.
  - sysMFibLimit - system MFIB limit has been exceeded
  - addPending - adding pending due to processing delays
- If the static-ISID is not configured, but present in the MFIB:
  - delPending - cleanup pending due to processing delays.

The following output shows some of these possible states:

```
*A:PE-5# show service id 40 sap lag-1:40 static-isids mfib
===================================
ISID Detail
===================================
ISID           Status
-----------------------------------
41             useDefMCTree
42             useDefMCTree
---snip---
80             ok
81             ok
---snip---
===================================
```

Conclusion

SR OS supports an efficient SPBM implementation in the context of a B-VPLS, where system BMACs, vMEP OAM BMACs and SAP-BMACs are advertised in SPB IS-IS. SPBM provides a simple solution where no other control plane protocol is required in the B-VPLS to take care of the resiliency, load-balancing and multicast optimization. The SPBM implementation in the SR OS provides scale optimization through the use of control and user B-VPLSs, allows the interworking between SPB networks and PBB networks, as well as the optimization of the MFIB resources and advertisement of ISIDs through the use of ISID-policies.

Virtual Ethernet Segments

This chapter provides information about Virtual Ethernet Segments.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

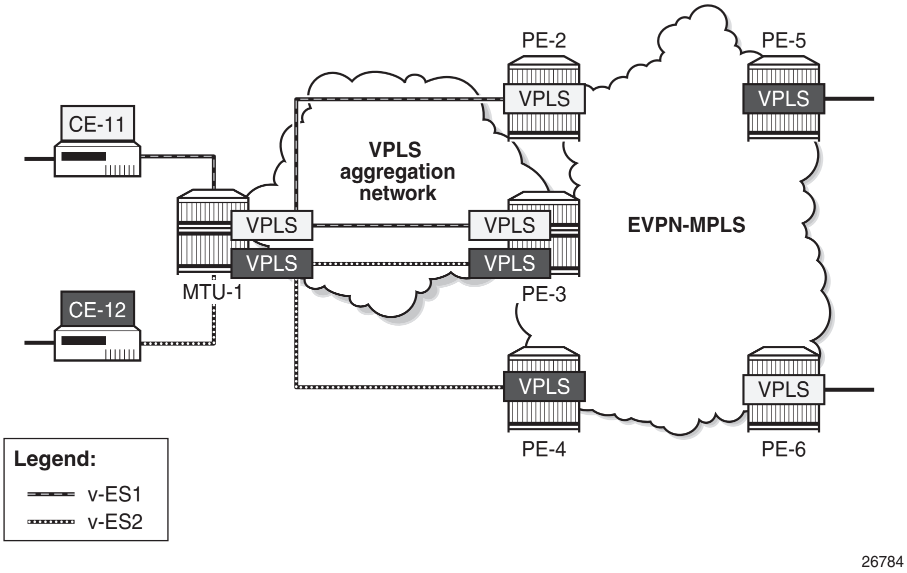
Applicability

The information and configuration in this chapter are based on SR OS Release 15.0.R3. Virtual Ethernet segments are supported in SR OS Release 15.0.R1, and later.

Overview

RFC7432 describes the use and procedures for Ethernet segments (ESs) that can be associated with physical Ethernet ports and LAGs. The SR OS implementation also allows an ES to be associated with SDPs. ESs meet the redundancy requirements of directly connected CEs. However, ESs will not work when an aggregation network exists between CEs and ES PEs, which requires different ESs to be defined for the port, LAG, or SDP. *Draft-sajassi-bess-evpn-virtual-eth-segment* describes how virtual ESs (vESs) can be defined with an Attachment Circuit (AC) level granularity. Figure 136 shows an example where vES definition at the pseudowire (PW) granularity level is required:

Figure 136 **vESs for PWs**



When a Layer 2 aggregation network is used to get access to EVPN, the association of ACs that belong to the same ES and physical ports or SDPs can be arbitrary. For example, the SDP between MTU-1 and PE-3 (Figure 1) cannot be associated with only one ES, because it is being used by two different CEs that require different ESs. The association must be at spoke-SDP level. The RFC7432 port/lag-based ES definition is not sufficient, so vESs need to be defined. Virtual ESs can be configured with up to eight ranges of one or more:

- VC-IDs (spoke-SDPs)
- Q-tags (dot1q)
- S-tags (qinq)
- C-tags for a fixed S-tag (qinq)

Mesh-SDPs are not allowed for an SDP used by a vES.

Virtual ESs are configured as Ethernet segments with the creation-time keyword **virtual**:

```
*A:PE-2# configure service system bgp-evpn ethernet-segment
  - ethernet-segment <name> [create] [virtual]
  - no ethernet-segment <name>

 <name>               : [28 chars max]
 <virtual>            : keyword

      dot1q           + Configure dot1q port or lag information
 [no] es-activation-* - Configure ethernet segment activation timer
 [no] esi             - Configure ethernet segment identifier
 [no] lag             - Configure lag for service BGP EVPN ethernet segment
 [no] multi-homing    - Configure multi-homing for service BGP EVPN ethernet segment
 [no] port            - Configure port for service BGP EVPN ethernet segment
      qinq            + Configure qinq port or lag information
 [no] sdp             - Configure sdp for service BGP EVPN ethernet segment
      service-carving + Configure service carving mode for BGP EVPN ethernet segment
 [no] shutdown        - Enable/disable administrative state of the ethernet segment
 [no] source-bmac-lsb - Configure source  BMAC address LSB information
 [no] vc-id-range     - Configure VC ID range
```

Virtual ES "vESI-23_600" is associated with LAG 1 and one service-delimiting VLAN range is defined for the S-tag, as follows:

```
configure
    service
        system
            bgp-evpn
                ethernet-segment "vESI-23_600" virtual create
                    esi 01:00:00:00:00:23:06:00:00:01
                    es-activation-timer 3
                    service-carving
                        mode manual
                        manual
                            evi 2
                        exit
                    exit
                    multi-homing all-active
                    lag 1
                    qinq
                        s-tag-range 600 to 602
                    exit
                    no shutdown
                exit
```

The configured ES will match all the SAPs for which the top (outer) service-delimiting tag is within the 600 to 602 range.

When the ES is created as virtual, a port, LAG, or SDP needs to be created before any VLAN or VC-ID can be associated.

- For VC-ID, only spoke-SDPs are allowed, no mesh-SDPs. Manual spoke-SDP VC-IDs and BGP-AD VC-IDs can be included in the range.
- For dot1q, only those SAPs that match the service-delimiting VLAN range will be associated with the vES
- For qinq, the following two commands can be configured, with a mutually exclusive S-tag:
  - **s-tag-range <qtag1> to <qtag1>** - associates all qinq SAPs with outer tag between the configured qtags.

- **s-tag <qtag1> c-tag-range <qtag2> to <qtag2>** - associates all qinq SAPs with outer qtag1 and inner qtag between the configured qtag2 values to the vES

A mutually exclusive S-tag means that a value for the S-tag can be configured in either of the two commands, but not in both.

Table 11 shows the supported examples for qtag values between 1 and 4094; Table 12 shows the supported examples for qtag values 0, *, and null:

Table 11 **Supported Examples for Q-tag Values between 1 and 4094**

vES configuration for port 1/1/1	SAP association
dot1q qtag-range 100	1/1/1:100
dot1q qtag-range 100 to 102	1/1/1:100, 1/1/1:101, 1/1/1:102
qinq s-tag 100 c-tag-range 200	1/1/1:100.200
qinq s-tag 100 c-tag-range 200 to 202	1/1/1:100.200, 1/1/1:100.201, 1/1/1:100.202
qinq s-tag-range 100	All SAPs 1/1/1:100.x (x being 1 to 4094, 0, or *)
qinq s-tag-range 100 to 102	All SAPs 1/1/1:100.x, 1/1/1:101.x, 1/1/1:102.x (x being 1 to 4094, 0, or *)

Table 12 **Supported Examples for Q-tag Values 0, *, and null**

vES configuration for port 1/1/1	sap association
dot1q qtag-range 0	1/1/1:0
dot1q qtag-range *	1/1/1:*
qinq s-tag 0 c-tag-range *	1/1/1:0.*
qinq s-tag * c-tag-range *	1/1/1:*.*
qinq s-tag * c-tag-range null	1/1/1:*.null

Considerations:

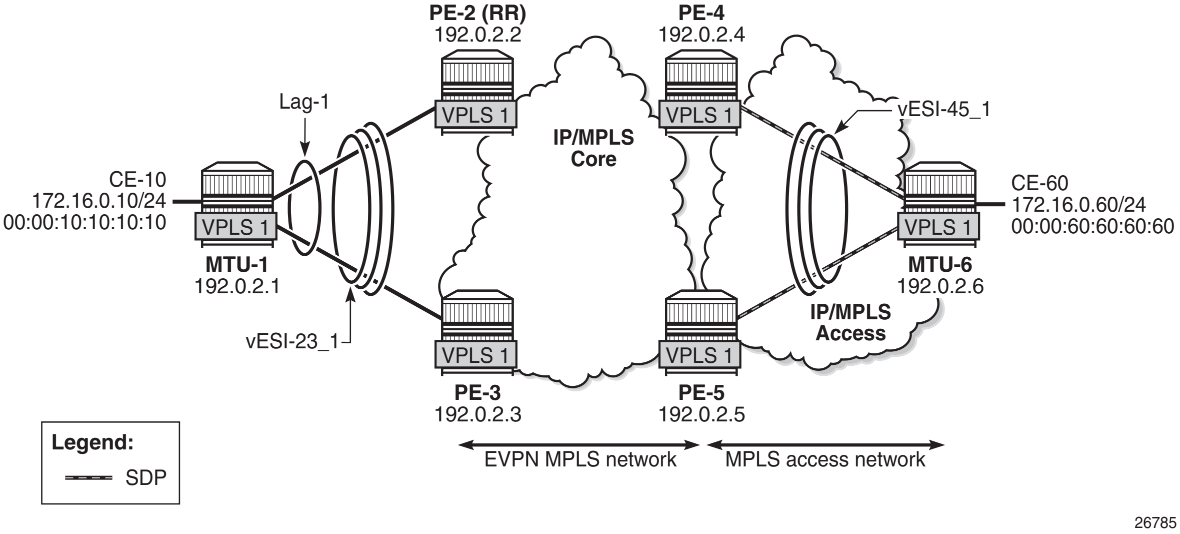
- The ranges can be modified on the fly: qtag-range, s-tag/c-tag-range, vc-id-range.
- For port-based vESs, PXC sub-ports are supported. For more information about PXC, see chapter *Port Cross-Connect (PXC)*.

- Virtual ESs are supported in EVPN-MPLS, PBB-EVPN, and EVPN-VPWS
- Virtual ESs are supported in single-active and all-active EVPN multi-homing
  - Two all-active vESs must use different ES-BMACs, even if they are defined in the same LAG.
- Virtual ESs implement CMAC flush procedures described in RFC7623. Optionally, ISID-based CMAC-flush can be used where the single-active vES does not use ES-BMAC allocation. See chapter PBB-EVPN ISID-based CMAC Flush.
- Connection-profile-vlan SAPs (CP-SAPs) cannot be associated with a vES and cannot be configured on ports where vESs are defined. For more information about CP-SAPs, see chapter VLAN Range SAPs for VPLS and Epipe Services.

Configuration

Figure 137 shows the example topology with four core PEs in an EVPN-MPLS network and two MTUs. VPLS 1 is configured in all the nodes. EVPN is configured on the core PEs, not on the MTUs. LAG 1 is configured on MTU-1, PE-2, and PE-3 and associated with an all-active vES "ESI-23_1" on PE-2 and PE-3. A single-active vES "ESI-45_1" is configured on PE-4 and PE-5, associated with SDPs.

Figure 137 **Example Topology**



The configuration is similar to the one in chapter EVPN for MPLS Tunnels, where the parameters are described in detail.

The initial configuration on the nodes includes the following:

- Cards, MDAs, ports
- Router interfaces
- IS-IS (alternatively, OSPF can be configured)
- LDP in the IP/MPLS core and IP/MPLS access network

LAG 1 is configured with qinq encapsulation. The LAG configuration on PE-1 is as follows:

```
configure
    lag 1
        mode access
        encap-type qinq
        port 1/1/1
        port 1/1/2
        lacp active administrative-key 32768
        no shutdown
```

BGP is configured on all PEs for address family EVPN. PE-2 is the Route Reflector (RR) and is configured as follows.

```
configure
    router
        autonomous-system 64500
        bgp
            vpn-apply-import
            vpn-apply-export
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            split-horizon
            rapid-update evpn
            group "internal"
                family evpn
                cluster 1.1.1.1
                peer-as 64500
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
                neighbor 192.0.2.5
                exit
            exit
```

VPLS 1 is configured on all nodes. On the PEs, BGP-EVPN is enabled for MPLS. The following is configured on PE-2:

```
configure
    service
        vpls 1 customer 1 create
            bgp
            exit
```

```
                bgp-evpn
                    evi 1
                    mpls
                        ingress-replication-bum-label
                        ecmp 2
                        auto-bind-tunnel
                            resolution any
                        exit
                        no shutdown
                    exit
                exit
                sap lag-1:1.1 create
                exit
                no shutdown
            exit
```

The configuration on the other PEs is similar, but on PE-4 and PE-5, a spoke-SDP is configured instead of a SAP. The service configuration on PE-4 is as follows:

```
configure
    service
        sdp 46 mpls create
            far-end 192.0.2.6
            ldp
            no shutdown
        exit
        vpls 1 customer 1 create
            bgp
            exit
            bgp-evpn
                evi 1
                mpls
                    ingress-replication-bum-label
                    ecmp 2
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            spoke-sdp 46:1 create
            exit
            no shutdown
        exit
```

Virtual ESs must be created with the **virtual** keyword; if not, the following error is raised after an attempt to define a range:

```
*A:PE-2>config>service>system>bgp-evpn>eth-seg>qinq#  s-tag-range 1
MINOR: SVCMGR #8064 Cannot create range - ethernet-segment is not virtual
```

On PE-2 and PE-3, the two following two all-active multi-homing vESs are created, each with a unique ESI:

```
configure
    service
```

```
system
    bgp-evpn
        ethernet-segment "vESI-23_1" virtual create
            esi 01:00:00:00:00:23:01:00:00:01
            es-activation-timer 3
            service-carving
                mode auto
            exit
            multi-homing all-active
            lag 1
            qinq
                s-tag-range 1
                s-tag-range 500 to 501
                s-tag 495 c-tag-range 100 to 102
            exit
            no shutdown
        exit
        ethernet-segment "vESI-23_600" virtual create
            esi 01:00:00:00:00:23:06:00:00:01
            es-activation-timer 3
            service-carving
                mode manual
                manual
                    evi 2
                exit
            exit
            multi-homing all-active
            lag 1
            qinq
                s-tag-range 600 to 602
            exit
            no shutdown
        exit
```

When attempting to configure another vES with the ESI of an existing ES/vES, the following error is raised:

```
*A:PE-2>config>service>system>bgp-evpn# ethernet-segment "vESI-
23_610" virtual create
*A:PE-2>config>service>system>bgp-evpn>eth-seg# esi 01:00:00:00:00:23:06:00:00:01
MINOR: SVCMGR #8047 Ethernet segment id is not valid -
 ESI already in use by another ethernet segment
```

Multiple vESs can be defined on the same LAG. However, the ranges should not overlap. The following error is raised after attempting to configure an additional range in vES "ESI-23_600" that uses S-tag 600 in combination with a range of C-tags. S-tag 600 is already included in the first range: **s-tag-range 600 to 602**. The error message points out that this range is of a different type: the existing range defines only S-tags, whereas the new range defines a range of C-tags for S-tag 600.

```
*A:PE-2>config>service>system>bgp-evpn>eth-seg>qinq# s-tag 600 c-tag-
range 100 to 111
MINOR: SVCMGR #8064 Cannot create range -
 range overlaps with existing range of a different type
```

When attempting to define **s-tag-range 1** in "vESI-23_2", when S-tag 1 is already defined in "vESI-23_1", the following error is raised:

```
*A:PE-2>config>service>system>bgp-evpn>eth-seg>qinq# s-tag-range 1
MINOR: SVCMGR #8064 Cannot create range -
 range overlaps with existing range in ethernet-segment vESI-23_1
```

On PE-4, the following single-active multi-homing vESs are configured. The configuration on PE-5 contains a different SDP.

```
configure
    service
        system
            bgp-evpn
                ethernet-segment "vESI-45_1" virtual create
                    esi 01:00:00:00:00:45:01:00:00:01
                    es-activation-timer 3
                    service-carving
                        mode auto
                    exit
                    multi-homing single-active
                    sdp 46
                    vc-id-range 1
                    vc-id-range 500 to 501
                    no shutdown
                exit
                ethernet-segment "vESI-45_2" virtual create
                    esi 01:00:00:00:00:45:02:00:00:01
                    es-activation-timer 3
                    service-carving
                        mode manual
                        manual
                            evi 2
                        exit
                    exit
                    multi-homing single-active
                    sdp 46
                    vc-id-range 2
                    no shutdown
                exit
```

The configured ESs and vESs can be retrieved as follows:

```
*A:PE-2# show service system bgp-evpn ethernet-segment

===============================================================================
Service Ethernet Segment
===============================================================================
Name                             ESI                           Admin     Oper
-------------------------------------------------------------------------------
vESI-23_1                        01:00:00:00:00:23:01:00:00:01 Enabled   Up
vESI-23_600                      01:00:00:00:00:23:06:00:00:01 Enabled   Up
-------------------------------------------------------------------------------
Entries found: 2
===============================================================================
*A:PE-2#
```

The following information for the first entry in the list shows that it is a virtual ES.

```
*A:PE-2# show service system bgp-evpn ethernet-segment name "vESI-23_1"

===============================================================================
Service Ethernet Segment
===============================================================================
Name                    : vESI-23_1
Eth Seg Type            : Virtual
Admin State             : Enabled            Oper State         : Up
ESI                     : 01:00:00:00:00:23:01:00:00:01
Multi-homing            : allActive          Oper Multi-homing  : allActive
ES SHG Label            : 262136
Source BMAC LSB         : <none>
Lag Id                  : 1
ES Activation Timer     : 3 secs
Svc Carving             : auto               Oper Svc Carving   : auto
Cfg Range Type          : primary
===============================================================================
*A:PE-2#
```

Virtual ES "vESI-23_1" on PE-2 has the following S-tag ranges and S/C-tag ranges:

```
*A:PE-2# show service system bgp-evpn ethernet-segment name "vESI-23_1" virtual-
ranges

===============================================================================
Q-Tag Ranges
===============================================================================
Q-Tag Start         Q-Tag End           Last Changed
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
No entries found
===============================================================================
===============================================================================
VC-Id Ranges
===============================================================================
VC-Id Start         VC-Id End           Last Changed
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
No entries found
===============================================================================
===============================================================================
S-Tag Ranges
===============================================================================
S-Tag Start         S-Tag End           Last Changed
-------------------------------------------------------------------------------
1                   1                   06/28/2017 12:15:41
500                 501                 06/28/2017 12:15:41
-------------------------------------------------------------------------------
Number of Entries: 2
===============================================================================
===============================================================================
S-Tag C-Tag Ranges
===============================================================================
S-Tag Start         C-Tag Start         C-Tag End       Last Changed
-------------------------------------------------------------------------------
495                 100                 102             06/28/2017 12:17:35
```

```
-------------------------------------------------------------------------------
Number of Entries: 1
===============================================================================
*A:PE-2#
```

The ranges in the vES can be modified while the vES is operationally up, for example, an S-tag range can be added as follows:

```
*A:PE-2# configure service system bgp-evpn ethernet-segment "vESI-23_1" qinq s-tag-
range 10
```

The S-tag ranges can be verified with the following command. Compared with the preceding output, the S-tag 10 has been added:

```
*A:PE-2# show service system bgp-evpn ethernet-segment name "vESI-23_1" virtual-
ranges | match S-Tag post-lines 10
S-Tag Ranges
===============================================================================
S-Tag Start         S-Tag End           Last Changed
-------------------------------------------------------------------------------
1                   1                   06/28/2017 12:15:41
10                  10                  06/28/2017 12:34:04
500                 501                 06/28/2017 12:15:41
-------------------------------------------------------------------------------
Number of Entries: 3
===============================================================================
===============================================================================
S-Tag C-Tag Ranges
===============================================================================
S-Tag Start         C-Tag Start         C-Tag End       Last Changed
-------------------------------------------------------------------------------
495                 100                 102             06/28/2017 12:17:35
-------------------------------------------------------------------------------
Number of Entries: 1
===============================================================================
*A:PE-2#
```

On PE-4, the same **show** command shows the range of VC-IDs, as follows:

```
*A:PE-4# show service system bgp-evpn ethernet-segment name "vESI-45_1" virtual-
ranges
===============================================================================
Q-Tag Ranges
===============================================================================
Q-Tag Start         Q-Tag End           Last Changed
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
No entries found
===============================================================================
===============================================================================
VC-Id Ranges
===============================================================================
VC-Id Start         VC-Id End           Last Changed
-------------------------------------------------------------------------------
1                   1                   06/28/2017 12:15:04
500                 501                 06/28/2017 12:15:04
```

```
-------------------------------------------------------------------------------
Number of Entries: 2
===============================================================================
===============================================================================
S-Tag Ranges
===============================================================================
S-Tag Start         S-Tag End           Last Changed
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
No entries found
===============================================================================
===============================================================================
S-Tag C-Tag Ranges
===============================================================================
S-Tag Start         C-Tag Start         C-Tag End       Last Changed
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
No entries found
===============================================================================
*A:PE-4#
```

Connection-profile-vlan SAPs (CP-SAPs) cannot be associated with a vES and cannot be configured on ports where vESs are defined. CP-SAP 10 is created on PE-3, as follows:

```
configure
    connection-profile-vlan 10 create
        vlan-range 5 to 100
        vlan-range 495
    exit
```

The following vES is configured on PE-3:

```
configure
    service
        system
            bgp-evpn
                ethernet-segment "vESI-23_10" virtual create
                    esi 01:00:00:00:00:23:10:00:00:01
                    es-activation-timer 3
                    service-carving
                        mode auto
                    exit
                    multi-homing single-active
                    port 1/2/3
                    qinq
                        s-tag-range 100
                    exit
                    no shutdown
                exit
```

This vES can only be configured when no CP-SAPs are defined on port 1/2/3. The following error message is raised when a CP-SAP is configured on port 1/2/3 already and the vES is configured afterward:

```
*A:PE-3>config>service>system>bgp-evpn>eth-seg#  port 1/2/3
MINOR: SVCMGR #8048 Ethernet segment access port/lag/sdp is not valid -
 unsupported sap type configured on port in service:1
```

When attempting to configure CP-SAP 1/2/3:cp-10 in VPLS 1 with port 1/2/3 associated with a vES, the following error message is raised.

```
*A:PE-3# configure service vpls 1 sap 1/2/3:100.cp-10 create
MINOR: SVCMGR #6044 Cannot create sap -
 sap type not allowed when port is associated with virtual ethernet-segment
```

Conclusion

Regular ESs and vESs can be associated with ports, LAGs, and SDPs; in case of vES, ranges of Q-tags, S-tags, C-tags, or VC-IDs can be defined. The granularity for vES is per AC. Multiple vESs with different ESIs can be defined on the same port, LAG, or SDP.

VLAN Range SAPs for VPLS and Epipe Services

This chapter provides information about VLAN Range SAPs for VPLS and Epipe Services.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SR OS Release 14.0.R6, but the CLI in the current edition is based on SR OS Release 15.0.R2. Connection-Profile VLAN SAPs (CP SAPs) are supported in SR OS Release 14.0.R1, and later.

Overview

Backhaul services through metro Ethernet networks require bundled interface support. In SR OS terminology, bundling refers to Connection-Profile VLAN SAPs (CP SAPs)—special SAPs that capture the traffic of a range of CE VLAN IDs (VIDs) entering an Ethernet port. CP SAPs are fully compatible with Metro Ethernet Forum (MEF) 10.3 bundling service attributes and RFC 7432 EVPN VLAN bundle service interfaces. CP SAPs are supported in Layer 2 services only, and can be configured together with other SAPs and/or SDP-bindings.

For frames with an ingress VID contained in the range configured in the SAP's CP, the behavior is similar to default SAPs, such as 1/1/1:*, where "*" spans the entire VID range from 0 to 4095 and serves as a wildcard. However, unlike a default SAP, a CP SAP cannot co-exist with a VLAN SAP that is in the same range and on the same port or LAG. For example, 1/1/1:* and 1/1/1:100 can co-exist whereas 1/1/2:cp-1 (where cp-1 corresponds to the VLAN range from 1 to 200) and 1/1/2:100 cannot co-exist.

The VLAN manipulation between VLAN SAPs, default SAPs, and CP SAPs is compared in Table 13.

Table 13 **VLAN Manipulation in SAPs**

	VLAN SAP	Default SAP	CP SAP
Service-delimiting VLAN	Yes For example: VLAN 100 in 1/1/1:100	No	No
Push/pop VLAN tags in egress/ingress frames	Yes	No	No
VLAN translation	Yes	No	No

Figure 138 shows how dot1q VLAN SAPs pop the customer VLAN tag in ingress frames and push the VLAN tag in egress frames. Therefore, frames are untagged between PE-1 and PE-2. VLAN translation is possible when the VIDs in the VLAN tags that are popped or pushed at the SAPs are different at ingress and egress, as follows.

Figure 138 **Customer VID is Popped and Pushed by VLAN SAPs - VLAN Translation**

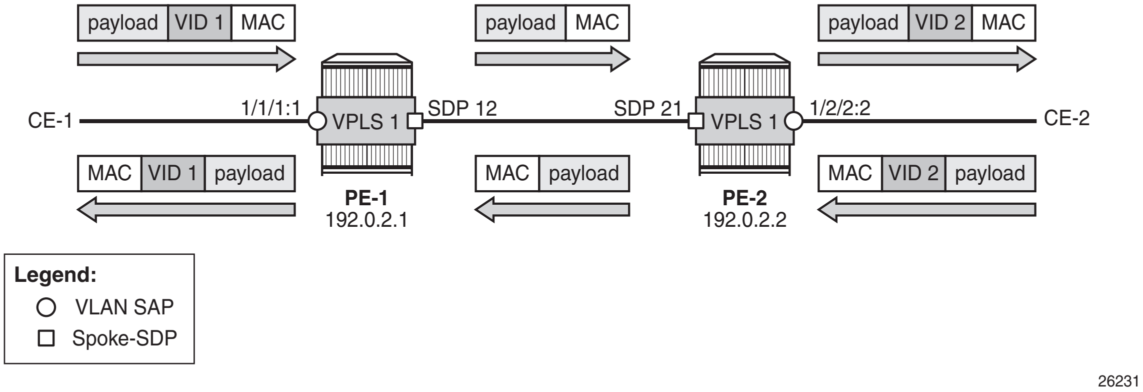


Figure 139 shows that dot1q CP SAPs do not pop or push the CE VID. Frames keep the same tag end-to-end; therefore, VLAN translation is not possible.

Figure 139 **Customer VID is Preserved between Dot1q CP SAPs - No VLAN Translation**

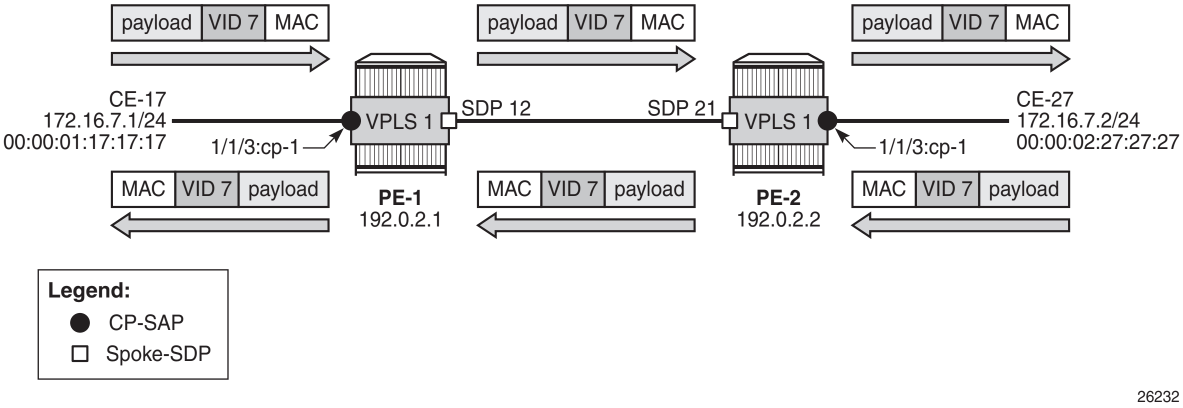
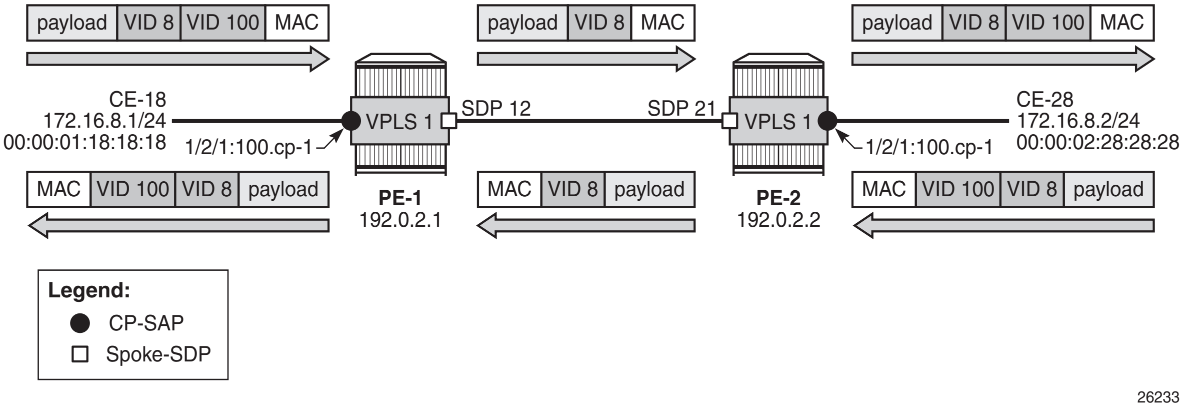


Figure 140 shows that QinQ CP SAPs only pop or push the service delimiting VID (VID 100), but not the customer VID in the CP range, as follows:

Figure 140 **Customer VID is Preserved between QinQ CP SAPs - No VLAN Translation**



VID 100 is service delimiting and can be different in both SAPs, but the customer VID in the VLAN range of the CP is not.

Connection Profile VLAN

Note: The **connection-profile-vlan** context is different from the connection-profile used for ATM connectivity.

CP SAPs refer to connection profiles (connection-profile-vlan) that can contain up to 32 ranges of customer VIDs. Connection profiles are configured with the following command:

```
*A:PE-1# configure connection-profile-vlan
  - connection-profile-vlan <conn-prof-id> [create]
  - no connection-profile-vlan <conn-prof-id>

 <conn-prof-id>       : [1..8000]

 [no] description     - Configure a connection profile VLAN description
 [no] vlan-range      - Configure a connection profile vlan range
```

VLAN ranges in a CP contain one or more consecutive VIDs, as follows:

```
*A:PE-1# configure connection-profile-vlan 10 vlan-range
  - no vlan-range <from>
  - vlan-range <from> [to <to>]

 <from>               : [1..4094]
 <to>                 : [1..4094]
```

Following is an example of a CP configuration containing three non-overlapping VLAN ranges:

```
configure
    connection-profile-vlan 10 create
        vlan-range 5 to 100
        vlan-range 150 to 300
        vlan-range 350
    exit
exit
```

Overlapping ranges are not allowed within the same CP, as follows:

```
*A:PE-1# configure connection-profile-vlan 10 vlan-range 7 to 9
MINOR: ATM #2304 Overlapping range
```

New non-overlapping VLAN ranges can be added to the CP defined in an existing and operationally up SAP. The CP's VLAN ranges can also be removed on the fly. When a user wants to extend a VLAN range, for example, VLAN range 350 becoming a range from 350 to 400, the existing VLAN range is not overwritten and a message is raised indicating that the VLAN ranges overlap, as follows:

```
*A:PE-1# configure connection-profile-vlan 10 vlan-range 350 to 400
MINOR: ATM #2304 Overlapping range
```

The existing VLAN range of 350 can be preserved when the CP SAP is operational and a new VLAN range from 351 to 400 can be added, as follows:

```
*A:PE-1# configure connection-profile-vlan 10 vlan-range 351 to 400
```

The following example shows four VLAN ranges in CP 10, with a timestamp of the last change for each VLAN range:

```
*A:PE-1# show connection-profile-vlan 10

===============================================================================
Connection Profile 10 Information
===============================================================================
Description : (Not Specified)
Last Change : 05/15/2017 08:16:52

===============================================================================
Connection Profile Vlan Eth Information
===============================================================================
Range Start         Range End                     Last Change
-------------------------------------------------------------------------------
5                   100                           05/15/2017 09:15:20
150                 300                           05/15/2017 09:15:20
350                 350                           05/15/2017 09:15:20
351                 400                           05/15/2017 09:15:25
===============================================================================
===============================================================================
*A:PE-1#
```

If a VLAN tag combination matches different SAPs, the highest priority SAP will be picked regardless of the operational status. For completeness, the following two tables show the SAP lookup matching order for dot1q and QinQ ports.

Table 14 **SAP Lookup Order for Dot1q Ports**

Incoming frame qtag VID value	**SAP lookup precedence order (:0 and :* are mutually exclusive on the same port)**			
	:X	**:CP**	**:0**	**:***
X (belongs to the CP range)	1st	1st		2nd
0			1st	1st
<untagged>			1st	1st

Table 15 **SAP Lookup Order for QinQ Ports**

Incoming frame qtag1.qtag2	**System/port settings = new-qinq-untagged-sap SAP lookup precedence order (assumption: X and Y are defined in CP ranges)**							
	:X.Y	**:X.0**	**:X.CP**	**:CP.***	**:X.***	**:0.***	**:*.null**	**:*.***
X.Y	1st		1st	2nd	2nd			3rd
X.0		1st		2nd	2nd			3rd

Table 15 **SAP Lookup Order for QinQ Ports (Continued)**

Incoming frame qtag1.qtag2	System/port settings = new-qinq-untagged-sap SAP lookup precedence order (assumption: X and Y are defined in CP ranges)							
	:X.Y	:X.0	:X.CP	:CP.*	:X.*	:0.*	:*.null	:*.*
0.Y						1st		2nd
0.0						1st		2nd
X		1st		2nd	2nd		3rd	4th
0						1st	2nd	3rd
<untagged>						1st	2nd	3rd

For example, ingress frames with VIDs 100.20 are classified as part of CP SAP 1/2/1:100.cp-10, not of CP SAP 1/2/3:cp-10.*. Only when SAP 1/2/1:100.cp-10 is removed from the configuration, frames with VIDs 100.20 will go to SAP 1/2/3:cp-10.*.

CPs can be created without any VLAN range, but such empty CPs cannot be used in a CP SAP, as follows:

```
*A:PE-1# configure connection-profile-vlan 11 create

*A:PE-1# show connection-profile-vlan

===============================================================================
Connection Profile Vlan Summary Information
===============================================================================
CP Index                                Number of Members
-------------------------------------------------------------------------------
---snip---
11                                      0
===============================================================================

*A:PE-1# show connection-profile-vlan 11

===============================================================================
Connection Profile 11 Information
===============================================================================
Description : (Not Specified)
Last Change : 05/15/2017 07:32:31

===============================================================================
Connection Profile Vlan Eth Information
===============================================================================
Range Start         Range End                     Last Change
-------------------------------------------------------------------------------
===============================================================================
Connection Profile Vlan Eth members do not exist.
===============================================================================
*A:PE-1#

*A:PE-1# configure service vpls 1 sap 1/1/3:cp-11 create
MINOR: SVCMGR #1611 Invalid encapsulation value for the port's encapsulation type -
 connection profile must have at least one range
```

Assign CP SAPs to VPLS or Epipe Services

Like ordinary SAPs, CP SAPs can be assigned to VPLS or Epipe services, as follows. The VPLS and Epipe can be EVPN services or not. In the following example, VPLS 1 has BGP-EVPN enabled, whereas Epipe 2 does not:

```
configure
    service
        sdp 12 mpls create
            far-end 192.0.2.2
            ldp
            no shutdown
        exit
        vpls 1 customer 1 create
            bgp
            exit
            bgp-evpn
                evi 1
                mpls
                    ingress-replication-bum-label
                    auto-bind-tunnel
                        resolution any
                    exit
                    no shutdown
                exit
            exit
            sap 1/1/3:cp-10 create
            exit
            sap 1/2/1:1.11 create
            exit
            sap 1/2/1:100.cp-10 create
            exit
            sap 1/2/3:cp-10.* create
            exit
            no shutdown
        exit
        epipe 2 customer 1 create
            sap 1/2/1:200.cp-10 create
            exit
            spoke-sdp 12:2 create
            exit
            no shutdown
        exit
```

CP SAPs are configured in the same way as VLAN SAPs and default SAPs, with the following restrictions:

- A CP can be defined for inner or outer tags as shown in the preceding configuration, but not both at the same time, as follows:

```
*A:PE-1# configure service vpls 1 sap 1/2/1:cp-3.cp-10 create
MINOR: CLI SAP-id has an invalid port number or encapsulation value.
```

- If a CP is defined for the outer VID, the inner VID cannot be a specific VID. The inner VID can only be a "*" (where the inner tag can have any value) or a "0" (where the inner tag can be 0 or null), as follows:

```
*A:PE-1# configure service vpls 1 sap 1/2/1:cp-3.4 create
MINOR: CLI SAP-id has an invalid port number or encapsulation value.


*A:PE-1# configure service vpls 1 sap 1/2/3:cp-1.* create
*A:PE-1>config>service>vpls>sap$ exit


*A:PE-1# configure service vpls 1 sap 1/2/1:cp-3.0 create
*A:PE-1>config>service>vpls>sap$ exit
```

- A CP must have at least one VLAN range before it can be associated with a CP SAP, as follows:

```
*A:PE-1# configure service vpls 1 sap 1/1/3:cp-11 create
MINOR: SVCMGR #1611 Invalid encapsulation value for the port's encapsulation type -
 connection profile must have at least one range
```

- No VLAN SAP can be added on a port in dot1q (or a combination of port and service-delimiting VLAN in case of QinQ) when the VLAN is included in the VLAN range in a CP SAP on the same port. One of the VLAN ranges in CP 10 contains all VIDs from 5 to 100. Therefore, it is not allowed to configure a VLAN SAP with VID 100 on port 1/1/3, where a CP SAP is configured with CP 10, as follows:

```
*A:PE-1# configure service vpls 1 sap 1/1/3:100 create
MINOR: SVCMGR #1602 The SAP-id is already in use - 1/1/3:100 is already configured
```

- No CP SAPs can be added with overlapping VLAN ranges on the same port for dot1q (or on the same port- and service-delimiting tag for QinQ), as follows. CP 1 contains VLAN range from 7 to 9, which overlaps with VLAN range from 5 to 100 in CP 10.

```
*A:PE-1# configure connection-profile-vlan 1 create vlan-range 7 to 9
*A:PE-1# configure service vpls 1 sap 1/1/3:cp-1 create
MINOR: SVCMGR #1602 The SAP-id is already in use - 1/1/3:cp-10 is already configured
*A:PE-1# configure service vpls 1 sap 1/2/1:100.cp-1 create
MINOR: SVCMGR #1602 The SAP-id is already in use - 1/2/1:100.cp-10 is already
configured
```

However, the CP can be referred to by SAPs on other ports for dot1q or for QinQ on other combinations of port and service-delimiting VLAN, as follows:

```
*A:PE-1# configure service vpls 1 sap 1/2/1:101.cp-1 create
*A:PE-1>config>service>vpls>sap$ exit
```

- CP SAPs can be added when they contain non-overlapping VLAN ranges on the same port, as follows. CP 3 contains one VLAN range with only one VID: 3. This VLAN range (3) does not overlap with any VLAN range in the CP SAPs assigned to VPLS 1.

```
*A:PE-1# configure connection-profile-vlan 3 create vlan-range 3
*A:PE-1# configure service vpls 1 sap 1/1/3:cp-3 create
*A:PE-1>config>service>vpls>sap# exit
*A:PE-1# configure service vpls 1 sap 1/2/1:100.cp-3 create
*A:PE-1>config>service>vpls>sap# exit
```

- It is possible to modify a CP that is associated with a SAP that is operationally up, but not possible to remove the last VLAN range from the CP, because a CP SAP must always have at least one VLAN range. CP 10 has four VLAN ranges and the system allows that all but one of these ranges can be removed when the CP is associated with CP SAPs. An error message is raised when a user attempts to remove the last VLAN range, as follows:

```
*A:PE-1# configure connection-profile-vlan 10
*A:PE-1>config>connprofvlan# info
----------------------------------------------
        vlan-range 5 to 100
        vlan-range 150 to 300
        vlan-range 350
        vlan-range 351 to 400
----------------------------------------------
*A:PE-1>config>connprofvlan# no vlan-range 5
*A:PE-1>config>connprofvlan# no vlan-range 150
*A:PE-1>config>connprofvlan# no vlan-range 350
*A:PE-1>config>connprofvlan# no vlan-range 351
MINOR: ATM #2306 Range SAP exists for this connection-profile-vlan id - Num saps = 3
```

VPLS 1 contains the following seven SAPs. There is no overlap between the VLAN ranges on a port (or port and service-delimiting tag for QinQ).

```
*A:PE-1# show service id 1 sap

===============================================================================
SAP(Summary), Service 1
===============================================================================
PortId                          SvcId      Ing.  Ing.    Egr.  Egr.   Adm  Opr
                                           QoS   Fltr    QoS   Fltr
-------------------------------------------------------------------------------
1/1/3:cp-3                      1          1     none    1     none   Up   Up
1/1/3:cp-10                     1          1     none    1     none   Up   Up
1/2/1:1.11                      1          1     none    1     none   Up   Up
1/2/1:101.cp-1                  1          1     none    1     none   Up   Up
1/2/1:100.cp-3                  1          1     none    1     none   Up   Up
1/2/1:100.cp-10                 1          1     none    1     none   Up   Up
1/2/3:cp-10.*                   1          1     none    1     none   Up   Up
-------------------------------------------------------------------------------
Number of SAPs : 7
```

Constraints to be considered when applying CP SAPs in Layer 2 services are described in the Release Notes, section "Known Limitations" - "Services General".

Consumed Resources for CP SAPs

Regular and default SAPs consume one SAP instance each, whereas CP SAPs consume a number of SAP instances equal to the number of VLANs in the range. The following shows that there are eight SAP entries (in this example, seven SAPs in VPLS 1 and one SAP in Epipe 2), which can be regular, default, or CP SAP entries:

```
*A:PE-1# tools dump resource-usage system

===============================================================================
Resource Usage Information for System
===============================================================================
                                                  Total   Allocated        Free
-------------------------------------------------------------------------------
                   SAP Ingress QoS Policies        3071           1        3070
                    SAP Egress QoS Policies        3071           1        3070
              Ingress Queue-Group Templates        2047           3        2044
               Egress Queue-Group Templates        2047           4        2043
          Egress Port Queue-Group Instances       40959           8       40951
           Ingress FP Queue-Group Instances       16383           0       16383
                          Egress Port VPort       40959           0       40959
          Dynamic Services Next-Hop Entries      511999           0      511999
                     IPSec Next-Hop Entries      262143           0      262143
                Subscriber Next-Hop Entries      199999           0      199999
                                SAP Entries      262143           8      262135
===============================================================================
*A:PE-1#
```

However, the number of SAP instances consumed for card 1 FP 1 exceeds the number of SAP entries in the system, as follows:

```
*A:PE-1# tools dump resource-usage card 1 fp 1

===============================================================================
Resource Usage Information for Card Slot #1 FP #1
===============================================================================
                                                  Total   Allocated        Free
-------------------------------------------------------------------------------
---snip---
                              SAP Instances       63999        1198       62801

===============================================================================
*A:PE-1#
```

The calculation of the number of SAP instances is as follows. In this example, CP 10 is used in four SAPs (three in VPLS 1 and one in Epipe 2) and contains the following VLAN ranges:

```
*A:PE-1# show connection-profile-vlan 10

===============================================================================
Connection Profile 10 Information
===============================================================================
Description : (Not Specified)
```

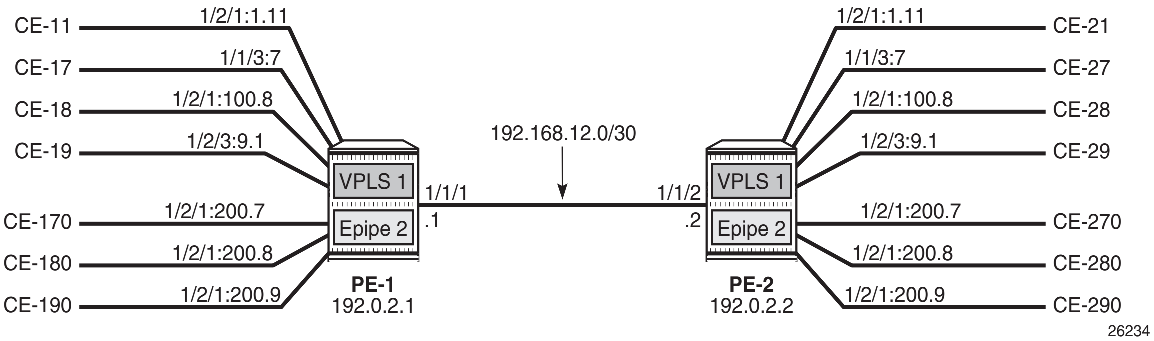
```
Last Change : 05/15/2017 08:16:52
===============================================================================
Connection Profile Vlan Eth Information
===============================================================================
Range Start         Range End                      Last Change
-------------------------------------------------------------------------------
5                   100                            05/15/2017 09:15:20
150                 300                            05/15/2017 09:15:20
350                 350                            05/15/2017 09:15:20
351                 400                            05/15/2017 09:15:25
===============================================================================
===============================================================================
*A:PE-1#
```

The number of VLANs in the VLAN ranges of CP 10 equals 298. For each of the four SAP entries with CP 10, 298 SAP instances are used, for a total of 1192. As well, there is one CP SAP using CP 1 with three VLANs in the VLAN range from 7 to 9 (for three more SAP instances). Two CP SAPs use CP 3 with only VID 3 in the VLAN range (for two more SAP instances), and one SAP is a regular SAP that consumes one SAP instance. Therefore, the total number of SAP instances is 1198.

Configuration

Figure 141 shows the example topology used in this chapter.

Figure 141 **Example Topology**



The initial configuration on the PEs includes the following:

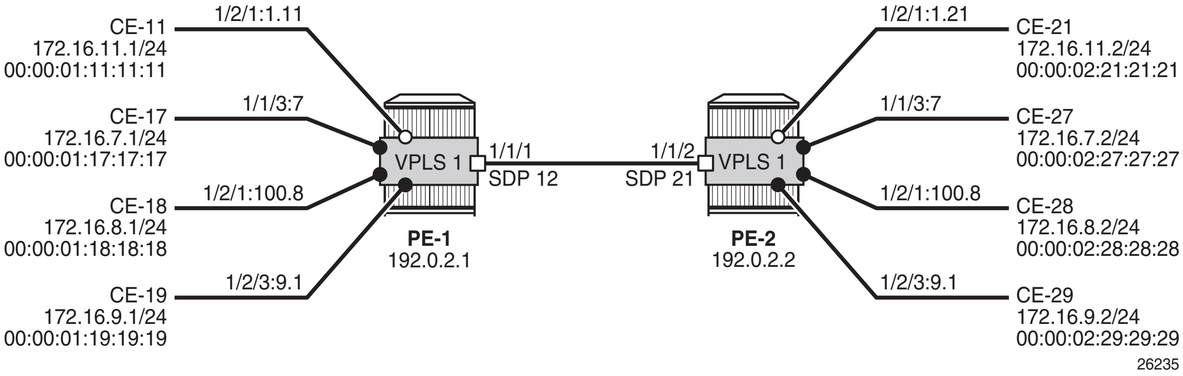
- Cards, MDAs, ports
- Router interfaces
- IS-IS (or OSPF) between the PEs
- LDP between the PEs

In this example, no BGP is configured and no BGP-EVPN will be configured in the VPLS and Epipe services. However, VLAN ranges can be applied in EVPN VPLS and EVPN Epipe services.

VLAN Ranges in VPLS Services

Figure 142 shows the example topology for VPLS 1 with a combination on VLAN SAPs and CP SAPs. The port:VID represents the port to which the CE is connected and the VID sent by the CE; for example, CE-17 is connected to port 1/1/3 on PE-1 and sends frames with VID 7. When VLAN ranges are used, the port:VID 1/1/3:7 does not represent the configured SAP, which is 1/1/3:cp-1.

Figure 142 **Example Topology for VLAN Ranges in VPLS 1**



The service configuration for VPLS 1 on PE-1 is as follows:

```
configure
    service
        sdp 12 mpls create
            far-end 192.0.2.2
            ldp
            no shutdown
        exit
        vpls 1 customer 1 create
            sap 1/1/3:cp-1 create
            exit
            sap 1/2/1:1.11 create
            exit
            sap 1/2/1:100.cp-1 create
            exit
            sap 1/2/3:cp-1.* create
            exit
            spoke-sdp 12:1 create
            exit
            no shutdown
        exit
```

The configuration of VPLS 1 on PE-2 is as follows:

```
configure
    service
        sdp 21 mpls create
            far-end 192.0.2.1
            ldp
            no shutdown
        exit
        vpls 1 customer 1 create
            sap 1/1/3:cp-1 create
            exit
            sap 1/2/1:1.21 create
            exit
            sap 1/2/1:100.cp-1 create
            exit
            sap 1/2/3:cp-1.* create
            exit
            spoke-sdp 21:1 create
            exit
            no shutdown
        exit
```

When the CEs send traffic to each other, such as ICMP echo requests, the MAC addresses are learned in the SAPs, and the forwarding database (FDB) on PE-1 is as follows:

```
*A:PE-1# show service id 1 fdb detail

===============================================================================
Forwarding Database, Service 1
===============================================================================
ServId    MAC               Source-Identifier        Type     Last Change
                                                     Age
-------------------------------------------------------------------------------
1         00:00:01:11:11:11 sap:1/2/1:1.11           L/0      05/15/17 10:34:42
1         00:00:01:17:17:17 sap:1/1/3:cp-1           L/0      05/15/17 10:34:29
1         00:00:01:18:18:18 sap:1/2/1:100.cp-1       L/0      05/15/17 10:34:33
1         00:00:01:19:19:19 sap:1/2/3:cp-1.*         L/0      05/15/17 10:34:37
1         00:00:02:21:21:21 sdp:12:1                 L/0      05/15/17 10:34:42
1         00:00:02:27:27:27 sdp:12:1                 L/0      05/15/17 10:34:29
1         00:00:02:28:28:28 sdp:12:1                 L/0      05/15/17 10:08:53
1         00:00:02:29:29:29 sdp:12:1                 L/0      05/15/17 10:34:37
-------------------------------------------------------------------------------
No. of MAC Entries: 8
-------------------------------------------------------------------------------
Legend:  L=Learned O=Oam P=Protected-MAC C=Conditional S=Static Lf=Leaf
===============================================================================
*A:PE-1#
```

VLAN Manipulation in Dot1q SAPs

Figure 143 shows the VLAN manipulation for VLAN SAPs. CE-17 and CE-18 are connected to VLAN SAPs, where the VLAN tag with VID 7 will be popped or pushed. VLAN translation is possible, but does not apply. The configuration of the SAPs in VPLS 1 on PE-1 and PE-2 is modified as follows:

```
*A:PE-1# configure service vpls 1 sap 1/1/3:cp-1 shutdown
*A:PE-1# configure service vpls 1 no sap 1/1/3:cp-1
*A:PE-1# configure service vpls 1 sap 1/1/3:7 create

*A:PE-2# configure service vpls 1 sap 1/1/3:cp-1 shutdown
*A:PE-2# configure service vpls 1 no sap 1/1/3:cp-1
*A:PE-2# configure service vpls 1 sap 1/1/3:7 create
```

Figure 143 **Customer VIDs are Popped and Pushed by Dot1q VLAN SAPs**

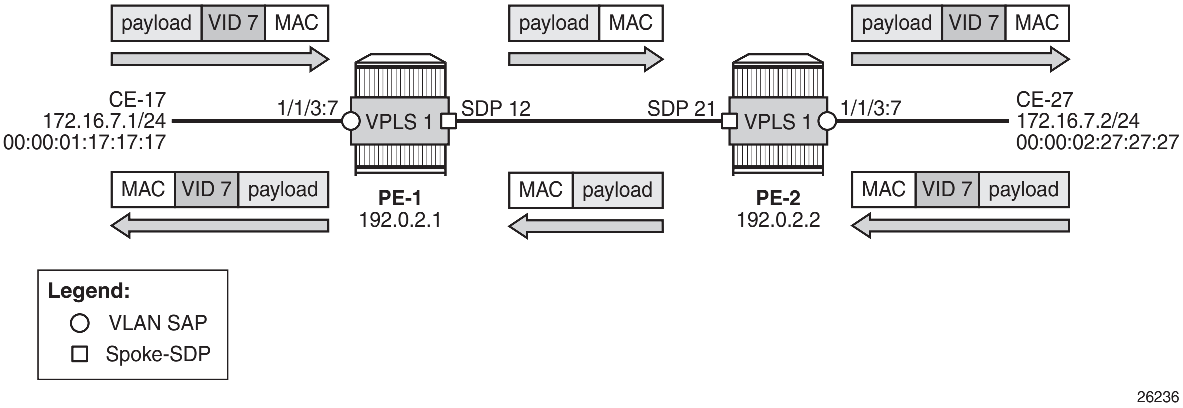
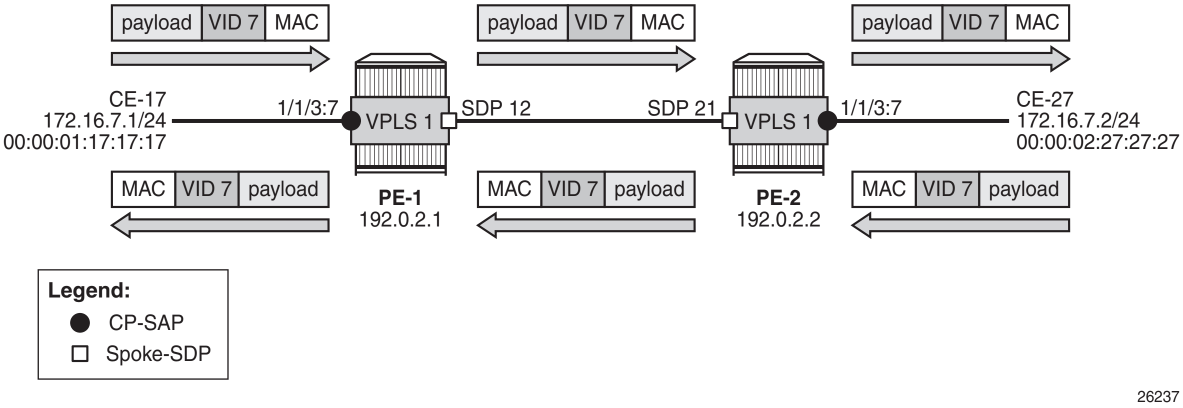


Figure 144 shows how the customer VID 7 is preserved between CE-17 and CE-27 when CP SAPs are used instead of VLAN SAPs. The configuration for the SAPs is modified as follows:

```
*A:PE-1# configure service vpls 1 sap 1/1/3:7 shutdown
*A:PE-1# configure service vpls 1 no sap 1/1/3:7
*A:PE-1# configure service vpls 1 sap 1/1/3:cp-1 create

*A:PE-2# configure service vpls 1 sap 1/1/3:7 shutdown
*A:PE-2# configure service vpls 1 no sap 1/1/3:7
*A:PE-2# configure service vpls 1 sap 1/1/3:cp-1 create
```

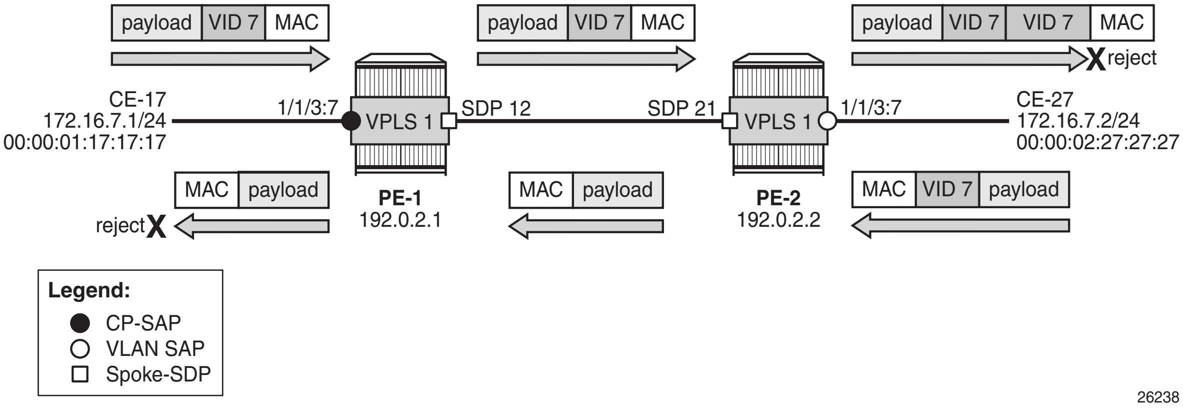
CE-17 sends frames with VID 7 to dot1q CP SAP 1/1/3:cp-1 in VPLS 1 on PE-1, and this CP SAP preserves the VLAN tag. When the frames with VID 7 reach the egress CP SAP 1/1/3:cp-1 of VPLS 1 on PE-2, the egress CP SAP preserves the VID, and the frames are forwarded to CE-27. Traffic in the opposite direction is treated in the same way: the customer VID is preserved between the CEs.

Figure 144 **Customer VID is Preserved between Two Dot1q CP SAPs**



No traffic is possible between a CP SAP in VPLS 1 on PE-1 and a VLAN SAP in VPLS 1 on PE-2, as shown in Figure 145.

Figure 145 **No Traffic Between Dot1q CP SAP and Dot1q VLAN SAP**



The CP SAP 1/1/3:cp-1 in VPLS 1 on PE-1 remains unchanged, whereas the SAP in VPLS 1 on PE-2 is reconfigured as VLAN SAP 1/1/3:7 for VLAN 7, as follows:

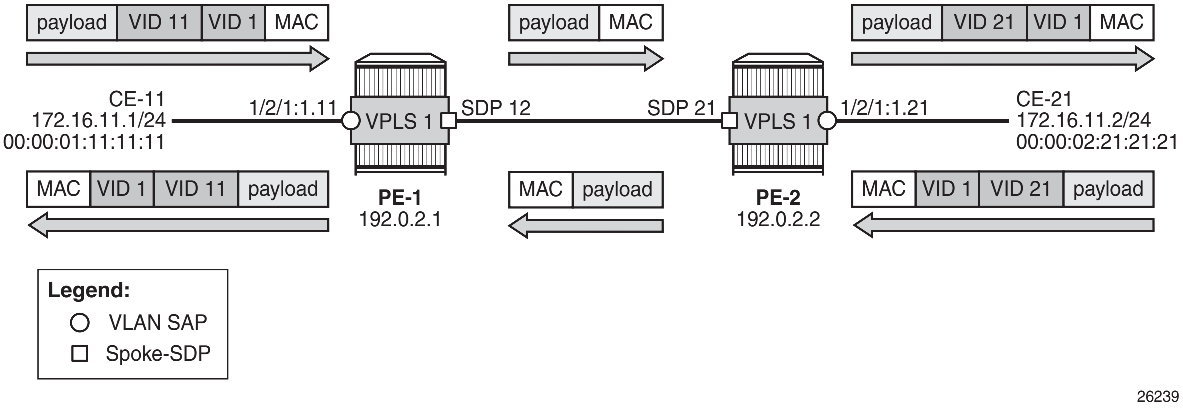
```
*A:PE-2# configure service vpls 1 sap 1/1/3:cp-1 shutdown
*A:PE-2# configure service vpls 1 no sap 1/1/3:cp-1
*A:PE-2# configure service vpls 1 sap 1/1/3:7 create
```

Frames from CE-17 are forwarded by CP SAP 1/1/3:cp-1 in VPLS 1 on PE-1 without any changes to the VLAN tag. The tagged frames reach the VLAN SAP 1/1/3:7, where another VLAN tag with VID 7 is pushed onto the frame. The receiver CE-27 rejects the double-tagged frame. When CE-27 sends traffic to CE-17, the VLAN SAP 1/1/3:7 in VPLS 1 on PE-2 pops the VLAN tag and the frame is forwarded untagged to PE-1. The CP SAP 1/1/3:cp-1 on PE-1 does not push any VLAN tag and the frame is forwarded untagged to CE-17, where it is rejected.

VLAN Manipulation in QinQ SAPs

Figure 146 shows the VLAN manipulation in QinQ VLAN SAPs that pop and push the VLAN labels. In the example, the customer VID is translated.

Figure 146 **Traffic between Two QinQ VLAN SAPs - VLAN Translation**



CE-11 sends double-tagged traffic to QinQ VLAN SAP 1/2/1:1.11 in VPLS 1 on PE-1. This VLAN SAP pops both labels and forwards the frame untagged to PE-2. The egress VLAN SAP 1/2/1:1.21 in VPLS 1 on PE-2 pushes a label stack with two labels: the inner label with VID 21 and the outer label with VID 1. Both VIDs can be translated, but in this example, only the inner label gets another VID.

Figure 147 shows that VLAN translation is not possible between two QinQ CP SAPs. In the example, the outer tag with VID 1 is popped by the CP SAPs (VLAN translation is possible for this VLAN tag, but not done here) and the inner tag with VID 11 or 21 is preserved by the CP SAPs, which implies that the received frames will be rejected.

In this example, CP 2 is configured on both PE-1 and PE-2 with one VLAN range with one VID (11 or 21), as follows:

```
*A:PE-1# configure connection-profile-vlan 2 create vlan-range 11
```

```
*A:PE-2# configure connection-profile-vlan 2 create vlan-range 21
```

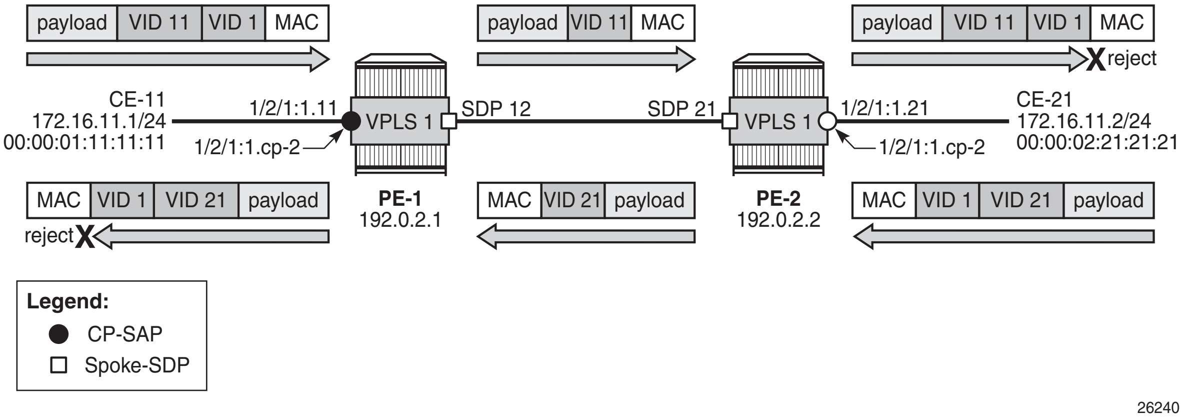
The VLAN SAP 1/2/1:1.11 is replaced by CP SAP 1/2/1:1.cp-2, as follows:

```
*A:PE-1# configure service vpls 1 sap 1/2/1:1.11 shutdown
*A:PE-1# configure service vpls 1 no sap 1/2/1:1.11
*A:PE-1# configure service vpls 1 sap 1/2/1:1.cp-2 create
```

Likewise, the VLAN 1/2/1:1.21 is replaced by CP SAP 1/2/1:1.cp-2, as follows:

```
*A:PE-2# configure service vpls 1 sap 1/2/1:1.21 shutdown
*A:PE-2# configure service vpls 1 no sap 1/2/1:1.21
*A:PE-2# configure service vpls 1 sap 1/2/1:1.cp-2 create
```

Figure 147 **No Traffic between Two QinQ CP SAPs - VLAN Translation Not Supported**

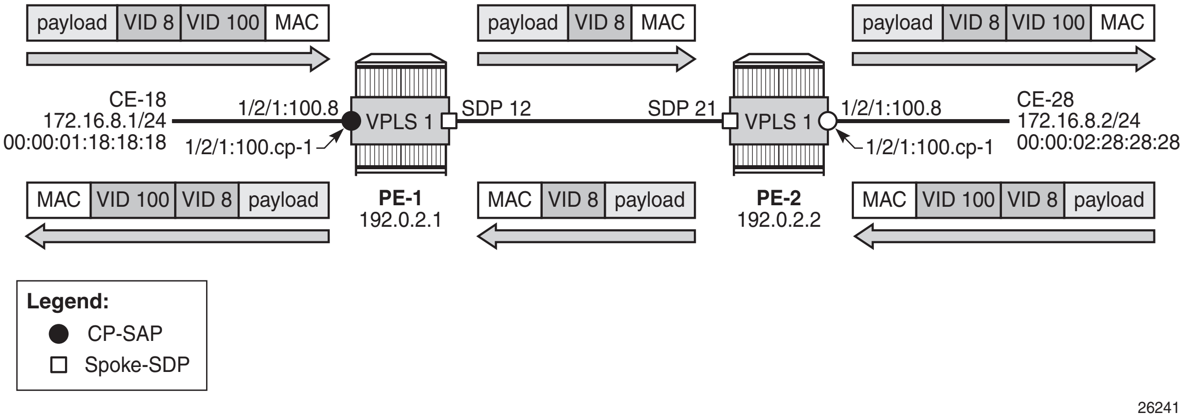


CE-11 sends double-tagged frames to SAP 1/2/1:1.cp-2 in VPLS 1 on PE-1. This CP SAP pops the outer tag with VID 1, but preserves the VLAN tag with VID 11. The single-tagged frame is sent to PE-2 where CP SAP 1/2/1:1.cp-2 pushes an outer tag with VID 1 onto the frame. This double-tagged frame is sent to CE-12 where it is rejected, because an inner label with VID 21 is expected.

When CE-21 sends frames to CE-11, the frames will be double-tagged with inner tag VID 21 and outer tag 1. The outer tag is popped by the ingress SAP 1/2/1:1.cp-2 in VPLS 1 on PE-2, but the inner tag is preserved. The egress SAP 1/2/1:1.cp-2 in VPLS 1 on PE-1 preserves the inner tag with VID 21 and pushes an outer tag with VID 1. This double-tagged frame is rejected by CE-11, because another inner tag is expected, with VID 11 instead of VID 21.

Figure 148 shows how traffic is sent between two QinQ CP SAPs without VLAN translation. Both CE-18 and CE-28 send double-tagged frames with inner tag VID 8 and outer tag VID 100. The tag with VID 100 need not be the same on both CEs, because it is popped and pushed by the CP SAPs; only the tag with VID 8 from the VLAN range must be unchanged.

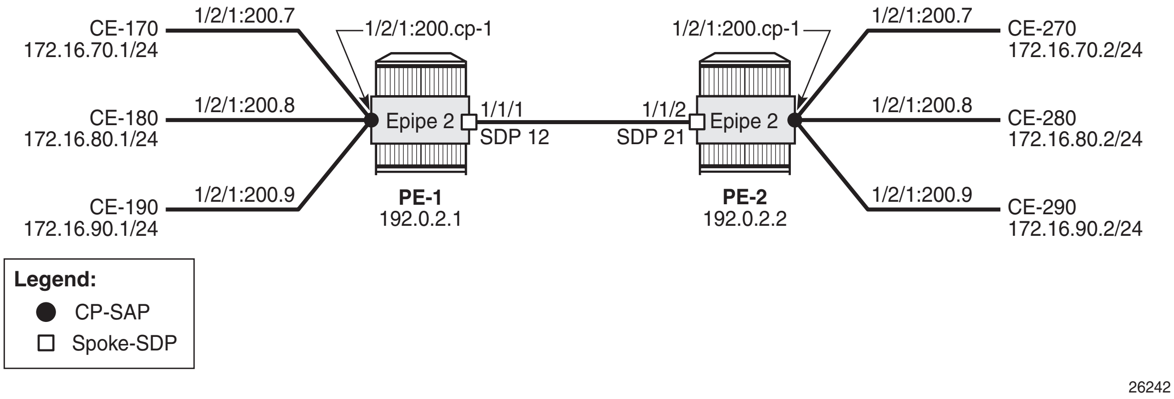
Figure 148 Traffic between Two QinQ CP SAPs - No VLAN Translation



VLAN Ranges in Epipe Services

Figure 149 shows the example topology for VLAN ranges in Epipe 2.

Figure 149 Example Topology for VLAN Ranges in Epipe 2



Epipe 2 is configured with one CP SAP and a spoke-SDP, as follows:

```
configure
    service
        sdp 12 mpls create
            far-end 192.0.2.2
            ldp
            no shutdown
        exit
        epipe 2 customer 1 create
```

```
            sap 1/2/1:200.cp-1 create
            exit
            spoke-sdp 12:2 create
            exit
            no shutdown
        exit
```

CE-170 and CE-270 send double-tagged frames with inner VID 7 and outer VID 200. The inner VID 7 is preserved by the CP SAPs; therefore, CE-170 can only communicate with CE-270, not with any other CE at the other end, because they have different customer VIDs.

Conclusion

CP SAPs can be used to build services that can be bundled as per MEF 10.3 and RFC 7432. Multiple customer VIDs can be mapped to one CP-SAP.

Layer 3 Services

In This Section

This section provides configuration information for the following topics:

- BGP Best External in a VPRN
- Carrier Supporting Carrier IP VPNs
- Layer 3 VPN: VPRN Type Spoke
- NG-MVPN Configuration with MPLS
- NG-MVPN Configuration with PIM
- NG-MVPN Sender-Only, Receiver-Only
- NG-MVPN Source Redundancy
- NG-MVPN Wildcard S-PMSI
- Rosen MVPN Core Diversity
- Rosen MVPN Inter-AS Option B
- Spoke Termination for IPv6-6VPE
- Traffic Leaking from VPRN to GRT
- VPRN Inter-AS VPN Model C

BGP Best External in a VPRN

This chapter provides information about BGP Best External in a VPRN.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

The information and configuration in this chapter was originally written for SR OS Release 14.0.R7. The CLI is updated to SR OS Release 15.0.R4.

Overview

By default, BGP speakers only advertise their best route for a destination. The BGP best external feature allows BGP speakers to advertise their best external route for a prefix/Network Layer Reachability Information (NLRI) to their iBGP peers when their best overall route for this prefix/NLRI is an internal route. This feature provides additional path visibility to the iBGP mesh. When two paths are available to reach a destination, and one is preferred, the availability of an alternate path in the RIB means that only a FIB update is required if the preferred next-hop fails. Also, the presence of two paths can reduce route oscillation.

BGP best external can be enabled in the base router with the following command:

```
*A:PE-2# configure router bgp
  - bgp
  - no bgp
 [no] add-paths       + Enable/Disable BGP ADD-PATHS
 [no] advertise-exte* - Enable/Disable Advertise Best External for the bgp family
 [no] advertise-inac* -
 Enable/disable advertising of inactive BGP routes to other BGP peers
---snip---
```

```
*A:PE-2# configure router bgp advertise-external ipv4
```

Chapter *BGP Add-Path* describes the use of the add-paths parameter for different address families. Chapter *BGP Fast Reroute* includes a configuration example with BGP best external enabled in the base router, whereas this chapter focuses on BGP best external in a VPRN context.

VPRN BGP best external can be configured with the following command:

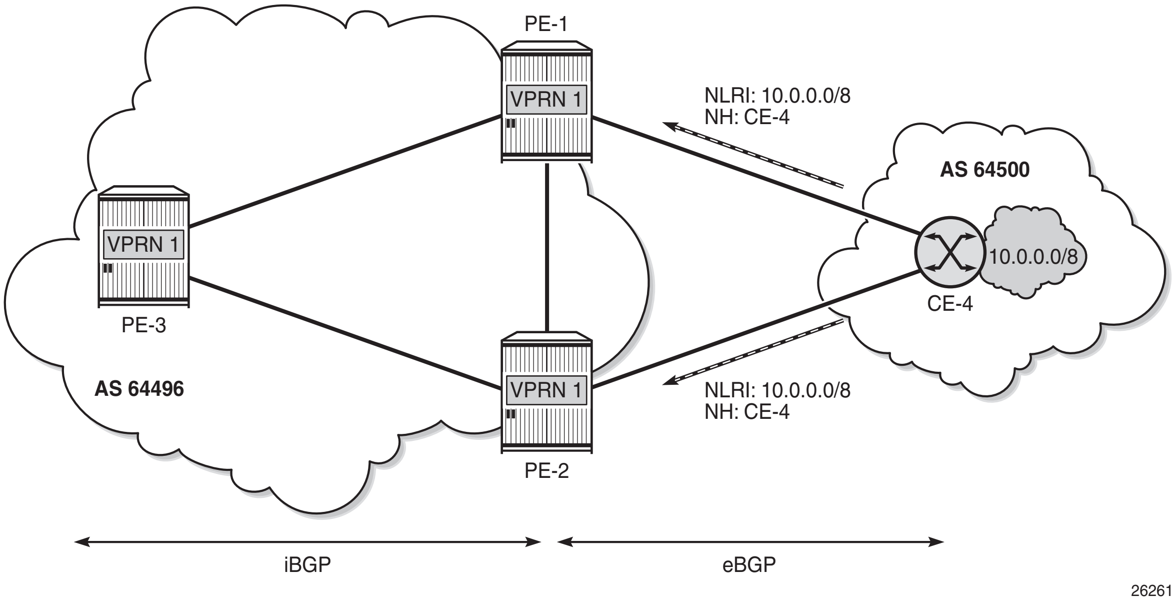
```
*A:PE-2# configure service vprn 1
  - vprn <service-id> [customer <customer-id>] [create]
  - no vprn <service-id>
---snip---
[no] export-inactiv* - Allow/Disallow exporting inactive BGP routes
---snip---
```

```
*A:PE-2# configure service vprn 1 export-inactive-bgp
```

VPRN BGP best external allows the best eBGP IPv4/IPv6 route learned by a VPRN to be exported as a BGP VPN-IPv4/IPv6 route, even when that eBGP IPv4/IPv6 route is inactive due to the presence of a preferred BGP VPN-IPv4/IPv6 route from another PE. This best external route advertisement is useful in active/standby multi-homing scenarios because it can ensure that all PEs have knowledge of the backup path provided by the standby PE, thus reducing convergence times. VPRN BGP best external can also be applied in combination with equal cost multipath (ECMP).

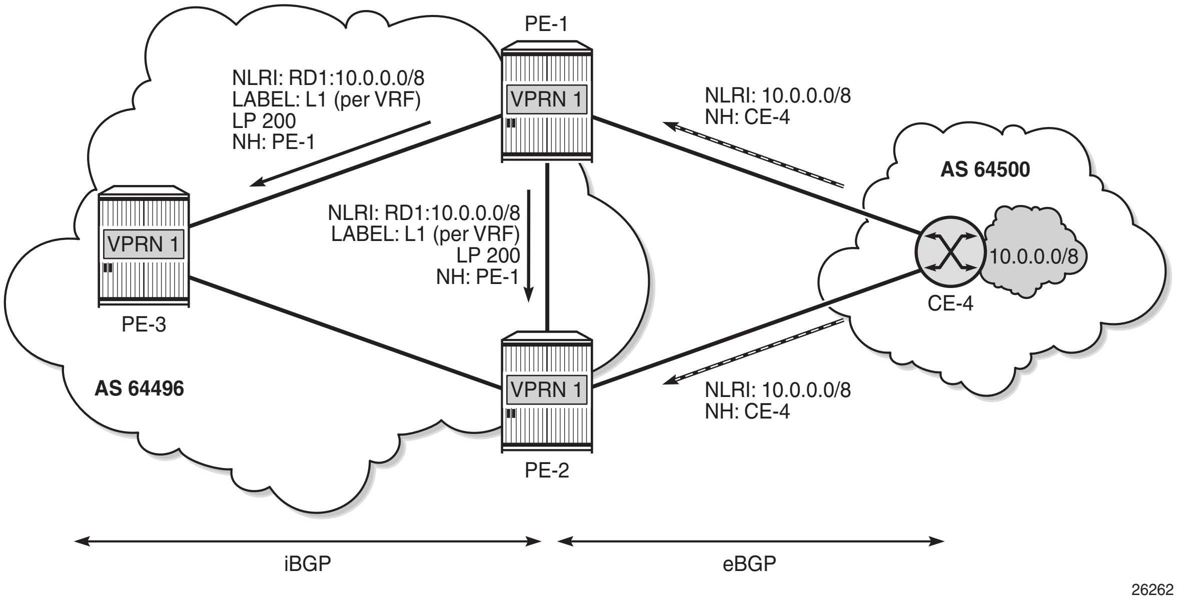
Figure 150 shows the example topology with CE-4 in autonomous system (AS) 64500 advertising prefix 10.0.0.0/8 to VPRN 1 in PE-1 and PE-2 in AS 64496.

Figure 150 **CE-4 Advertises prefix 10.0.0.0/8 to its eBGP Peers PE-1 and PE-2**



PE-1 is the primary PE for this prefix and it creates a corresponding BGP VPN-IPv4 route with a higher local preference (LP) value (for example, 200) compared to the default one (100). PE-1 advertises this BGP VPN-IPv4 route to its iBGP peers PE-2 and PE-3. PE-2 imports this BGP VPN-IPv4 route into its VRF, which deactivates the eBGP route received from CE-4, because it has the default LP of 100 (by BGP selection rules, highest LP wins). By default, BGP prevents PE-2 from exporting its inactive BGP IPv4 route from CE-4 and, therefore, PE-1 and PE-3 cannot learn a BGP VPN-IPv4 backup route for prefix 10.0.0.0/8, as shown in Figure 151.

Figure 151 **Default BGP Behavior: BGP Advertises Best Route Only**

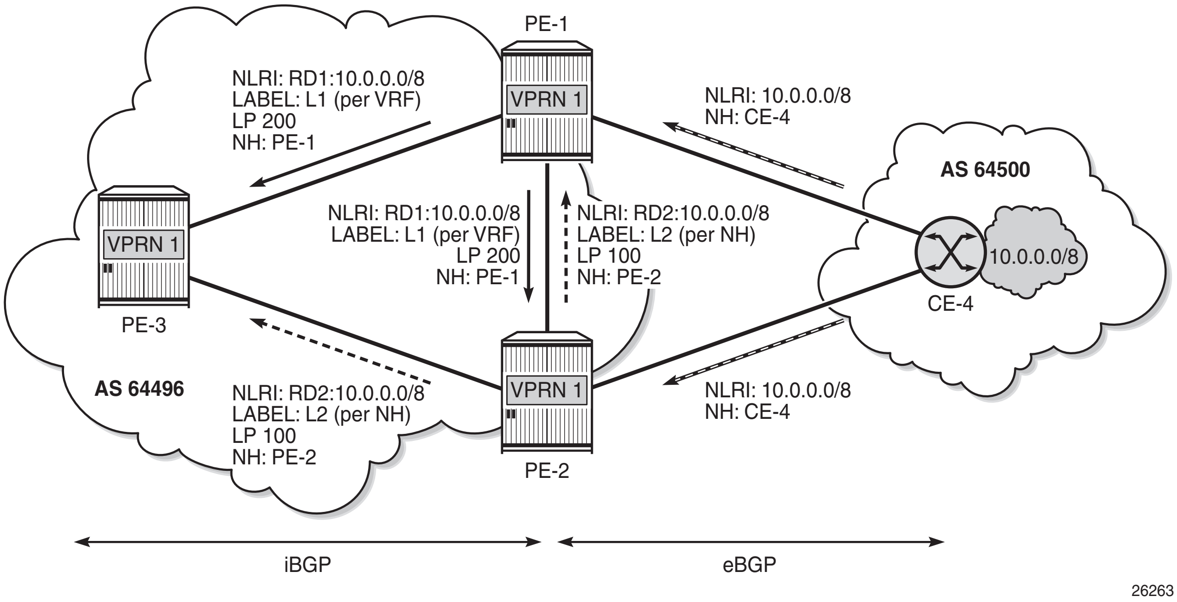


VPRN BGP best external allows PE-2 to advertise its best external route as backup on the following conditions:

- The option **export-inactive-bgp** is configured in VPRN 1 on PE-2 (or on all PEs in the multi-homed site).
- The BGP route from CE-4 must match the VRF export policy in PE-2.
- The BGP VPN-IPv4 route exported by PE-2 must have a unique NLRI (RD:IP prefix combination) that does not overlap with a BGP VPN-IPv4 route from another PE for the same prefix. Therefore, a different RD can be allocated to the VRF in each PE connected to the multi-homed site. For example, VPRN 1 in PE-1 has RD 64496:11 and VPRN 1 in PE-2 has RD 64496:12.

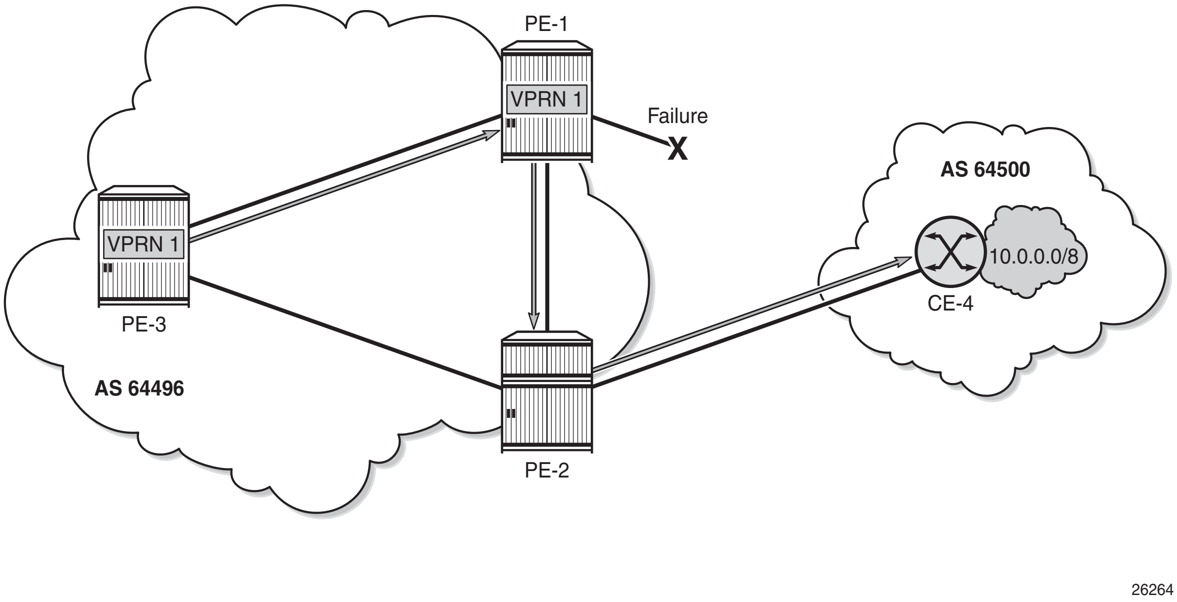
Figure 152 shows the BGP route advertisements when VPRN BGP best external is enabled. The BGP VPN-IPv4 route from PE-2 carries a per-next-hop label (meaning pop and forward to CE-4) regardless of the configured label mode of the VPRN service in PE-2.

Figure 152 **VPRN BGP Best External Enabled: BGP Advertises Active and Standby Routes**



The PEs support BGP fast reroute using BGP VPN-IPv4 routes; therefore, PE-1 and PE-3 can install the route advertised by PE-2 as a backup path for prefix 10.0.0.0/8 and use that path immediately after detecting that the primary path has failed. When the link between PE-1 and CE-4 fails, PE-1 will detect this link failure typically seconds before the other PEs do. Therefore, PE-3 keeps sending traffic toward the network 10.0.0.0/8 to PE-1 and PE-1 uses the repair path via PE-2, as shown in Figure 153.

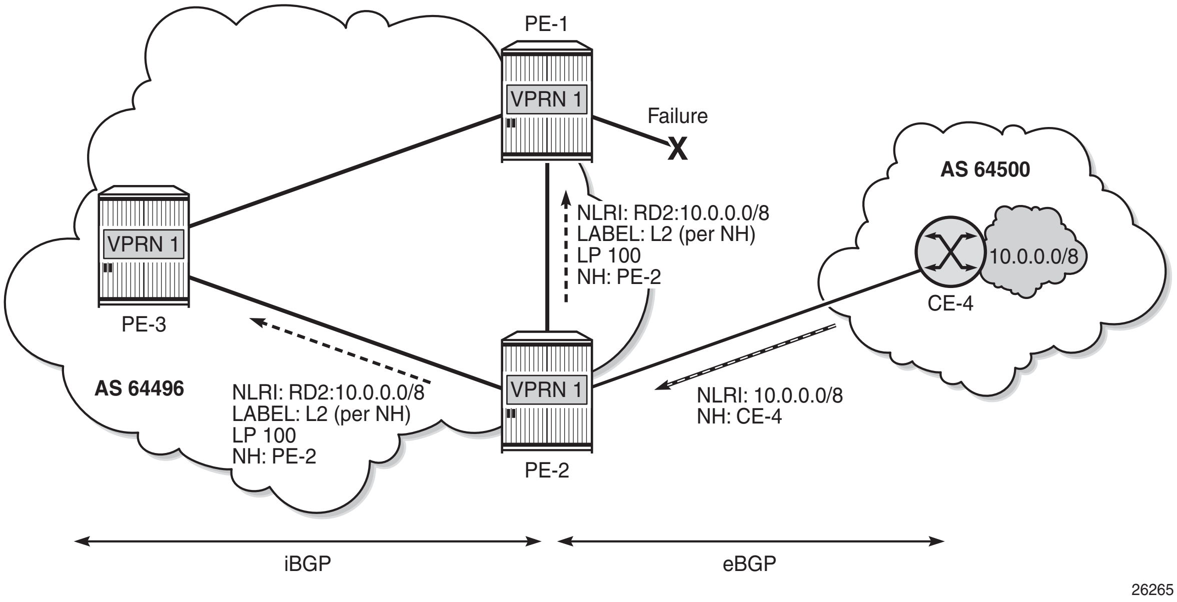
Figure 153 **BGP Fast Reroute on PE-1 after Failure of Active Link to CE**



Even when PE-2 is still unaware of the link failure between PE-1 and CE-4, PE-2 will not loop traffic back to PE-1. The reason is that PE-1 sends traffic to PE-2 with a per-next-hop label so that no FIB lookup occurs in PE-2. Traffic is forwarded correctly to CE-4.

When PE-2 receives the BGP VPN-IPv4 route withdrawal from PE-1 for prefix 10.0.0.0/8, it removes the route from its RIB-IN and reruns the decision process. In this example, the eBGP route to CE-4 becomes the new primary/best path. PE-2 will re-advertise its BGP VPN-IPv4 route for prefix 10.0.0.0/8. The difference is that the BGP VPN-IPv4 route is based on the export of an active/used route and, therefore, the advertised label value is based on the configured label mode of the VPRN service, as shown in Figure 154 for label mode VRF (default).

Figure 154 **PE-2 Re-Advertises VPN-IPv4 Route with Label Based on VRF**



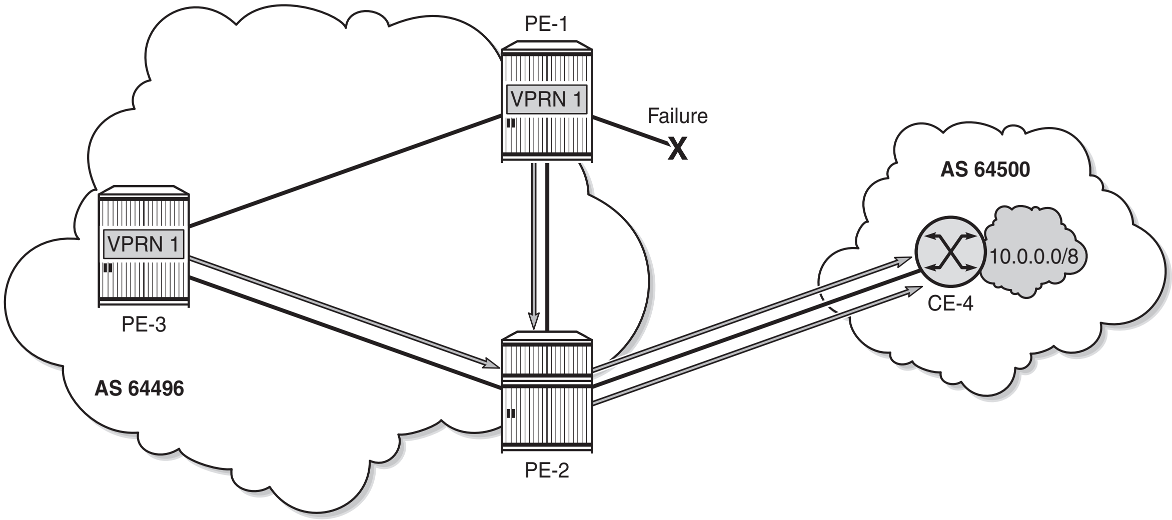
It takes time for this route to reach all ingress routers and for these routers to update their forwarding tables to use the per-VRF label value. For a while, there may still be traffic destined for prefix 10.0.0.0/8 that is received by PE-2 with the per-next-hop label L2. Traffic will be dropped if the per-next-hop label is deleted by the IOM as soon as PE-2 determines there are no more inactive/standby paths with CE-4 as next hop. Traffic loss can be avoided by delaying the deletion of per-next-hop labels in the IOM by configuring label retention for BGP labels with the following command:

```
*A:PE-2# configure router mpls-labels bgp-labels-hold-timer
  - bgp-labels-hold-timer <seconds>
  - no bgp-labels-hold-timer
 <seconds>            : [0..255]
```

```
*A:PE-2# configure router mpls-labels bgp-labels-hold-timer 60
```

Finally, all ingress routers have updated their forwarding tables based on the BGP update sent by PE-2, and PE-3 will send traffic for prefix 10.0.0.0/8 directly toward PE-2, as shown in Figure 155.

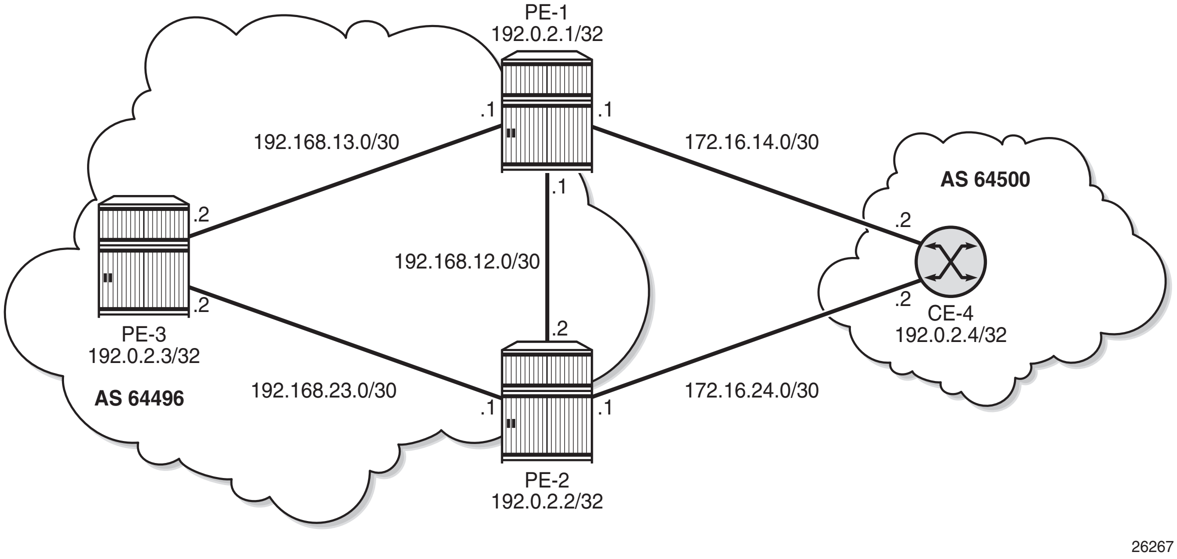
Figure 155 **Traffic Destined for Prefix 10.0.0.0/8 after Control Plane Convergence**



Configuration

Figure 156 shows the example topology with the used IP addresses.

Figure 156 **Example Topology**



The initial configuration includes the following:

- Cards, MDAs, ports
- Router interfaces
- IS-IS (or OSPF) as IGP within AS 64496
- LDP on all interfaces within AS 64496

BGP is configured in the base router context of all PEs for address family VPN-IPv4; for example, for PE-1 as follows:

```
configure
    router
        autonomous-system 64496
        bgp
            min-route-advertisement 1
            rapid-withdrawal
            group "iBGP"
                family vpn-ipv4
                peer-as 64496
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
            exit
```

The BGP configuration for the base router on the other two PEs is similar and a full mesh is established in AS 64496.

Configure VPRN without BGP Best External

VPRN 1 is created on all PEs with the following settings:

- Default label mode: label-mode vrf
- Ready for BGP fast reroute: enable-bgp-vpn-backup ipv4
- Different RDs in VPRN 1 for each PE: 64496:11 on PE-1, 64496:12 on PE-2, and 64496:13 on PE-3
- Auto-bind-tunnel with resolution any. In this example, LDP will be used.
- Loopback interface "lo0" with IP address 1.1.1.1/32 on PE-1, which is also defined as the router ID in VPRN 1. The same approach is used on PE-2 and PE-3: 2.2.2.2/32 and 3.3.3.3/32.
- iBGP between all PEs (full mesh) for address family IPv4
- eBGP between PE-1 and CE-4 and between PE-2 and CE-4
- BGP best external is disabled, by default.

The configuration of VPRN 1 on PE-3 is as follows:

```
configure
    service
        vprn 1 customer 1 create
            router-id 3.3.3.3
            autonomous-system 64496
            route-distinguisher 64496:13
            auto-bind-tunnel
                resolution any
            exit
            label-mode vrf  # default
            enable-bgp-vpn-backup ipv4
            vrf-target target:64496:1
            interface "lo0" create
                address 3.3.3.3/32
                loopback
            exit
            bgp
                min-route-advertisement 1
                rapid-withdrawal
                group "iBGP"
                    peer-as 64496
                    neighbor 1.1.1.1
                    exit
                    neighbor 2.2.2.2
                    exit
                exit
            exit
            no shutdown
```

On PE-1 and PE-2, the VPRN configuration includes an external interface toward CE-4, and eBGP is defined toward peer CE-4. The VPRN 1 configuration on PE-2 is as follows:

```
configure
    service
        vprn 1 customer 1 create
            router-id 2.2.2.2
            autonomous-system 64496
            route-distinguisher 64496:12
            auto-bind-tunnel
                resolution any
            exit
            enable-bgp-vpn-backup ipv4
            vrf-target target:64496:1
            interface "lo0" create
                address 2.2.2.2/32
                loopback
            exit
            interface "int-PE-2-CE-4_VPRN1" create
                address 172.16.24.1/30
                sap 1/1/3:1 create
                exit
            exit
            bgp
                min-route-advertisement 1
                rapid-withdrawal
                split-horizon
                group "eBGP"
                    peer-as 64500
                    neighbor 172.16.24.2
                    exit
                exit
                group "iBGP"
                    peer-as 64496
                    neighbor 1.1.1.1
                    exit
                    neighbor 3.3.3.3
                    exit
                exit
            exit
            no shutdown
```

PE-2 does not have an import policy that sets the LP and, therefore, the default LP of 100 is used for routes imported from eBGP peer CE-4.

The VPRN 1 configuration on PE-1 looks similar to the configuration on PE-2, but includes an import policy that assigns an LP of 200 to each prefix that is received from CE-4, as follows:

```
configure
    service
        vprn 1 customer 1 create
            router-id 1.1.1.1
            autonomous-system 64496
            route-distinguisher 64496:11
            auto-bind-tunnel
                resolution any
            exit
            enable-bgp-vpn-backup ipv4
            vrf-target target:64496:1
            interface "lo0" create
                address 1.1.1.1/32
                loopback
            exit
            interface "int-PE-1-CE-4_VPRN1" create
                address 172.16.14.1/30
                sap 1/1/3:1 create
                exit
            exit
            bgp
                min-route-advertisement 1
                rapid-withdrawal
                split-horizon
                group "eBGP"
                    import "import-bgp-LP200"
                    peer-as 64500
                    neighbor 172.16.14.2
                    exit
                exit
                group "iBGP"
                    peer-as 64496
                    neighbor 2.2.2.2
                    exit
                    neighbor 3.3.3.3
                    exit
                exit
            exit
            no shutdown
```

The import policy is defined on PE-1 as follows:

```
configure
    router
        policy-options
            begin
            policy-statement "import-bgp-LP200"
                default-action accept
                    local-preference 200
                exit
            exit
            commit
```

CE-4 has eBGP configured toward PE-1 and PE-2, as follows:

```
configure
    router
        interface "int-CE-4-PE-1_VPRN1"
            address 172.16.14.2/30
            port 1/1/1:1
        exit
        interface "int-CE-4-PE-2_VPRN1"
            address 172.16.24.2/30
            port 1/1/2:1
        exit
        interface "system"
            address 192.0.2.4/32
        exit
        interface "test_connectedNW"
            address 10.0.0.1/8
            loopback
        exit
        autonomous-system 64500
        bgp
            min-route-advertisement 1
            rapid-withdrawal
            split-horizon
            group "eBGP"
                export "export-bgp"
                peer-as 64496
                neighbor 172.16.14.1
                exit
                neighbor 172.16.24.1
                exit
            exit
        exit
```

CE-4 exports the prefix 10.0.0.0/8, as defined in the following export policy that is applied in the BGP context:

```
configure
    router
        policy-options
            begin
            prefix-list "10.0.0.0/8"
                prefix 10.0.0.0/8 longer
            exit
            policy-statement "export-bgp"
                entry 10
                    from
                        prefix-list "10.0.0.0/8"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

Initially, VPRN BGP best external is disabled and, so only the best BGP route will be advertised and iBGP peers will not learn backup paths. The following section shows which routes are exchanged. Afterward, VPRN BGP best external will be enabled and the same show commands will be used.

Verification - VPRN without BGP Best External

VPRN with BGP best external results in the following. PE-1 imports prefix 10.0.0.0/8, assigns LP 200 to it, and advertises a corresponding VPN-IPv4 route to its iBGP peers (PE-2 and PE-3). Toward PE-2, this is as follows:

```
11 2017/09/26 11:23:27.594 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 65
    Flag: 0x90 Type: 14 Len: 30 Multiprotocol Reachable NLRI:
        Address Family VPN_IPV4
        NextHop len 12 NextHop 192.0.2.1
        10.0.0.0/8 RD 64496:11 Label 262140
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 6 AS Path:
        Type: 2 Len: 1 < 64500 >
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 200
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64496:1
"
```

The NLRI includes the prefix 10.0.0.0/8 and the RD 64496:11, and the label is 262140. BGP prevents PE-2 from sending a similar BGP update for prefix 10.0.0.0/8 because that route is not active on PE-2. PE-3 receives a BGP VPN-IPv4 route for network 64496:11:10.0.0.0/8, and this route has PE-1 as next hop and LP 200. No route is received from PE-2 for network 64496:12:10.0.0.0/8; as follows:

```
*A:PE-3# show router bgp routes vpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  64496:11:1.1.1.1/32                                100         None
      192.0.2.1                                          None        262140
      No As-Path
u*>i  64496:11:10.0.0.0/8                                200         None
      192.0.2.1                                          None        262140
      64500
u*>i  64496:11:172.16.14.0/30                            100         None
      192.0.2.1                                          None        262140
      No As-Path
u*>i  64496:12:2.2.2.2/32                                100         None
```

```
      192.0.2.2                                          None        262140
      No As-Path
u*>i  64496:12:172.16.24.0/30                            100         None
      192.0.2.2                                          None        262140
      No As-Path
-------------------------------------------------------------------------------
Routes : 5
===============================================================================
*A:PE-3#
```

In a similar way, the list of BGP VPN-IPv4 routes on PE-2 includes prefix 64496:11:10.0.0.0/8 with LP 200 and next hop PE-1, as follows:

```
*A:PE-2# show router bgp routes vpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  64496:11:1.1.1.1/32                                100         None
      192.0.2.1                                          None        262140
      No As-Path
u*>i  64496:11:10.0.0.0/8                                200         None
      192.0.2.1                                          None        262140
      64500
u*>i  64496:11:172.16.14.0/30                            100         None
      192.0.2.1                                          None        262140
      No As-Path
u*>i  64496:13:3.3.3.3/32                                100         None
      192.0.2.3                                          None        262140
      No As-Path
-------------------------------------------------------------------------------
Routes : 4
===============================================================================
*A:PE-2#
```

The list of BGP IPv4 routes in VPRN 1 on PE-2 has two entries for prefix 10.0.0.0/8, but none of them is best or used, as follows:

```
*A:PE-2# show router 1 bgp routes
===============================================================================
 BGP Router ID:2.2.2.2          AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
```

```
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
*i    10.0.0.0/8                                         None        None
      172.16.24.2                                        None        -
      64500
i     10.0.0.0/8                                         200         None
      172.16.14.2                                        None        -
      64500
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-2#
```

The routing table for VPRN 1 on PE-2 and PE-3 for prefix 10.0.0.0/8 shows that the next hop is PE-1 and the protocol is BGP VPN, as follows:

```
*A:PE-2# show router 1 route-table 10.0.0.0/8

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.0.0.0/8                                    Remote  BGP VPN   00h06m30s  170
       192.0.2.1 (tunneled)                                         0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-2#
```

Instead of using an external route to CE-4, the route for prefix 10.0.0.0/8 is internal (BGP VPN), using an LDP transport tunnel to PE-1. There are no non-active routes, as can be shown by adding the keyword **all** to the preceding show command, as follows:

```
*A:PE-2# show router 1 route-table 10.0.0.0/8 all

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                         Active     Metric
-------------------------------------------------------------------------------
10.0.0.0/8                                    Remote  BGP VPN   00h06m30s  170
       192.0.2.1 (tunneled)                            Y            0
```

```
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
       E = Inactive best-external BGP route
===============================================================================
*A:PE-2#
```

There are no standby routes, because BGP only advertises the best used route.

On PE-1, the following BGP IPv4 route with next hop CE-4 is used for prefix 10.0.0.0/8 in VPRN 1:

```
*A:PE-1# show router 1 bgp routes
===============================================================================
 BGP Router ID:1.1.1.1          AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  10.0.0.0/8                                         200         None
      172.16.14.2                                        None        -
      64500
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-1#
```

The route for prefix 10.0.0.0/8 in the routing table of VPRN 1 has next hop CE-4, as follows:

```
*A:PE-1# show router 1 route-table 10.0.0.0/8 all
===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                         Active     Metric
-------------------------------------------------------------------------------
10.0.0.0/8                                    Remote  BGP       00h02m03s  170
       172.16.14.2                                     Y            0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
       E = Inactive best-external BGP route
===============================================================================
*A:PE-1#
```

There is no backup route because BGP prevents PE-2 from sending a standby route for prefix 10.0.0.0/8 to its iBGP peers.

Before VPRN BGP best external is enabled, PE-2 has advertised two VPN-IPv4 routes in the base router (the last number in Rcv/Act/Sent = Received/Active/Sent), as follows:

```
A:PE-2# show router bgp summary family vpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
BGP Admin State         : Up          BGP Oper State              : Up
--- snipped ---
===============================================================================
BGP VPN-IPv4 Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
                   AS PktRcvd PktSent  InQ OutQ Up/Down   State|Recv/Actv/Sent
-------------------------------------------------------------------------------
192.0.2.1
               64496     3330    3329    0    0 01d03h41m 3/3/2
192.0.2.3
               64496     3326    3329    0    0 01d03h40m 1/1/2
-------------------------------------------------------------------------------
A:PE-2#
```

Enable BGP Best External in VPRN

VPRN BGP best external is configured on PE-2 (or on all PEs in the multi-homing site) as follows:

```
configure service vprn 1 export-inactive-bgp
```

When configured, this command causes all IPv4 and IPv6 VPRN BGP best external routes to be exported in the multi-protocol BGP (MP-BGP) domain. Best external routes are BGP routes for which all the following conditions are met:

- The BGP route is matched by the VRF export policy.
- The BGP route is inactive because a more preferred BGP VPN route for the same prefix is present in the route table manager (RTM).
- This BGP route is best and valid considering only VPRN BGP routes.

PE-2 is advertising a best external route and is called the standby PE for prefix 10.0.0.0/8. PEs can be active for some IP prefixes and standby for other IP prefixes.

Best external routes are advertised to the BGP VPN-IPv4 neighbors. In this example, the BGP VPN-IPv4 neighbors are iBGP neighbors, but they can also be eBGP neighbors. The RD must be unique across the PEs exporting a BGP VPN-IP route for the same prefix; otherwise, the best external route may not be advertised. The advertised VPRN label is based on the next hop IP of the best external route, regardless of the label mode of the VPRN in the standby PE.

Verification - VPRN with BGP Best External - BGP FRR

VPRN with BGP best external BGP FRR results in the following. VPRN BGP best external is enabled (BGP Export Inactv) in VPRN 1 on PE-2:

```
*A:PE-2# show service id 1 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 1                   Vpn Id            : 0
Service Type      : VPRN
Name              : (Not Specified)
Description       : (Not Specified)
Customer Id       : 1                   Creation Origin   : manual

--- snipped ---

Max IPv6 Routes   : No Limit
Ignore NH Metric  : Disabled
Hash Label        : Disabled
Entropy Label     : Disabled
Vrf Target        : target:64496:1
--- snipped ---
Label mode        : vrf
BGP VPN Backup    : ipv4
BGP Export Inactv : Enabled

SAP Count         : 1                   SDP Bind Count    : 0
VSD Domain        : <none>

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/3:1                              q-tag        1578    1578    Up   Up
===============================================================================
*A:PE-2#
```

After VPRN BGP best external is enabled, PE-2 advertises its standby route for prefix 10.0.0.0/8 to its iBGP peers, as follows:

```
11 2017/09/26 11:31:19.117 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 65
    Flag: 0x90 Type: 14 Len: 30 Multiprotocol Reachable NLRI:
        Address Family VPN_IPV4
        NextHop len 12 NextHop 192.0.2.2
        10.0.0.0/8 RD 64496:12 Label 262139
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 6 AS Path:
        Type: 2 Len: 1 < 64500 >
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64496:1
"
```

The RD is 64496:12, the LP is 100, and the label is 262139. The BGP update shown is sent by PE-2 toward PE-3; the BGP update sent by PE-2 toward PE-1 is similar.

The number of BGP VPN-IPv4 routes sent by PE-2 to each iBGP peer increased from 2 to 3, as follows:

```
A:PE-2# show router bgp summary all

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.1
Def. Instance  64496     3260    0 01d03h06m 3/3/3 (VpnIPv4)
                         3259    0
192.0.2.3
Def. Instance  64496     3257    0 01d03h06m 1/1/3 (VpnIPv4)
                         3260    0

--- snipped ---
-------------------------------------------------------------------------------
A:PE-2#
```

PE-3 has two BGP VPN-IPv4 routes for prefix 10.0.0.0/8: one for network 64496:11:10.0.0.0/8 with LP 200 and next hop PE-1, and one for network 64496:12:10.0.0.0/8 with LP 100 and next hop PE-2, as follows:

```
*A:PE-3# show router bgp routes 10.0.0.0/8 vpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  64496:11:10.0.0.0/8                                200         None
      192.0.2.1                                          None        262140
      64500
u*>i  64496:12:10.0.0.0/8                                100         None
      192.0.2.2                                          None        262139
      64500
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-3#
```

PE-1 has one BGP VPN-IPv4 route for network 64496:12:10.0.0.0/8 with LP 100 and next hop PE-2; PE-2 has one BGP VPN-IPv4 route for network 64496:11:10.0.0.0/8 with LP 200 and next hop PE-1.

All PEs are ready for BGP FRR and the "B" flag indicates that a BGP VPN-IPv4 backup route is available. This flag is present when the VPRN is configured for BGP FRR (enable-bgp-vpn-backup) and a standby route has been received, as follows. The B flag was not present in the output for the routing table when VPRN BGP best external was disabled, as shown earlier.

```
*A:PE-1# show router 1 route-table 10.0.0.0/8

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                              Type    Proto     Age        Pref
      Next Hop[Interface Name]                                      Metric
-------------------------------------------------------------------------------
10.0.0.0/8 [B]                                  Remote  BGP       00h02m11s  170
       172.16.14.2                                                    0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

The active route on PE-1 has CE-4 as next hop.

On PE-3, the active BGP VPN-IPv4 route for prefix 10.0.0.0/8 uses an LDP transport tunnel to PE-1; a BGP VPN-IPv4 backup route is also available, as follows:

```
*A:PE-3# show router 1 route-table 10.0.0.0/8

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.0.0.0/8 [B]                                Remote  BGP VPN   00h10m06s  170
       192.0.2.1 (tunneled)                                         0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-3#
```

The active BGP VPN-IPv4 route on PE-2 uses an LDP transport tunnel to PE-1, but no BGP backup route is available:

```
*A:PE-2# show router 1 route-table 10.0.0.0/8

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.0.0.0/8                                    Remote  BGP VPN   00h10m03s  170
       192.0.2.1 (tunneled)                                         0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-2#
```

PE-2 has a standby BGP IPv4 route that is displayed with the following show command:

```
*A:PE-2# show router 1 route-table 10.0.0.0/8 all

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                         Active     Metric
-------------------------------------------------------------------------------
10.0.0.0/8 [E]                                Remote  BGP       00h02m12s  170
```

```
       172.16.24.2                                      N            0
10.0.0.0/8                                     Remote  BGP VPN   00h10m03s  170
       192.0.2.1 (tunneled)                             Y            0
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
       E = Inactive best-external BGP route
===============================================================================
*A:PE-2#
```

The "E" flag indicates that this route is an inactive best external BGP route.

VPRN 1 on PE-1 and PE-3 is ready for BGP FRR (**enable-bgp-vpn-backup**) and PE-2 advertised a standby BGP VPN-IPv4 route for prefix 10.0.0.0/8; therefore, PE-1 and PE-3 can add an alternative route to the routing table of VPRN 1, as follows:

```
*A:PE-1# show router 1 route-table 10.0.0.0/8 alternative

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                             Type    Proto     Age        Pref
      Next Hop[Interface Name]                                     Metric
      Alt-NextHop                                                 Alt-
                                                                 Metric
-------------------------------------------------------------------------------
10.0.0.0/8                                     Remote  BGP       00h02m11s  170
       172.16.14.2                                                   0
10.0.0.0/8 (Backup)                            Remote  BGP VPN   00h02m11s  170
       192.0.2.2 (tunneled)                                          0
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       Backup = BGP backup route
       LFA = Loop-Free Alternate nexthop
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

```
*A:PE-3# show router 1 route-table 10.0.0.0/8 alternative

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                             Type    Proto     Age        Pref
      Next Hop[Interface Name]                                     Metric
      Alt-NextHop                                                 Alt-
                                                                 Metric
-------------------------------------------------------------------------------
10.0.0.0/8                                     Remote  BGP VPN   00h10m06s  170
       192.0.2.1 (tunneled)                                          0
10.0.0.0/8 (Backup)                            Remote  BGP VPN   00h10m06s  170
       192.0.2.2 (tunneled)                                          0
-------------------------------------------------------------------------------
```

```
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       Backup = BGP backup route
       LFA = Loop-Free Alternate nexthop
       S = Sticky ECMP requested
===============================================================================
*A:PE-3#
```

The alternative BGP VPN-IPv4 route for prefix 10.0.0.0/8 in VPRN 1 uses an LDP transport tunnel toward PE-2.

Configure ECMP

Because BGP best external allows advertising of an alternative path, it can also be used for load-sharing. ECMP is configured with value 2 in VPRN 1 on all PEs, as follows:

```
configure service vprn 1 ecmp 2
```

Other than the ECMP configuration, the VPRN configuration is the same as in the previous example. If ECMP is configured, BGP FRR is not needed anymore:

```
configure service vprn 1 no enable-bgp-vpn-backup
```

On PE-3, the BGP decision process will prefer the route with the highest LP and, therefore, only the route via PE-1 with LP 200 will be used and there will be no load-sharing. To ensure that the routes via PE-1 and PE-2 have the same cost, the import policy in VPRN 1 on PE-1 that sets the LP to 200 is removed, as follows:

```
configure service vprn 1 bgp group "eBGP" no import
```

BGP best external is enabled on PE-1 and PE-2, as follows:

```
configure service vprn 1 export-inactive-bgp
```

Verification - VPRN with BGP Best External - ECMP

VPRN with BGP best external ECMP results in the following. With BGP best external enabled on the PEs in the multi-homing site (PE-2 and PE-3), the following two BGP VPN-IPv4 routes are used on PE-3:

```
*A:PE-3# show router bgp routes 10.0.0.0/8 vpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  64496:11:10.0.0.0/8                                100         None
      192.0.2.1                                          None        262140
      64500
u*>i  64496:12:10.0.0.0/8                                100         None
      192.0.2.2                                          None        262140
      64500
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-3#
```

The following BGP IPv4 routes are learned in VPRN 1 on PE-3, but they are not used:

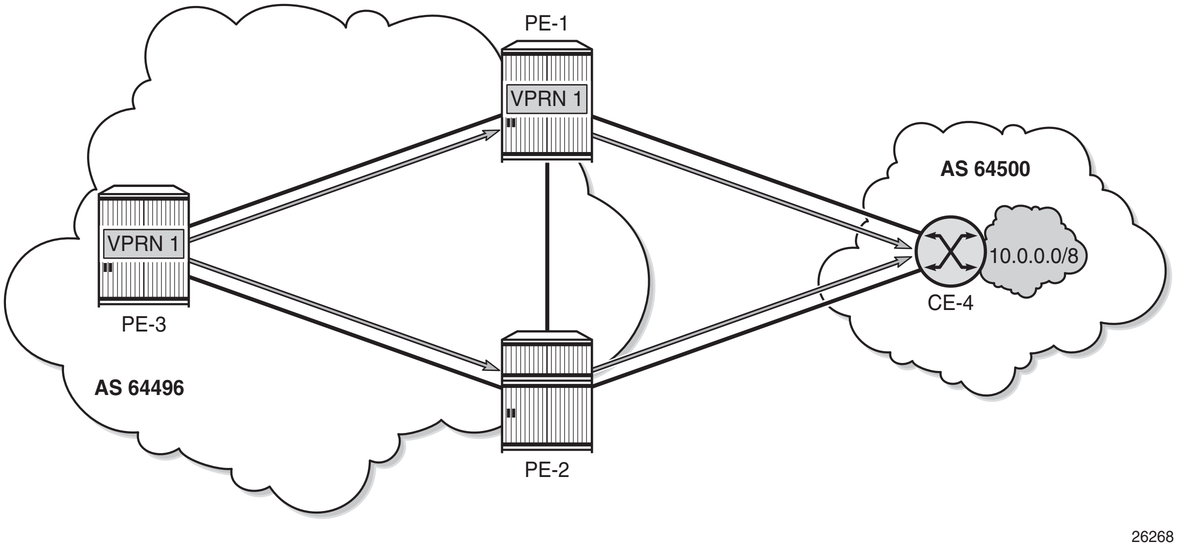
```
*A:PE-3# show router 1 bgp routes 10.0.0.0/8
===============================================================================
 BGP Router ID:3.3.3.3          AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
i     10.0.0.0/8                                         100         None
      172.16.14.2                                        None        -
      64500
i     10.0.0.0/8                                         100         None
      172.16.24.2                                        None        -
      64500
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-3#
```

When ECMP is enabled and the routes have the same LP, the routing table on PE-3 has two active routes for prefix 10.0.0.0/8, each using an LDP transport tunnel, as follows:

```
*A:PE-3# show router 1 route-table 10.0.0.0/8

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.0.0.0/8                                    Remote  BGP VPN   00h00m44s  170
       192.0.2.1 (tunneled)                                         0
10.0.0.0/8                                    Remote  BGP VPN   00h00m44s  170
       192.0.2.2 (tunneled)                                         0
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-3#
```

Figure 157 shows that traffic from VPRN 1 on PE-3 destined to prefix 10.0.0.0/8 is sprayed over two paths: one via PE-1 and one via PE-2.

Figure 157 Loadsharing for Traffic from PE-3 Destined to 10.0.0.0/8



Conclusion

VPRNs can be configured with the option **export-inactive-bgp**, which allows a BGP speaker to advertise its best external BGP route to its BGP peers even if that route is inactive due to the presence of a more preferred BGP VPN route from another PE. BGP best external in VPRN is useful in active/standby multi-homing scenarios because it allows the standby PE to advertise a backup path. The traffic failover time can be reduced when all PE routers have advance knowledge of the potential backup paths and do not have to wait for BGP route advertisements and/or withdrawals to reprogram their forwarding tables. VPRN BGP best external can also be used in combination with ECMP.

Carrier Supporting Carrier IP VPNs

This chapter provides information about carrier supporting carrier IP VPN configurations.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was originally written for and tested on 11.0.R1. The CLI in this guide corresponds to release 15.0.R5. The configuration for labeled IPv4 routes changed in SR OS release 14.0.R4, see chapter *Separate BGP RIBs for Labeled Routes*.

Overview

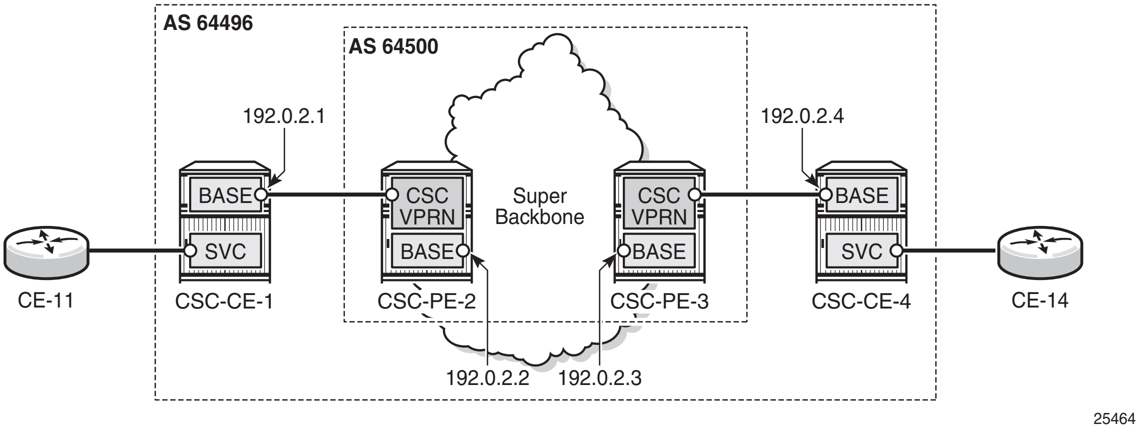
Carrier Supporting Carrier (CSC) is a solution that allows one service provider (the Customer Carrier) to use the IP VPN service of another service provider (the Super Carrier) for some or all of its backbone transport. RFC 4364 defines a Carrier Supporting Carrier solution for BGP/MPLS IP VPNs that uses MPLS at the interconnection points between the two service providers to provide a scalable and secure solution.

A simplified CSC network topology is shown in Figure 158. A CSC deployment involves the following types of devices:

- CE — Customer premises equipment dedicated to one particular business/enterprise.
- PE — Edge router managed and operated by the Customer Carrier that connects to CEs to provide business VPN or Internet services.
- CSC-CE — Peering router managed and operated by the Customer Carrier that is connected to CSC-PEs for purposes of using the associated CSC IP VPN services for backbone transport. The CSC-CE may attach directly to CEs if it is also configured to be a PE for business VPN services.

- CSC-PE — A PE router managed and operated by the Super Carrier that supports one or more CSC IP VPN services possibly in addition to other traditional PE services.

Figure 158 **CSC Network Topology**



In the CSC solution, the CSC-CE and CSC-PE are directly connected by a link that supports MPLS. The CSC-CE distributes an MPLS label for every /32 IPv4 prefix it and any downstream PE uses as a BGP next-hop in routes associated with services offered by the Customer Carrier. BGP *must be used* as the label distribution protocol between CSC-CE and CSC-PE if the latter device is a 7x50. Typically, the Customer Carrier and Super Carrier operate as two different Autonomous Systems (AS) and therefore BGP, more specifically EBGP, is the best label distribution protocol, even if other options are available. The BGP session between CSC-CE and CSC-PE must be single-hop EBGP (or IBGP) if either device is a 7x50.

In a 7x50 CSC-PE, the interface to a CSC-CE is a special type of IP/MPLS interface that belongs to a VPRN configured for CSC mode. This special type of interface is called a CSC VPRN interface throughout the remainder of this chapter. The CSC VPRN interface has many of the same characteristics as a network interface of the base router but its association with a VRF ensures that the traffic and control plane routes of the Customer Carrier are kept separate from other services.

When a 7x50 CSC-PE receives a labeled-IPv4 route (with label L1, next-hop N1) from a CSC-CE BGP peer, the following actions take place in the CSC-PE:

1. The BGP route is installed into the routing table of the CSC VPRN (assuming the BGP route is the best route to the destination).
2. If the BGP route matches the VRF export policy, it is advertised to the core MP-BGP peers as a VPN-IPv4 route. The advertised label value is changed to L2.

3. BGP programs the line cards with an MPLS forwarding entry that swaps L2 for L1 and sends the MPLS packet over the CSC VPRN interface associated with next-hop N1.

When a 7x50 CSC-PE receives a VPN-IPv4 route (with label L2, next-hop N2) the following actions take place in the CSC-PE:

1. If the VPN-IPv4 route matches the VRF import policy of a CSC VPRN, it is installed into the routing table of that CSC VPRN.
2. If the imported (BGP-VPN) route matches the BGP export policy associated with a CSC-CE BGP peer, it is advertised to that peer as a labeled-IPv4 route. The advertised label value is changed to L3.
3. BGP programs the line cards with an MPLS forwarding entry that swaps L3 for L2 and sends the packet inside the MPLS tunnel to next-hop N2.

Once a CSC-CE has learned a labeled-IPv4 route for a remote CSC-CE and vice versa, the two CSC-CEs can set up a BGP session between themselves and exchange VPN routes over this session if they are both PEs with services. Typically this BGP session will be an IBGP session because the local and remote CSC-CEs belong to the same Autonomous System (AS). The Layer 2 VPN and Layer 3 VPN routes exchanged by the CSC-CEs are resolved by the labeled-IPv4 routes they have for each other’s /32 IPv4 address.

Configuration

This section will walk through the steps to configure the CSC solution shown in Figure 158. The IPv4 addresses in Figure 158 are the system IP addresses of the routers. The steps are the following:

- Configure CSC-CE-1
- Configure CSC service on CSC-PE-2
- Verify exchange of labeled IPv4 routes between CSC-CE-1 and CSC-PE-2
- Configure core connectivity for CSC-PE-2
- Configure core connectivity for CSC-PE-3
- Configure CSC service on CSC-PE-3
- Verify exchange of VPN-IPv4 routes between CSC-PE-2 and CSC-PE-3
- Configure CSC-CE-4
- Verify exchange of labeled IPv4 routes between CSC-PE-3 and CSC-CE-4
- Configure BGP session between CSC-CE-1 and CSC-CE-4
- Verify exchange of VPN-IPv4 routes between CSC-CE-1 and CSC-CE-4

Step 1. Configure CSC-CE-1

This example assumes that CSC-CE-1 is a PE router with Layer 2 and Layer 3 VPN services that must extend across the CSC VPN service; assume that there are no further downstream PEs in AS 64496. The configuration of one such Layer 3 VPN service in CSC-CE-1 is as follows:

```
# on CSC-CE-1
configure
    service
        vprn 1 customer 1 create
            route-distinguisher 64496:11
            auto-bind-tunnel
                resolution any
            exit
            vrf-target target:64496:1
            interface "loopback-1" create
                address 10.11.30.2/24
                loopback
            exit
            --- snipped ---
```

For brevity, the above configuration sample omits commands related to SAP IP interfaces, spoke-SDP IP interfaces, PE-CE routing protocols, QoS, IP filters, etc. The loopback interface is used to test whether this prefix will be learned at the remote CSC-CE-4.

The base routing instance of the CSC-CE should be configured with the appropriate router-ID and autonomous-system number and the system interface should be given an IPv4 address (usually the same as the router-id). If the router-ID is not configured, by default the system IP address is used as the router-ID. The interface to CSC-PE-2 should then be created and configured. The base router configuration of CSC-CE-1 is as follows:

```
# on CSC-CE-1
configure
    router
        interface "int-CSC-CE-1-CSC-PE-2"
            address 192.168.12.1/30
            port 1/1/1
            no shutdown
        exit
        interface "system"
            address 192.0.2.1/32
            no shutdown
        exit
        autonomous-system 64496
        --- snipped ---
    exit
```

BGP should be configured as the control plane protocol running on the interface to CSC-PE-2, as follows:

```
# on CSC-CE-1
configure
    router
        bgp
            group "CSC-PE"
                peer-as 64500
                neighbor 192.168.12.2
                    family label-ipv4
                    export "static-to-BGP"
                    split-horizon
                exit
            exit
            no shutdown
        exit
    exit
```

The peer type is EBGP (**peer-as** is different from the locally configured **autonomous-system)**

The address family for the EBGP session is **label-ipv4** (the **neighbor** address is an IPv4 address). Family label-IPv4 causes MP-BGP negotiation of the address family for AFI=1 and SAFI=4 (IPv4 NLRI with MPLS labels), as can be observed from the following debug trace (using the command **debug router bgp open**) of the OPEN message from CSC-CE-1.This message can obviously only be seen when the BGP peer is up. The configuration for CSC-PE-2 will be shown later, but in order to have the trace message, it must be configured already.

```
2 2017/10/19 07:38:09.783 UTC MINOR: DEBUG #2001 Base BGP
"BGP: OPEN
Peer 1: 192.168.12.2 - Received BGP OPEN: Version 4
   AS Num 64500: Holdtime 90: BGP_ID 192.0.2.2: Opt Length 16
   Opt Para: Type CAPABILITY: Length = 14: Data:
     Cap_Code MP-BGP: Length 4
       Bytes: 0x0 0x1 0x0 0x4
     Cap_Code ROUTE-REFRESH: Length 0
     Cap_Code 4-OCTET-ASN: Length 4
       Bytes: 0x0 0x0 0xfb 0xf4
"
```

The **split-horizon** command is optional. It prevents a best BGP route from the CSC-PE peer from being re-advertised back to that peer.

The **export** command applies a BGP export policy to the session.

The configuration of the policy is as follows:

```
# on CSC-CE-1
configure
    router
        policy-options
            begin
            prefix-list "system-IP"
                prefix 192.0.2.1/32 exact
            exit
            policy-statement "static-to-BGP"
                entry 10
                    from
                        protocol direct
                        prefix-list "system-IP"
                    exit
                    action accept
                    exit
                exit
                default-action drop
                exit
            exit
            commit
        exit
```

The purpose of the BGP export policy is to advertise the system IP address of CSC-CE-1 as a labeled-IPv4 BGP route toward the CSC-PE(s).

Step 2. Configure CSC service on SCS-PE-2

CSC-PE-2 must be configured with a VPRN in **carrier-carrier-vpn** mode in order to provide CSC service to CSC-CE-1. The entire configuration of the VPRN is shown below:

```
# on CSC-PE-2
configure
    service
        vprn 1 customer 1 create
```

```
            carrier-carrier-vpn
            router-id 192.0.2.2
            autonomous-system 64500
            route-distinguisher 64500:12
            auto-bind-tunnel
                resolution any
            exit
            vrf-target target:64500:1
            network-interface "int-CSC-PE-2-CSC-CE-1" create
                address 192.168.12.2/30
                port 1/1/2
                no shutdown
            exit
            bgp
                group "CSC-CE"
                    as-override
                    export "BGP-VPN-routes"
                    peer-as 64496
                    neighbor 192.168.12.1
                        family label-ipv4
                        split-horizon
                    exit
                exit
                no shutdown
            exit
            no shutdown
        exit
    exit
```

The **carrier-carrier-vpn** command is mandatory. It cannot be configured if the VPRN currently has any SAP or spoke-SDP access interfaces configured; they must first be shutdown if necessary and then deleted.

```
*A:CSC-PE-2# configure service vprn 1 carrier-carrier-vpn
INFO: PIP #1195 Cannot toggle carrier-carrier-vpn - service interfaces present
*A:CSC-PE-2#
```

The **auto-bind**-tunnel command should be set appropriately for the type of transport desired to other CSC-PEs, but note that GRE is not supported.

```
*A:CSC-PE-2# configure service vprn 1 auto-bind-tunnel resolution-filter gre
MINOR: SVCMGR #1538 auto-bind config not supported - Autobind gre not supported for
carrier-carrier vprn
*A:CSC-PE-2#
```

The interface to CSC-CE-1 must be a **network-interface**. A **network-interface** can be associated with an entire Ethernet port (as shown in the previous example), a VLAN sub-interface of an Ethernet port, an entire LAG or a VLAN sub-interface of a LAG. In all cases, the associated Ethernet ports must be configured in network or hybrid mode and must reside on FP2 or higher based cards/systems.

The peer type is EBGP (**peer-as** is different from the locally configured **autonomous-system**).

The address family for the EBGP session is **label-ipv4** (the **neighbor** address is an IPv4 address). Address family label-ipv4 causes MP-BGP negotiation of the address family for AFI=1 and SAFI=4 (IPv4 NLRI with MPLS labels).

The **split-horizon** command is optional. It prevents a best BGP route from the CSC-CE peer from being re-advertised back to that peer.

The **as-override** command replaces CSC-CE-1’s AS number (64496) with CSC-PE-2’s AS number (64500) in the AS_PATH attribute of routes advertised to CSC-CE-1. Without this configuration, CSC-CE-1 may reject routes originated by CSC-CE-4 as invalid due to an AS-path loop.

The **export** command applies a BGP export policy to the session. The configuration of the policy is as follows:

```
# on CSC-PE-2
configure
    router
        policy-options
            begin
            policy-statement "BGP-VPN-routes"
                entry 10
                    from
                        protocol bgp-vpn
                    exit
                    action accept
                    exit
                exit
                default-action drop
                exit
            exit
            commit
        exit
    exit
```

The effect of the BGP export policy is to re-advertise VPN-IPv4 routes imported into the CSC VPRN (and used for forwarding) to CSC-CE-4.

Step 3. Verify exchange of labeled IPV4 routes

When steps 1 and 2 have been completed properly, CSC-CE-1 should be advertising the labeled-IPv4 route for its system IP address to CSC-PE-2. This can be checked from the perspective of CSC-CE-1 as follows:

```
*A:CSC-CE-1# show router bgp routes 192.0.2.1/32 label-ipv4 hunt
================================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
================================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
```

```
===============================================================================
BGP Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
Network        : 192.0.2.1/32
Nexthop        : 192.168.12.1
Path Id        : None
To             : 192.168.12.2
Res. Nexthop   : n/a
Local Pref.    : n/a                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : No Community Members
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.2
IPv4 Label     : 262142                 Label Type     : POP
Origin         : IGP
AS-Path        : 64496
Route Tag      : 0
Neighbor-AS    : 64496
Orig Validation: NotFound
Source Class   : 0                      Dest Class     : 0

-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:CSC-CE-1#
```

CSC-CE-1 has advertised a label value of 262142 with the prefix.

The following output shows the received route from the perspective of CSC-PE-2:

```
*A:CSC-PE-2# show router 1 bgp routes 192.0.2.1/32 label-ipv4 hunt
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : 192.0.2.1/32
Nexthop        : 192.168.12.1
Path Id        : None
```

```
From           : 192.168.12.1
Res. Nexthop   : 192.168.12.1
Local Pref.    : None                   Interface Name : int-CSC-PE-2-CSC-CE-1
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : No Community Members
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.1
Fwd Class      : None                   Priority       : None
IPv4 Label     : 262142                 Label Type     : SWAP
Flags          : Used  Valid  Best  IGP
Route Source   : External
AS-Path        : 64496
Route Tag      : 0
Neighbor-AS    : 64496
Orig Validation: NotFound
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h02m29s

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:CSC-PE-2#
```

Step 4. Configure core connectivity for CSC-PE-2

The next step is to configure the base router instance of CSC-PE-2 so that it can exchange VPN-IPv4 routes with CSC-PE-3 (and potentially other CSC-PEs). At a minimum this requires:

- Router-id and autonomous-system configuration.
- Network interface creation and configuration, including assignment of an IPv4 address to the system interface.
- Configuration of the IGP protocol. In this example, IS-IS is used.
- Configuration of the LDP protocol (optional).
- Configuration of RSVP LSPs used to reach remote CSC-PE devices (optional).
- Configuration of the BGP protocol.

The base router configuration of CSC-PE-2 is as follows:

```
# on CSC-PE-2
configure
    router
        interface "int-CSC-PE-2-CSC-PE-3"
            address 192.168.23.1/30
            port 1/1/1
            no shutdown
        exit
        interface "system"
            address 192.0.2.2/32
            no shutdown
        exit
        autonomous-system 64500
        isis 0
            level-capability level-2
            area-id 49.01
            level 2
                wide-metrics-only
            exit
            interface "system"
                passive
                no shutdown
            exit
            interface "int-CSC-PE-2-CSC-PE-3"
                interface-type point-to-point
                no shutdown
            exit
            no shutdown
        exit
        ldp
            interface-parameters
                interface "int-CSC-PE-2-CSC-PE-3" dual-stack
                    ipv4
                        no shutdown
                    exit
                    no shutdown
                exit
            exit
            targeted-session
            exit
            no shutdown
        exit
        bgp
            group "core"
                type internal
                neighbor 192.0.2.3
                    family vpn-ipv4
                exit
            exit
            no shutdown
        exit
```

The peer type is IBGP (**type internal**. It is also possible to configure this in a similar way as for eBGP, with the same value for **peer-as** as the locally configured **autonomous-system**).

The transport for the IBGP session is IPv4 (the **neighbor** address is an IPv4 address).

The **family vpn-ipv4** command causes MP-BGP negotiation of the address family for AFI=1 and SAFI=128 (=0x80), as can be observed from the following debug trace of the OPEN message from CSC-PE-2.

```
1 2017/10/19 07:41:51.579 UTC MINOR: DEBUG #2001 Base BGP
"BGP: OPEN
Peer 1: 192.0.2.3 - Send (Passive) BGP OPEN: Version 4
   AS Num 64500: Holdtime 90: BGP_ID 192.0.2.2: Opt Length 16
   Opt Para: Type CAPABILITY: Length = 14: Data:
     Cap_Code MP-BGP: Length 4
       Bytes: 0x0 0x1 0x0 0x80
     Cap_Code ROUTE-REFRESH: Length 0
     Cap_Code 4-OCTET-ASN: Length 4
       Bytes: 0x0 0x0 0xfb 0xf4
"
```

Step 5. Configure core connectivity for CSC-PE-3

The next step is to configure the base router instance of CSC-PE-3 so that it can exchange VPN-IPv4 routes with CSC-PE-2 and potentially other CSC-PEs. At a minimum, this requires:

- Router-id and autonomous-system configuration.
- Network interface creation and configuration, including assignment of an IPv4 address to the system interface.
- Configuration of the IGP protocol. In this example IS-IS is used.
- Configuration of the LDP protocol (optional).
- Configuration of RSVP LSPs used to reach remote CSC-PE devices (optional).
- Configuration of the BGP protocol.

The base router configuration of CSC-PE-3 then is as follows:

```
# on CSC-PE-3
configure
    router
        interface "int-CSC-PE-3-CSC-PE-2"
            address 192.168.23.2/30
            port 1/1/2
            no shutdown
        exit
        interface "system"
            address 192.0.2.3/32
            no shutdown
        exit
        autonomous-system 64500
        isis 0
            level-capability level-2
            area-id 49.01
            level 2
                wide-metrics-only
            exit
            interface "system"
                passive
                no shutdown
            exit
            interface "int-CSC-PE-3-CSC-PE-2"
                interface-type point-to-point
                no shutdown
            exit
            no shutdown
        exit
        ldp
            interface-parameters
                interface "int-CSC-PE-3-CSC-PE-2" dual-stack
                    ipv4
                        no shutdown
                    exit
                    no shutdown
                exit
            exit
            targeted-session
            exit
            no shutdown
        exit
        bgp
            group "core"
                type internal
                cluster 192.0.2.3
                neighbor 192.0.2.2
                    family vpn-ipv4
                    split-horizon
                exit
            exit
            no shutdown
        exit
```

The peer type is IBGP (**type internal**. If the configuration would include **peer-as** instead, the value would be the same as the locally configured **autonomous-system**).

The transport for the IBGP session is IPv4 (the **neighbor** address is an IPv4 address).

The **family vpn-ipv4** command causes MP-BGP negotiation of the address family for AFI=1 and SAFI=128.

The **cluster** command configures CSC-PE-2 as a route reflector for clients in the BGP group called “core”. This is not required and in a more typical deployment, the route reflector would be a separate router from any CSC-PE.

Step 6. Configure CSC service on CSC-PE-3

CSC-PE-3 must be configured with a VPRN in **carrier-carrier-vpn** mode in order to provide CSC service to CSC-CE-4. The entire configuration of the VPRN is shown below:

```
# on CSC-PE-3
configure
    service
        vprn 1 customer 1 create
            carrier-carrier-vpn
```

```
                router-id 192.0.2.3
                autonomous-system 64500
                route-distinguisher 64500:13
                auto-bind-tunnel
                    resolution any
                exit
                vrf-target target:64500:1
                network-interface "int-CSC-PE-3-CSC-CE-4" create
                    address 192.168.34.1/30
                    port 1/1/1
                    no shutdown
                exit
                bgp
                    group "CSC-CE"
                        as-override
                        export "BGP-VPN-routes"
                        peer-as 64496
                        neighbor 192.168.34.2
                            family label-ipv4
                            split-horizon
                        exit
                    exit
                    no shutdown
                exit
                no shutdown
            exit
        exit
```

The **carrier-carrier-vpn** command is mandatory. It cannot be configured if the VPRN currently has any SAP or spoke-SDP “access” interfaces configured; they must first be shutdown if necessary and then deleted.

The **auto-bind-tunnel** command should be set appropriately for the type of transport desired to other CSC-PEs, but GRE is not supported.

The interface to CSC-CE-4 must be a **network-interface**. A **network-interface** can be associated with an entire Ethernet port (as shown in the example above), a VLAN sub-interface of an Ethernet port, an entire LAG or a VLAN sub-interface of a LAG. In all cases, the associated Ethernet ports must be configured in network or hybrid mode and must reside on FP2 or higher based cards/systems.

The peer type is EBGP (**peer-as** is different from the locally configured **autonomous-system**).

The address family for the EBGP session is **label-ipv4** (the **neighbor** address is an IPv4 address). Address family label-ipv4 causes MP-BGP negotiation of the address family for AFI=1 and SAFI=4 (IPv4 NLRI with MPLS labels).

The **split-horizon** command is optional. It prevents a best BGP route from the CSC-CE peer from being re-advertised back to that peer.

The **as-override** command replaces CSC-CE-4’s AS number (64496) with CSC-PE-3’s AS number (64500) in the AS_PATH attribute of routes advertised to CSC-CE-4. Without this configuration, CSC-CE-4 may reject routes originated by CSC-CE-1 as invalid due to an AS-path loop.

The **export** command applies a BGP export policy to the session. The configuration of the policy is as follows:

```
# on CSC-PE-3
configure
    router
        policy-options
            begin
            policy-statement "BGP-VPN-routes"
                entry 10
                    from
                        protocol bgp-vpn
                    exit
                    action accept
                    exit
                exit
                default-action drop
                exit
            exit
            commit
        exit
    exit
```

The effect of the BGP export policy is to re-advertise VPN-IPv4 routes imported into the CSC VPRN (and used for forwarding) to CSC-CE-4.

Step 7. Verify exchange of VPN-IPv4 routes between CSC-PE-2 and CSC-PE-3.

When the preceding steps have been completed properly, CSC-PE-2 should now be advertising the labeled-IPv4 route for 192.0.2.1/32 (the system IP address of CSC-CE-1) to CSC-PE-3. This can be checked from the perspective of CSC-PE-2 as follows:

```
*A:CSC-PE-2# show router bgp routes 192.0.2.1/32 vpn-ipv4 hunt
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
```

```
-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
Network        : 192.0.2.1/32
Nexthop        : 192.0.2.2
Route Dist.    : 64500:12               VPN Label      : 262140
Path Id        : None
To             : 192.0.2.3
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : target:64500:1
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.3
Origin         : IGP
AS-Path        : 64496
Route Tag      : 0
Neighbor-AS    : 64496
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0

-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:CSC-PE-2#
```

CSC-PE-2 has advertised a label value of 262140 with the prefix.

The following output shows the received route from the perspective of CSC-PE-3:

```
*A:CSC-PE-3# show router bgp routes 192.0.2.1/32 vpn-ipv4 hunt
===============================================================================
 BGP Router ID:192.0.2.3        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : 192.0.2.1/32
Nexthop        : 192.0.2.2
Route Dist.    : 64500:12               VPN Label      : 262140
Path Id        : None
From           : 192.0.2.2
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : int-CSC-PE-3-CSC-PE-2
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
```

```
Connector       : None
Community       : target:64500:1
Cluster         : No Cluster Members
Originator Id   : None                    Peer Router Id : 192.0.2.2
Fwd Class       : None                    Priority       : None
Flags           : Used  Valid  Best  IGP
Route Source    : Internal
AS-Path         : 64496
Route Tag       : 0
Neighbor-AS     : 64496
Orig Validation: N/A
Source Class    : 0                       Dest Class     : 0
Add Paths Send : Default
Last Modified   : 00h00m51s
VPRN Imported   :  1

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:CSC-PE-3#
```

The label swap entries that BGP programmed in the line cards of CSC-PE-2 based on the received labeled-IPv4 route from CSC-CE-1 (Label Origin = ExtCarCarVpn) and the advertised VPN-IPv4 route to CSC-PE-3, as follows:

```
*A:CSC-PE-2# show router bgp inter-as-label

===============================================================================
BGP Inter-AS labels
Flags: B - entry has backup, P - entry is promoted
===============================================================================
NextHop                       Received       Advertised     Label
                              Label          Label          Origin
-------------------------------------------------------------------------------
192.168.12.1                  262142         262140         ExtCarCarVpn
-------------------------------------------------------------------------------
Total Labels allocated:   1
===============================================================================
*A:CSC-PE-2#
```

Step 8. Configure CSC-CE-4

This chapter assumes that CSC-CE-4 is a PE router with Layer 2 and Layer 3 VPN services that must extend across the CSC VPN service. The configuration of one such Layer 3 VPN service in CSC-CE-4 is as follows:

```
# on CSC-CE-4
configure
    service
        vprn 1 customer 1 create
            route-distinguisher 64496:14
            auto-bind-tunnel
                resolution any
            exit
            vrf-target target:64496:1
            interface "loopback-1" create
                address 10.14.30.2/24
                loopback
            exit
            ---snip---
        exit
```

For brevity, the preceding configuration sample omits commands related to SAP IP interfaces, spoke-SDP IP interfaces, PE-CE routing protocols, QoS, IP filters, and so on.

The base routing instance of CSC-CE-4 should be configured with the appropriate router ID and autonomous system number and the system interface should be given an IPv4 address (usually the same as the router-id). The interface to CSC-PE-3 should then be created and configured. The base router configuration of CSC-CE-4 is as follows:

```
# on CSC-CE-4
configure
    router
        interface "int-CSC-CE-4-CSC-PE-3"
            address 192.168.34.2/30
            port 1/1/2
            no shutdown
        exit
        interface "system"
            address 192.0.2.4/32
            no shutdown
        exit
        autonomous-system 64496
        --- snipped ---
    exit
```

BGP should be configured as the control plane protocol running on the interface to CSC-PE-4 as follows:

```
# on CSC-CE-4
configure
    router
        bgp
            group "CSC-PE"
                peer-as 64500
                neighbor 192.168.34.1
                    family label-ipv4
                    export "static-to-BGP"
                    split-horizon
                exit
            exit
            no shutdown
        exit
```

The peer type is EBGP (**peer-as** is different from the locally configured **autonomous-system**).

The address family for the EBGP session is **label-ipv4** (the **neighbor** address is an IPv4 address). Address family label-ipv4 causes MP-BGP negotiation of the address family for AFI=1 and SAFI=4 (IPv4 NLRI with MPLS labels)

The **split-horizon** command is optional. It prevents a best BGP route from the CSC-PE peer from being re-advertised back to that peer.

The **export** command applies a BGP export policy to the session. The configuration of the policy is shown below:

```
# on CSC-CE-4
configure
    router
        policy-options
            begin
            prefix-list "system-IP"
                prefix 192.0.2.4/32 exact
            exit
            policy-statement "static-to-BGP"
                entry 10
                    from
                        protocol direct
                        prefix-list "system-IP"
                    exit
                    action accept
                    exit
                exit
                default-action drop
                exit
            exit
            commit
        exit
```

The purpose of the BGP export policy is to advertise the system IP address of CSC-CE-4 as a labeled-IPv4 BGP route toward CSC-PE-3.

Step 9. Verify exchange of labeled IPv4 routes between CSC-PE-3 and CSC-CE-4

When the preceding steps have been completed, CSC-PE-3 should be advertising the labeled-IPv4 route for 192.0.2.1/32 to CSC-CE-4. This can be checked from the perspective of CSC-PE-3, as follows:

```
*A:CSC-PE-3# show router 1 bgp routes 192.0.2.1/32 label-ipv4 hunt
===============================================================================
 BGP Router ID:192.0.2.3        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
```

```
===============================================================================
BGP Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
Network        : 192.0.2.1/32
Nexthop        : 192.168.34.1
Path Id        : None
To             : 192.168.34.2
Res. Nexthop   : n/a
Local Pref.    : n/a                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : target:64500:1
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.4
IPv4 Label     : 262140                 Label Type     : SWAP
Origin         : IGP
AS-Path        : 64500 64500
Route Tag      : 0
Neighbor-AS    : 64500
Orig Validation: NotFound
Source Class   : 0                      Dest Class     : 0

-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:CSC-PE-3#
```

CSC-PE-3 has advertised a label value of 262140 with the prefix.

The following output shows the received route from the perspective of CSC-CE-4:

```
*A:CSC-CE-4# show router bgp routes 192.0.2.1/32 label-ipv4 hunt
===============================================================================
 BGP Router ID:192.0.2.4        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : 192.0.2.1/32
Nexthop        : 192.168.34.1
```

```
Path Id        : None
From           : 192.168.34.1
Res. Nexthop   : 192.168.34.1
Local Pref.    : None                   Interface Name : int-CSC-CE-4-CSC-PE-3
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : target:64500:1
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.3
Fwd Class      : None                   Priority       : None
IPv4 Label     : 262140                 Label Type     : SWAP
Flags          : Used  Valid  Best  IGP
Route Source   : External
AS-Path        : 64500 64500
Route Tag      : 0
Neighbor-AS    : 64500
Orig Validation: NotFound
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h00m53s

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:CSC-CE-4#
```

The BGP distributed labels are programmed in the line cards of CSC-PE-3 based on the received VPN-IPv4 routes from CSC-PE-2 (Label Origin = Internal) and the advertised labeled-IPv4 routes to CSC-CE-4:

```
*A:CSC-PE-3# show router 1 bgp inter-as-label

===============================================================================
BGP Inter-AS labels
Flags: B - entry has backup, P - entry is promoted
===============================================================================
NextHop                       Received       Advertised     Label
                              Label          Label          Origin
-------------------------------------------------------------------------------
192.0.2.2                     262140         262140         Internal
192.0.2.2                     262141         262139         Internal
-------------------------------------------------------------------------------
Total Labels allocated:   2
===============================================================================
*A:CSC-PE-3#
```

In the preceding output, the first entry for NextHop 192.0.2.2 corresponds to the prefix 192.0.2.1/32; recall from Step 7 that CSC-PE-3 received the VPN-IPv4 route with label value 262140 and it can be seen from this step that it re-advertised the route to CSC-CE-4 with the same label value 262140.

Step 10. Configure BGP session between CSC-CE-1 and CSC-CE-4

The final step in the setup of the CSC solution shown in Figure 158 is the creation of a BGP session between CSC-CE-1 and CSC-CE-4 so that they can exchange routes belonging to VPN services they support. The configuration of this BGP session from the perspective of CSC-CE-1 is as follows:

```
# on CSC-CE-1
configure
    router
        bgp
            group "CSC-CE"
                type internal
                neighbor 192.0.2.4
                    family vpn-ipv4
                exit
            exit
            no shutdown
        exit
    exit
exit
```

The configuration of the BGP session from the perspective of CSC-CE-4 is very similar, as shown below.

```
# on CSC-CE-4
configure
    router
        bgp
            group "CSC-CE"
                type internal
                neighbor 192.0.2.1
                    family vpn-ipv4
                exit
            exit
            no shutdown
        exit
    exit
exit
```

The configuration of the BGP session between CSC-CE-1 and CSC-CE-4 has the following properties:

- The peer type is IBGP (**type internal**. Alternatively, **peer-as** can be configured with the same value as the locally configured **autonomous-system**).
- The transport for the IBGP session is IPv4 (the **neighbor** address is an IPv4 address).
- The **family vpn-ipv4** command causes MP-BGP negotiation of the address family for AFI=1 and SAFI=128.

Step 11. Verify exchange of VPN-IPv4 routes

When the preceding steps have been completed properly, CSC-PE-3 should now be able to advertise a VPN-IPv4 route for some IP prefix (for example, 10.11.30.0/24) to CSC-CE-4. This can be checked from the perspective of CSC-CE-4 as follows:

```
*A:CSC-CE-4# show router bgp routes 10.11.30.0/24 vpn-ipv4 hunt
===============================================================================
 BGP Router ID:192.0.2.4        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : 10.11.30.0/24
Nexthop        : 192.0.2.1
Route Dist.    : 64496:11               VPN Label      : 262143
Path Id        : None
From           : 192.0.2.1
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : target:64496:1
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.1
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h00m04s
VPRN Imported  :  1

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:CSC-CE-4#
```

It is also possible to check that CSC-CE-4 has properly installed the preceding VPN-IPv4 route into the routing table of the importing VPRN service, as follows:

```
*A:CSC-CE-4# show router 1 route-table
```

```
===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.11.30.0/24                                 Remote  BGP VPN   00h00m45s  170
       192.0.2.1 (tunneled:BGP)                                     0
10.14.30.0/24                                 Local   Local     00h03m51s  0
       loopback-1                                                   0
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:CSC-CE-4#
```

Conclusion

Carrier Supporting Carrier is a scalable and secure solution for using an infrastructure IP VPN to transport traffic between dispersed CSC-CE devices belonging to an ISP or other service provider. Many different topology models are supported by SR OS. This chapter has explored one simplified configuration that can serve as the basis for more complicated setups.

Layer 3 VPN: VPRN Type Spoke

This chapter provides information about Layer 3 VPRN CE hub and spoke architecture.

Topics in this chapter include:

- Applicability
- Summary
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SR OS release 12.0. However, the CLI in the current edition is based on release 14.0.R5.

Summary

This chapter provides a basic technology overview and configuration examples of a network topology used for a Layer 3 VPRN CE hub and spoke architecture.

Knowledge of Nokia’s Layer 3 VPN concepts is assumed throughout this document.

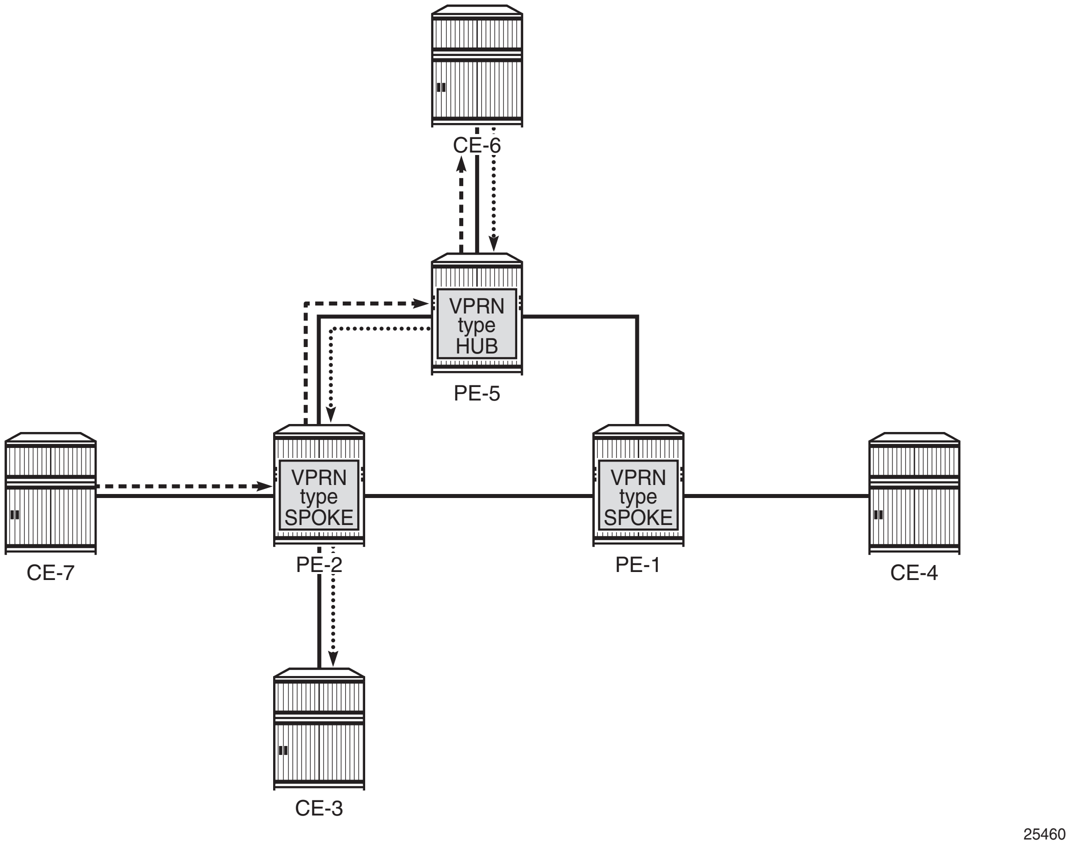
Overview

In SR OS releases earlier than 12.0, a **CE hub and spoke** architecture was partially supported. Internal optimization was available for the hub sites connected to the same PE router only. This feature is known as VPRN **type hub**. If, on the other hand, multiple spoke sites were connected to the same PE router, separate VPRN instances had to be created to maintain the split horizon forwarding behavior. This approach was complex, hard to maintain and consumed extra VPRN instances.

Release 12.0.R1 added new functionality to overcome these limitations. Introducing the VPRN **type spoke** feature allows multiple spoke sites to be kept within the same VPRN instance while at the same time maintaining the split horizon approach such that spoke sites cannot send traffic directly to each other.

The primary goal of the feature is to allow multiple spoke sites to be part of a single VPRN instance without allowing direct communication between the spoke CE sites which are part of that VPRN (of type spoke). The packet flow is demonstrated in Figure 159.

Figure 159 **CE Hub and Spoke Data Path**



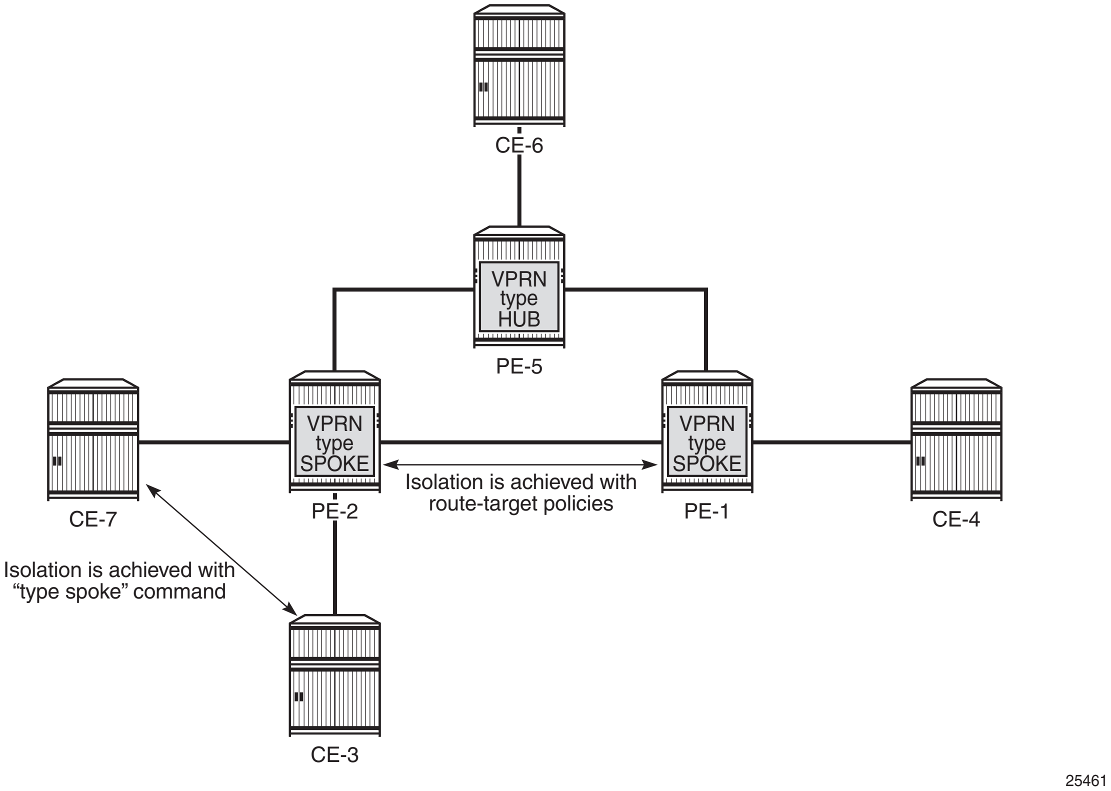
The only way for CE-7 to communicate with CE-3 is via hub site CE-6. The same applies to the communication between CE-7 and CE-4. The VPRN on PE-2 is configured as **type spoke** and has IP interfaces using SAPs or spoke SDPs that are considered spoke sites only. No direct communication between any of the spoke CE sites in the network is allowed.

This is achieved using two techniques (Figure 160).

- Use the **type spoke** command under the VPRN context as explained later.

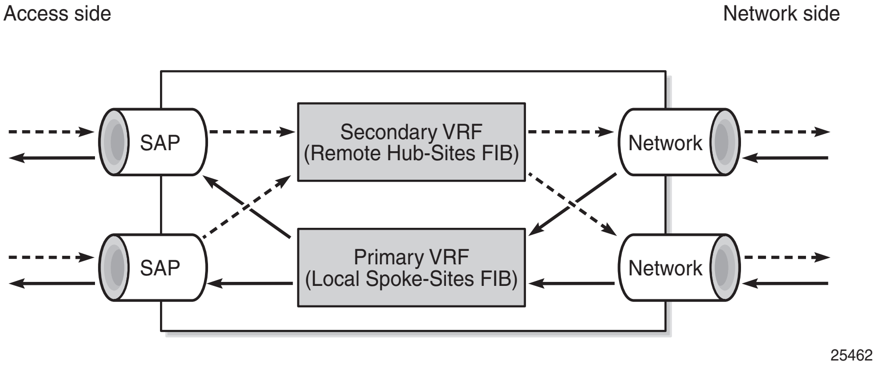
- The extended community configuration using route-target policies (this is not covered in detail in this chapter).

Figure 160 **CE Hub and Spoke Control Plane Isolation**


CE-6
VPRN type HUB
PE-5
VPRN type SPOKE
PE-2
VPRN type SPOKE
PE-1
CE-7
CE-4
Isolation is achieved with route-target policies
Isolation is achieved with “type spoke” command
CE-3
25461

When a VPRN on a PE router is configured as **type spoke,** then the internal forwarding logic changes as demonstrated in Figure 161.

Figure 161 **Internal VPRN Logic on a PE Router**

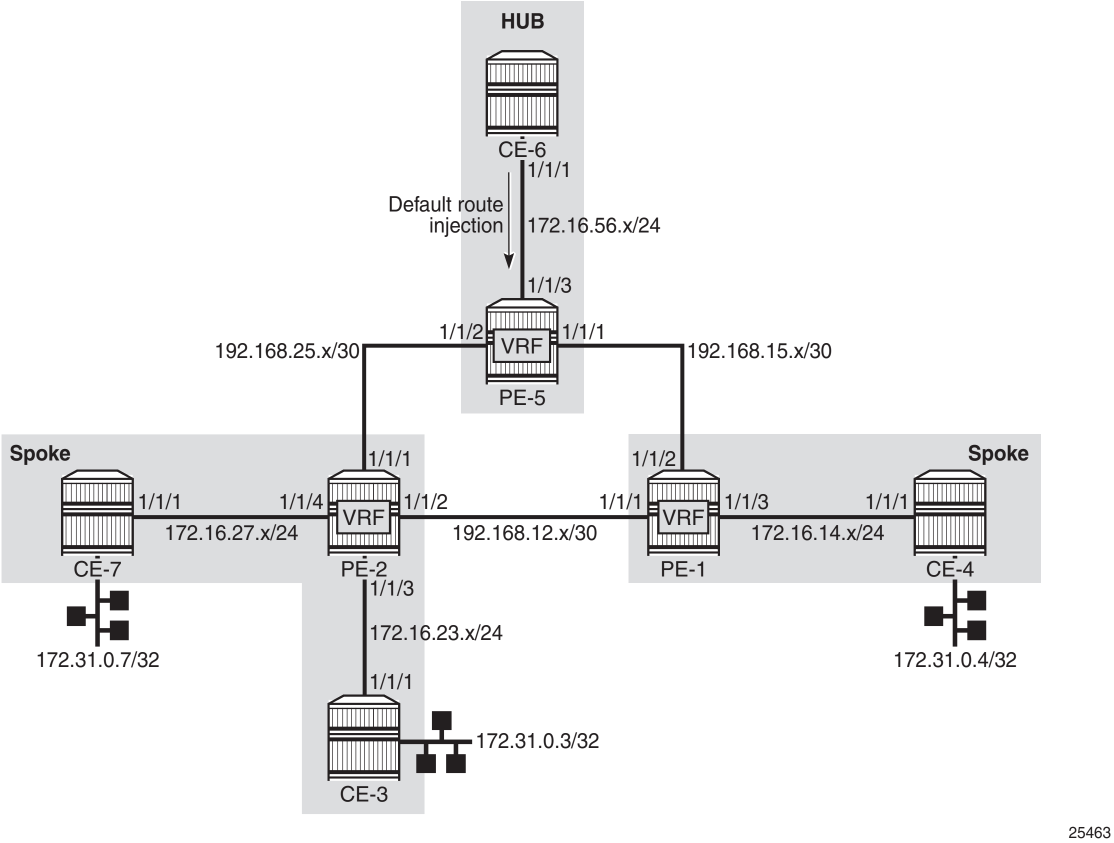

Access side
Network side
SAP
Secondary VRF (Remote Hub-Sites FIB)
Network
SAP
Primary VRF (Local Spoke-Sites FIB)
Network
25462

- VPRNs of type spoke create a primary and a secondary VRF internally to the VPRN:
  - The primary VRF is used for forwarding traffic from the network interfaces toward the IP interfaces using SAPs or spoke SDPs. This VRF is populated with routes learned from the spoke CE sites connected to the local PE through IP interfaces using SAPs or spoke SDPs.
  - The secondary VRF is used for forwarding traffic from the IP interfaces using SAPs or spoke SDPs toward the network interfaces or other VPRN instances. This VRF is populated with routes learned via MP-BGP from Hub sites.
- VPRNs of type spoke export routes using a specific extended community (for instance spoke-ext-comm) via an export policy to identify them as spoke site originated routes.
  - This community is not hard-coded and has to be configured manually (see configuration example later).
- VPRNs of type spoke import routes (using an import policy) received from other PEs or VPRN instances with a hub specific community only (for example hub-ext-comm). Routes with spoke-ext-comm community are ignored.
  - This community is not hard-coded and has to be configured manually (see configuration example later).
- Multiple VPRNs of type spoke and hub can coexist on the same PE if they use different VPRN instances.
- The configuration of type hub and type spoke is mutually exclusive within one VPRN instance.

Configuration

The physical topology and addressing scheme are presented in Figure 162.

Figure 162 **CE Hub and Spoke Topology and Addressing Scheme**



The configuration of PE-2 and PE-5 are the main focus of this example. The configuration of PE-1 is similar to that of PE-2.

Hub Site Configuration

Only the essential part of the configuration is provided for the hub site.

Vrf-import and export policies are used to manipulate the vrf-target in order to achieve logical isolation between the spoke sites in the network.

```
*A:PE-5# configure
    router
        policy-options
            begin
            community "hub-ext-comm" members "target:64500:11"
            community "spoke-ext-comm" members "target:64500:12"
            policy-statement "vrf-export"
```

```
                default-action accept
                    community add "hub-ext-comm"
                exit
            exit
            policy-statement "vrf-import"
                entry 10
                    from
                        community "spoke-ext-comm"
                    exit
                    action accept
                    exit
                exit
                default-action drop
                exit
            exit
            policy-statement "export-ospf"
                entry 10
                    from
                        protocol direct
                    exit
                    action accept
                    exit
                exit
                default-action accept
                exit
            exit
            commit
        exit
```

PE-5 is configured with VPRN 1 providing OSPF connectivity to customer CE-6.

```
*A:PE-5# configure
    service
        vprn 1 customer 1 create
            vrf-import "vrf-import"
            vrf-export "vrf-export"
            route-distinguisher 64500:15
            type hub
            auto-bind-tunnel
                resolution any
            exit
            interface "int-PE-5-CE-6" create
                address 172.16.56.1/24
                sap 1/1/3:1 create
                exit
            exit
            ospf
                export "export-ospf"
                area 0.0.0.0
                    interface "int-PE-5-CE-6"
                        interface-type point-to-point
                        mtu 1500
                        no shutdown
                    exit
                exit
                no shutdown
            exit
            no shutdown
        exit
```

At the same time, CE-6 is configured to advertise a default route which is used by all remote spoke CE sites to forward traffic via CE-6.

```
*A:CE-6# configure
    router
        policy-options
            begin
            policy-statement "export-ospf-default"
                entry 10
                    from
                        protocol static
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit
```

```
*A:CE-6# configure
    service
        vprn 1 customer 1 create
            route-distinguisher 64500:16
            interface "int-CE-6-PE-5" create
                address 172.16.56.2/24
                sap 1/1/1:1 create
                exit
            exit
            interface "lo0" create
                shutdown
                address 172.31.0.6/32
                loopback
            exit
            static-route-entry 0.0.0.0/0
                black-hole
                    no shutdown
                exit
            exit
            ospf 192.0.2.6
                export "export-ospf-default"
                ignore-dn-bit
                suppress-dn-bit
                area 0.0.0.0
                    interface "int-CE-6-PE-5"
                        interface-type point-to-point
                        mtu 1500
                        no shutdown
                    exit
                    interface "lo0"
                        no shutdown
                    exit
                exit
                no shutdown
            exit
            no shutdown
        exit
```

Spoke Site Configuration

According to the example topology, two spoke VPRNs are present: one VPRN with two CE spoke sites connected is located on PE-2, and another VPRN with one spoke CE site is located on PE-1. The service configuration for PE-2 is as follows with the one for PE-1 being similar.

Vrf-import and export policies are used to build a hub-and-spoke topology in order to achieve a logical isolation between spoke sites connected to different PE routers.

```
*A:PE-2# configure
    router
        policy-options
            begin
            community "hub-ext-comm" members "target:64500:11"
            community "spoke-ext-comm" members "target:64500:12"
            policy-statement "vrf-export"
                default-action accept
                    community add "spoke-ext-comm"
                exit
            exit
            policy-statement "vrf-import"
                entry 10
                    from
                        community "hub-ext-comm"
                    exit
                    action accept
                    exit
                exit
                default-action drop
                exit
            exit
            policy-statement "export-ospf"
                default-action accept
                exit
            exit
            commit
        exit
```

PE-2 is configured with VPRN 1, which has OSPF connectivity to the customer CE-3 and CE-7. The dedicated command **type spoke** is used to prevent direct CE spoke to CE spoke communications for this VPRN.

```
*A:PE-2# configure
    service
        vprn 1
            vrf-import "vrf-import"
            vrf-export "vrf-export"
            route-distinguisher 64500:12
            type spoke
            auto-bind-tunnel
                resolution any
            exit
            interface "int-PE-2-CE-3" create
                address 172.16.23.1/24
                sap 1/1/3:1 create
                exit
            exit
            interface "int-PE-2-CE-7" create
                address 172.16.27.1/24
                sap 1/1/4:1 create
                exit
            exit
            ospf
                export "export-ospf"
                area 0.0.0.0
                    interface "int-PE-2-CE-3"
                        interface-type point-to-point
                        mtu 1500
                        no shutdown
                    exit
                    interface "int-PE-2-CE-7"
                        interface-type point-to-point
                        mtu 1500
                        no shutdown
                    exit
                exit
                no shutdown
            exit
            no shutdown
```

For connectivity verification purposes, CE-3, CE-4, and CE-7 are configured to advertise their internal loopback interfaces via OSPF:

- CE-3 advertises 172.31.0.3/32
- CE-4 advertises 172.31.0.4/32
- CE-7 advertises 172.31.0.7/32

```
*A:CE-3# configure service
        vprn 1 customer 1 create
            route-distinguisher 64500:13
            interface "int-CE-3-PE-2" create
                address 172.16.23.2/24
                sap 1/1/1:1 create
                exit
            exit
            interface "lo0" create
                address 172.31.0.3/32
                loopback
            exit
            ospf 192.0.2.3
                ignore-dn-bit
                suppress-dn-bit
                area 0.0.0.0
                    interface "int-CE-3-PE-2"
                        interface-type point-to-point
                        mtu 1500
                    exit
                    interface "lo0"
                    exit
                exit
                no shutdown
            exit
            no shutdown
        exit
```

Hub Site Verification

The routing table (RIB) for VPRN 1 on PE-5 (hub site) lists all reachable networks.

```
*A:PE-5# show router 1 route-table
===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
0.0.0.0/0                                     Remote  OSPF      02d17h02m  150
       172.16.56.2                                                   1
172.16.14.0/24                                Remote  BGP VPN   00h32m41s  170
       192.0.2.1 (tunneled)                                          0
172.16.23.0/24                                Remote  BGP VPN   00h32m37s  170
       192.0.2.2 (tunneled)                                          0
172.16.27.0/24                                Remote  BGP VPN   00h32m37s  170
       192.0.2.2 (tunneled)                                          0
172.16.56.0/24                                Local   Local     02d17h31m  0
       int-PE-5-CE-6                                                 0
172.31.0.3/32                                 Remote  BGP VPN   00h03m30s  170
       192.0.2.2 (tunneled)                                          0
172.31.0.4/32                                 Remote  BGP VPN   00h02m21s  170
       192.0.2.1 (tunneled)                                          0
172.31.0.7/32                                 Remote  BGP VPN   00h01m31s  170
       192.0.2.2 (tunneled)                                          0
-------------------------------------------------------------------------------
No. of Routes: 8
```

The forwarding table (FIB) for the primary VRF of VPRN 1 is displayed using following command. All remote spoke and hub sites are reachable via this VRF.

```
*A:PE-5# show router 1 fib 1
===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
  NextHop
-------------------------------------------------------------------------------
0.0.0.0/0                                                   OSPF
  172.16.56.2 (int-PE-5-CE-6)
172.16.14.0/24                                              BGP_VPN
  192.0.2.1 (VPRN Label:262142 Transport:LDP)
172.16.23.0/24                                              BGP_VPN
```

```
  192.0.2.2 (VPRN Label:262142 Transport:LDP)
172.16.27.0/24                                              BGP_VPN
  192.0.2.2 (VPRN Label:262142 Transport:LDP)
172.16.56.0/24                                              LOCAL
  172.16.56.0 (int-PE-5-CE-6)
172.31.0.3/32                                               BGP_VPN
  192.0.2.2 (VPRN Label:262142 Transport:LDP)
172.31.0.4/32                                               BGP_VPN
  192.0.2.1 (VPRN Label:262142 Transport:LDP)
172.31.0.7/32                                               BGP_VPN
  192.0.2.2 (VPRN Label:262142 Transport:LDP)
-------------------------------------------------------------------------------
Total Entries : 8
-------------------------------------------------------------------------------
===============================================================================
*A:PE-5#
```

The forwarding table for the secondary VRF of VPRN 1 is displayed using following command, including the **secondary** keyword. All local hub CE sites are reachable via this VRF.

```
*A:PE-5# show router 1 fib 1 secondary
===============================================================================
FIB Display
===============================================================================
Prefix                                                      Protocol
    NextHop
-------------------------------------------------------------------------------
0.0.0.0/0                                                   OSPF
    172.16.56.2 (int-PE-5-CE-6)
172.16.56.0/24                                              LOCAL
    172.16.56.0 (int-PE-5-CE-6)
-------------------------------------------------------------------------------
Total Entries : 2
-------------------------------------------------------------------------------
===============================================================================
```

Spoke Site Verification

The RIB for VPRN 1 on PE-2 (spoke VPRN) lists all reachable networks.

The other spoke sites connected to the remote PEs (only CE-4 here, for example: 172.31.0.4/32)) are not present in the routing table.

The local interface addresses of PE-2 (172.16.23.1/32 and 172.16.27.1/32) are present in the routing table of VPRN 1, as follows. From a FIB point of view, these are reachable from any spoke VPRN, but the spoke CE's router host addresses are not. This fact does not influence the data plane isolation for the customer networks.

```
*A:PE-2# show router 1 route-table
===============================================================================
Route Table (Service: 1)
```

```
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
0.0.0.0/0                                     Remote  BGP VPN   00h04m33s  170
       192.0.2.5 (tunneled)                                         0
172.16.23.0/24                                Local   Local     02d17h47m  0
       int-PE-2-CE-3                                                0
172.16.23.1/32                                Local   Host      02d17h47m  0
       int-PE-2-CE-3                                                0
172.16.27.0/24                                Local   Local     02d17h47m  0
       int-PE-2-CE-7                                                0
172.16.27.1/32                                Local   Host      02d17h47m  0
       int-PE-2-CE-7                                                0
172.16.56.0/24                                Remote  BGP VPN   00h04m33s  170
       192.0.2.5 (tunneled)                                         0
172.31.0.3/32                                 Remote  OSPF      01h47m42s  10
       172.16.23.2                                                  10
172.31.0.7/32                                 Remote  OSPF      01h45m45s  10
       172.16.27.2                                                  10
-------------------------------------------------------------------------------
No. of Routes: 8
```

The FIB for the primary VRF of VPRN 1 shows all local spoke sites are reachable via this VRF, as follows:

```
*A:PE-2# show router 1 fib 1
===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
  NextHop
-------------------------------------------------------------------------------
172.16.23.0/24                                              LOCAL
  172.16.23.0 (int-PE-2-CE-3)
172.16.23.1/32                                              HOST
  Blackhole
172.16.27.0/24                                              LOCAL
  172.16.27.0 (int-PE-2-CE-7)
172.16.27.1/32                                              HOST
  Blackhole
172.31.0.3/32                                               OSPF
  172.16.23.2 (int-PE-2-CE-3)
172.31.0.7/32                                               OSPF
  172.16.27.2 (int-PE-2-CE-7)
-------------------------------------------------------------------------------
Total Entries : 6
```

The FIB for the secondary VRF of VPRN 1 shows the remote hub site (address 172.16.56.0/24) is reachable via this VRF, as follows:

```
*A:PE-2# show router 1 fib 1 secondary
===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
  NextHop
```

```
-------------------------------------------------------------------------------
0.0.0.0/0                                                   BGP_VPN
  192.0.2.5 (VPRN Label:262141 Transport:LDP)
172.16.23.1/32                                              HOST
  Blackhole
172.16.27.1/32                                              HOST
  Blackhole
172.16.56.0/24                                              BGP_VPN
  192.0.2.5 (VPRN Label:262141 Transport:LDP)
-------------------------------------------------------------------------------
Total Entries : 4
```

Spoke Sites Connectivity Verification

Without the VPRN spoke type configuration in VPRN 1 on PE-2, CE-3 takes the shortest path to CE-7, which violates the “hub and spoke” design approach explained earlier.

The VPRN has to be shut down in order to modify the type.

```
*A:PE-2# configure service vprn 1 no type
INFO: PIP #1162 Instance must be 'shutdown'
*A:PE-2# configure service vprn 1 shutdown
*A:PE-2# configure service vprn 1 no type
*A:PE-2# configure service vprn 1 no shutdown
```

Note: In this setup, a VPRN is configured on CE-3, but that is not necessary.

```
*A:CE-3# traceroute router 1 172.31.0.7
traceroute to 172.31.0.7, 30 hops max, 40 byte packets
  1  172.16.23.1 (172.16.23.1)    1.46 ms  0.642 ms  0.547 ms
  2  172.31.0.7 (172.31.0.7)    2.55 ms  1.70 ms  1.79 ms
```

After enabling the **type spoke** feature on PE-2, CE-3 takes the longest path via hub CE-6 to reach CE-7, as it should.

```
*A:PE-2# configure service vprn 1 shutdown
*A:PE-2# configure service vprn 1 type spoke
*A:PE-2# configure service vprn 1 no shutdown
```

```
*A:CE-3# traceroute router 1 172.31.0.7 no-dns
traceroute to 172.31.0.7, 30 hops max, 40 byte packets
  1  172.16.23.1    0.691 ms  0.642 ms  0.638 ms
  2  0.0.0.0  * * *
  3  172.16.56.2    2.32 ms  2.56 ms  2.25 ms
  4  172.16.56.1    2.46 ms  1.86 ms  2.27 ms
  5  172.16.27.1    1.89 ms  1.95 ms  1.99 ms
  6  172.31.0.7   4.44 ms  4.43 ms  4.90 ms
```

```
*A:CE-3#
```

Similarly, the long path is taken by CE-3 to reach CE-4, as follows. This is unrelated to the VPRN type. It is achieved by policies.

```
*A:CE-3# traceroute router 1 172.31.0.4 no-dns
traceroute to 172.31.0.4, 30 hops max, 40 byte packets
  1  172.16.23.1    0.677 ms  0.545 ms  0.546 ms
  2  0.0.0.0  * * *
  3  172.16.56.2    2.18 ms  2.41 ms  2.30 ms
  4  172.16.56.1    2.05 ms  1.94 ms  2.02 ms
  5  172.16.14.1    2.70 ms  3.13 ms  3.36 ms
  6  172.31.0.4    3.78 ms  3.76 ms  4.57 ms
*A:CE-3#
```

Conclusion

The VPRN type spoke feature completes the **CE hub and spoke** solution. It brings a new level of simplicity, scalability, and flexibility to operators using this VPRN architecture for their customers.

NG-MVPN Configuration with MPLS

This chapter provides information about NG-MVPN configuration with MPLS.

Topics in this chapter include:

- Applicability
- Summary
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SR OS release 9.0.R5, but the CLI in this edition corresponds to release 15.0.R5. There are no prerequisites for this configuration.

Summary

Multicast VPN (MVPN) or Next Generation IP Multicast in an IP-VPN (NG-MVPNs) architectures describe a set of virtual routing and forwarding (VRFs) or virtual private routed networks (VPRNs) that support the transport of multicast traffic across a provider network. MVPNs are defined in RFC 6513, *Multicast in MPLS/BGP IP VPNs*, and RFC 6514, *BGP Encodings and Procedures for Multicast in MPLS/IP VPNs*.

Initial MVPN deployments were originally based on Rosen MVPN (RFC 6037) which described the protocols and procedures required to support an IP Multicast VPN. There were a number of limitations with the Rosen MVPN implementation including, but not limited to:

- Rosen MVPN requires a set of multicast distribution trees (MDTs) per VPN, which requires a PIM state per MDT. There is no option to aggregate MDTs across multiple VPNs.

- Customer signaling. Initially, PE discovery and Data MDT signaling were all PIM-based because there was no mechanism available to decouple these. Now, PE discovery is supported using a BGP MDT address family identifier/subsequent address family identifier (AFI/SAFI), however, the data MDT still needs PIM.
- There is no mechanism for using MPLS to encapsulate multicast traffic in the VPN. GRE is the only encapsulation method available in Rosen MVPN.
- Rosen MPVN multicast trees are signaled using PIM only. MVPN allows the use of mLDP and RSVP P2MP LSPs.
- PE to PE protocol exchanges for Rosen MVPN is achieved using PIM only. MVPN allows for the use of BGP signaling as per unicast Layer 3 VPNs.

NG-MVPN addresses these limitations by extending the idea of the per-VRF tree by introducing the idea of provider multicast service interfaces (PMSIs). These are equivalent to the default MDTs of Rosen MVPN. NG-MVPN allows the decoupling of the mechanisms required to create a multicast VPN, such as PE auto-discovery (which PEs are members of which VPN), PMSI signaling (creation of tunnels between PEs), and customer multicast signaling (multicast signaling —IGMP/PIM— received from customer edge routers). Two types of PMSI exist:

- Inclusive (I-PMSI) — Contains all the PEs for a given MVPN, I-PMSI is the default multicast data path between all PEs of the same VPN.
- Selective (S-PMSI) — Contains only a subset of PEs of a given MVPN, used to optimize multicast stream distribution to only the PEs with active receivers for those streams.

The NG-MVPN Configuration with PIM chapter contains the VPN configuration required for the provider multicast domain using PIM Any Source Multicast (ASM) with auto-discovery based on PIM or BGP auto-discovery (AD), PIM used for the customer multicast signaling and PIM Source Specific Multicast (SSM) used for the S-PMSI creation. The customer domain configuration covers the following cases:

- PIM ASM with the Rendezvous Point (RP) in the provider PE
- PIM ASM using anycast RP on the provider RPs
- PIM SSM

This chapter introduces some of the features that were not supported at the time of writing of chapter NG-MVPN Configuration with PIM (Release 7.0). It provides configuration details to implement:

- Multicast LDP (mLDP) and RSVP-TE Point to Multipoint (P2MP) for building customer trees (C-trees) which are using MPLS instead of PIM techniques.
- MVPN source redundancy.
- MDT AFI/SAFI (to fully interoperate with Cisco networks).

PIM SSM is the only case addressed in this example, other PIM customer domain configurations are out of the scope, for more information refer to NG-MVPN Configuration with PIM.

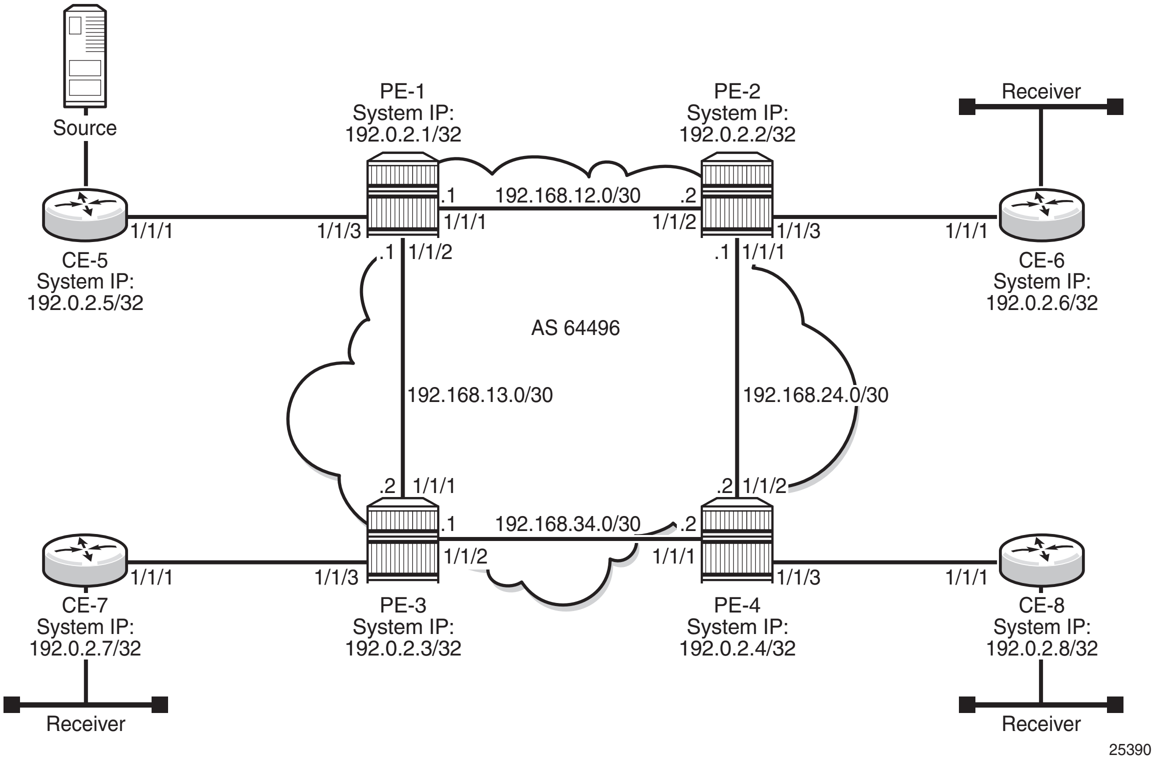
Overview

The network topology is shown in Figure 163. The setup consists of four 7750 SRs acting as provider edge (PE) routers within a single autonomous system (AS).

- Full mesh IS-IS in the AS (OSPF could be used instead)
- LDP on all interfaces in each AS (RSVP could be used instead)
- MP-iBGP sessions between the PE routers in the AS (route reflectors (RRs) could also be used).
- Layer 3 VPN on all PEs with identical route targets.

Connected to each PE is a single 7750 SR acting as a customer edge (CE) router. CE-5 has a multicast source connected, and PE-2, PE-3, and PE-4 each have a single receiver connected which will receive the multicast streams from the source. In this setup, each receiver is IGMPv3 capable. If the receiver is IGMPv3 capable, it will issue IGMPv3 reports that may include a list of required source addresses.

Figure 163 **Network Topology**



When the receiver wishes to become a member of any group, the source address of the group must be known to the CE. As a result, the source address must be IP reachable by each CE, so it is advertised using BGP by CE-5 to the PEs with attachment circuits in the VPRN. Static routes are then configured on the receiver CEs to achieve IP reachability to the source address of the multicast group.

Multicast traffic from the source is streamed toward router CE-5. Receivers connected to PE-2, PE-3 and PE-4 are interested in joining this multicast group.

CEs 5 to 8 are PIM enabled routers, which form a PIM adjacency with their nearest PE. Between the PEs across the provider network, there are no PIM adjacencies, because BGP auto-discovery and BGP signaling are used. Selective PMSI using mLDP or RSVP P2MP are out of the scope of this chapter. Selective PMSI using PIM SSM is supported too. I-PMSI and S-PMSI must use the same tunneling technology, either PIM/GRE, or mLDP, or RSVP P2MP.

Configuration

The configuration is divided into the following sections:

- Provider common configuration
  - PE global configuration
- PE VPRN configuration and PE VPRN multicast configuration for NG-MVPN
  - PMSI using mLDP
  - PMSI using RSVP-TE
  - UMH (upstream multicast hop)
- PE VPRN configuration and PE VPRN multicast configuration for Rosen MVPN using MDT AFI SAFI
  - Auto discovery using BGP MDT AFI SAFI as per Rosen MVPN version 9 with MDT using PIM SSM

Provider Common Configuration

PE Global Configuration

This section describes the common configuration required for each PE within the provider multicast domain, regardless of the MVPN PE auto-discovery or customer signaling methods. This includes interior gateway protocol (IGP) and VPRN service configuration.

The configuration tasks can be summarized as follows:

- PE global configuration.

  This includes configuration of the IGP (IS-IS will be used); configuration of link layer LDP between PEs (LDP will be used here as the method to interconnect VPRNs); configuration of iBGP between PEs to facilitate VPRN route learning.
- VPRN configuration on the PEs.

  This includes configuration of basic VPRN parameters (route-distinguisher, route target communities), configuration of attachment circuits toward CEs, configuration of VRF routing protocol and any routing policies.
- PIM within the VRF and MVPN parameters — I-PMSI
- CE configuration.

Step 1.

Configure the interfaces, the IGP (IS-IS) in all PE nodes (where IS-IS redistributes route reachability to all routers) and LDP in the interfaces (link layer LDP). To facilitate the IS-IS configuration, all routers are Level2-Level1 capable within the same ISIS area-id, so there is only a single topology area in the network (all routers share the same topology). The configuration for PE-1 is displayed below.

```
# on PE-1
configure
    router
        interface "int-PE-1-PE-2"
            address 192.168.12.1/30
            port 1/1/1
            no shutdown
        exit
        interface "int-PE-1-PE-3"
            address 192.168.13.1/30
            port 1/1/2
            no shutdown
        exit
        interface "system"
            address 192.0.2.1/32
            no shutdown
        exit
        autonomous-system 64496
        isis 0
            area-id 49.0001
            traffic-engineering
            interface "system"
                passive
                no shutdown
            exit
            interface "int-PE-1-PE-2"
                interface-type point-to-point
                no shutdown
            exit
            interface "int-PE-1-PE-3"
                interface-type point-to-point
                no shutdown
            exit
            no shutdown
        exit
        ldp
            interface-parameters
                interface "int-PE-1-PE-2" dual-stack
                exit
                interface "int-PE-1-PE-3" dual-stack
                exit
            exit
        exit
    exit
exit
```

The configuration for the rest of nodes is similar. The IP addresses can be derived from Figure 163.

Step 2.

Verify that IS-IS adjacencies and LDP peer sessions are formed.

```
*A:PE-1# show router isis adjacency

===============================================================================
Rtr Base ISIS Instance 0 Adjacency
===============================================================================
System ID                Usage State Hold Interface                     MT-ID
-------------------------------------------------------------------------------
PE-2                     L1L2  Up    22   int-PE-1-PE-2                 0
PE-3                     L1L2  Up    22   int-PE-1-PE-3                 0
-------------------------------------------------------------------------------
Adjacencies : 2
===============================================================================
*A:PE-1#
```

```
*A:PE-1# show router ldp session ipv4

==============================================================================
LDP IPv4 Sessions
==============================================================================
Peer LDP Id         Adj Type  State         Msg Sent  Msg Recv  Up Time
------------------------------------------------------------------------------
192.0.2.2:0         Link      Established   21        22        0d 00:00:33
192.0.2.3:0         Link      Established   19        20        0d 00:00:24
------------------------------------------------------------------------------
No. of IPv4 Sessions: 2
==============================================================================
*A:PE-1#
```

Step 3.

Configure iBGP full mesh between the PEs for VPRN routing (Route Reflectors could also be an option).

```
# on PE-1
configure
    router
        bgp
            min-route-advertisement 1
            rapid-withdrawal
            rapid-update mvpn-ipv4 mdt-safi
            group "INTERNAL"
                family vpn-ipv4 mvpn-ipv4 mdt-safi
                type internal
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
            exit
            no shutdown
        exit
```

The families configured under the group “INTERNAL” are vpn-ipv4, mvpn-ipv4, and mdt-safi, since the three families are referenced in this chapter.

The mdt-safi parameter is not needed for NG-MVPN (mLDP/RSVP scenarios) and is only required for Rosen MVPN with MDT AFI SAFI.

Rapid withdrawal (configured on all PEs) disables the minimum route advertisement interval (MRAI) interval on sending BGP withdrawals. Rapid update (configured for MVPN-IPv4 and MDT AFI/SAFI address families) disables the MRAI interval on sending BGP update messages for the address family MVPN-IPv4 and MDT-SAFI).

Step 4.

Verify that BGP peer relationships are established.

```
*A:PE-1# show router bgp summary
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
BGP Admin State         : Up          BGP Oper State              : Up
Total Peer Groups       : 1           Total Peers                 : 3
Total VPN Peer Groups   : 0           Total VPN Peers             : 0
Total BGP Paths         : 15          Total Path Memory           : 3960

Total IPv4 Remote Rts   : 0           Total IPv4 Rem. Active Rts  : 0
Total IPv6 Remote Rts   : 0           Total IPv6 Rem. Active Rts  : 0

--- snipped ---

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
                   AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.2
               64496        4    0 00h00m55s 0/0/0 (VpnIPv4)
                            4    0           0/0/0 (MvpnIPv4)
                                             0/0/0 (MdtSafi)
192.0.2.3
               64496        4    0 00h00m46s 0/0/0 (VpnIPv4)
                            4    0           0/0/0 (MvpnIPv4)
                                             0/0/0 (MdtSafi)
192.0.2.4
               64496        4    0 00h00m38s 0/0/0 (VpnIPv4)
                            4    0           0/0/0 (MvpnIPv4)
                                             0/0/0 (MdtSafi)
-------------------------------------------------------------------------------
*A:PE-1#
```

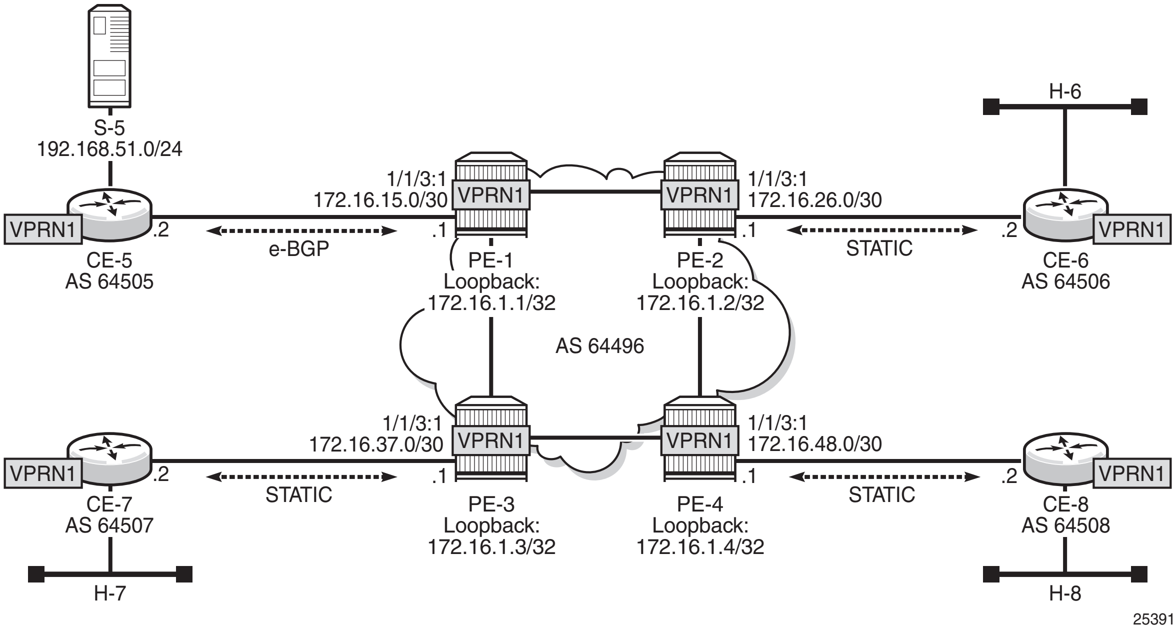
PE VPRN Configuration and PE VPRN Multicast Configuration

A VPRN is created on each PE per service (the different services using mLDP, RSVP-TE, and AFI/SAFI with PIM); these are the multicast VPRNs. PE-1 is the PE containing the attachment circuit toward CE-5. CE-5 is the CE nearest to the source. PE-2, PE-3, and PE-4 contain attachment circuits toward CE-6, CE-7, and CE-8 respectively. Each CE has a receiving host attached.

PMSI using mLDP

Figure 164 shows the details of the topology for VPRN 1.

Figure 164 **VPRN 1 Topology used for mLDP**



Unicast

Step 1.

Create VPRN 1 on each PE, containing a route-distinguisher of 64496:10X (where X= number of PE) and vrf-target of 64496:100. The autonomous system number is 64496. For the next hop tunnel route resolution to connect the VPRNs between the PEs, manually configured spoke SDPs are created (other methods such as auto-bind-tunnel resolution-filter LDP resolution filter could also be used). LDP was already enabled.

```
# on PE-1
configure
    service
        sdp 12 mpls create
            far-end 192.0.2.2
            ldp
            no shutdown
        exit
        sdp 13 mpls create
            far-end 192.0.2.3
            ldp
            no shutdown
        exit
        sdp 14 mpls create
            far-end 192.0.2.4
            ldp
            no shutdown
        exit
        vprn 1 customer 1 create
            description "mLDP"
            autonomous-system 64496
            route-distinguisher 64496:101
            vrf-target target:64496:100
            spoke-sdp 12 create
            exit
            spoke-sdp 13 create
            exit
            spoke-sdp 14 create
            exit
```

Step 2.

Create an attachment circuit interface toward the CE and a loopback (the loopback is not mandatory, but it is configured to aid troubleshooting the routers).

```
# on PE-1
configure
    service
        vprn 1
            interface "loopback" create
                address 172.16.1.1/32
                loopback
            exit
            interface "int-PE-1-CE-5" create
                address 172.16.15.1/30
                sap 1/1/3:1 create
                exit
            exit
```

Step 3.

The source address of the multicast stream will need to be reachable by all routers (PEs and CEs) within the VPN. This will be advertised within BGP from CE-5 to PE-1. Create a BGP peering relationship with the CE as follows:

```
# on PE-1
configure
    service
        vprn 1
            bgp
                group "EXTERNAL"
                    type external
                    peer-as 64505
                    neighbor 172.16.15.2
                    exit
                exit
                no shutdown
            exit
```

Step 4.

On CE-5, create a VPRN to support the connection of the source to CE-5 and the connection from CE-5 to PE-1. Two attachment circuits are required as well as a BGP peering relationship with the PE. This uses a default BGP address family of ipv4.

```
# on CE-5
configure
    service
        vprn 1 customer 1 create
            autonomous-system 64505
            route-distinguisher 64505:1
            interface "int-CE-5-PE-1" create
                address 172.16.15.2/30
                sap 1/1/1:1 create
                exit
            exit
            interface "int-CE-5-S-5" create
                address 192.168.51.1/24
                sap 1/1/3 create
                exit
            exit
            bgp
                group "EXTERNAL"
                    type external
                    peer-as 64496
                    neighbor 172.16.15.1
                    exit
                exit
                no shutdown
            exit
            no shutdown
        exit
```

Step 5.

In order for the subnet on the CE connecting to the source to be advertised within BGP, a route policy is required. The subnet containing the multicast source is 192.168.51.0/24, so a prefix-list can be used to define a match, and then used within a route policy to inject into BGP.

```
# on CE-5
configure
    router
        policy-options
            begin
            prefix-list "SOURCE-PREFIX"
                prefix 192.168.51.0/24 exact
            exit
            policy-statement "EXPORT-SOURCE-PREFIX-TO-BGP"
                entry 10
                    from
                        prefix-list "SOURCE-PREFIX"
                    exit
                    to
                        protocol bgp
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit


configure
    service
        vprn 1
            bgp
                export "EXPORT-SOURCE-PREFIX-TO-BGP"
            exit
        exit
```

Step 6.

Check that the route is seen in PE-1:

```
*A:PE-1# show router 1 route-table 192.168.51.0/24

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.168.51.0/24                               Remote  BGP       00h01m29s  170
       172.16.15.2                                                  0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
================================================================================
*A:PE-1#
```

This prefix will also be automatically advertised within the BGP VPRN to all other PEs, and will be installed in VRF 1.

For example, on PE-4, the source subnet 192.168.51.0/24 is received via BGP VPN with a next-hop of PE-1 (192.0.2.1):

```
*A:PE-4# show router 1 route-table 192.168.51.0/24

================================================================================
Route Table (Service: 1)
================================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
--------------------------------------------------------------------------------
192.168.51.0/24                               Remote  BGP VPN   00h01m40s  170
       192.0.2.1 (tunneled)                                         0
--------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
================================================================================
*A:PE-4#
```

Each CE containing a multicast receiver must be able to reach the source. As an example on CE-6, a static route is configured with next hop 172.16.26.1 of interface int-PE-2-CE-6.

```
# on CE-6
configure
    service
        vprn 1 customer 1 create
            autonomous-system 64506
            route-distinguisher 64506:1
            interface int-CE-6-H-6 create
                address 192.168.61.1/24
                sap 1/1/2:1 create
                exit
            exit
            interface int-CE-6-PE-2 create
                address 172.16.26.2/30
                sap 1/1/1:1 create
                exit
            exit
            static-route-entry 192.168.51.0/24
                next-hop 172.16.26.1
                    no shutdown
                exit
            exit
            no shutdown
        --- snipped ---
```

After **Steps 1** to **6**, all required unicast routing is provisioned. The following sections show the configuration of the multicast in the VPRN.

Auto-Discovery and mLDP PMSI Establishment

The MP-BGP based auto-discovery is implemented with a dedicated address family defined in RFC 4760 MP_REACH_NLRI/MP_UNREACH_NLRI attributes, with AFI 1 (IPv4) or 2 (IPv6) SAFI 5 (temporary value assigned by IANA). This is the mechanism by which each PE advertises the presence of an MVPN to other PEs. This can be achieved using PIM (like in Rosen MVPN) or using BGP. With the default parameter, BGP is automatically chosen because the PMSIs are mLDP and PIM is not an option in this case. Any PE that is a member of an MVPN will advertise to the other PEs using a BGP multi-protocol network layer reachability information (NLRI) update that is sent to all PEs within the AS. This update will contain an Intra-AS I-PMSI auto-discovery route type, also known as an Intra-AD. These use an address family mvpn-ipv4, so each PE must be configured to originate and accept such updates (this was done earlier when configuring the families).

At this step (auto-discovery), the information about the PMSI is exchanged, but the PMSI is not instantiated.

As each PE contains a CE which will be part of the multicast VRF, it is necessary to enable PIM on each interface containing the attachment circuit toward a CE, and to configure the I-PMSI multicast tunnel for the VRF. In order for the BGP routes to be accepted into the VRF, a route-target community is required (vrf-target). This is configured in the **configure service vprn 1 mvpn** context and, in this case is set to the same value as the unicast vrf-target (the vrf-target community as the **configure service vprn 1 vrf-target** context).

On each PE, the PIM and MVPN context within the VPRN instance are configured as follows:

```
# on PE-4
configure
    service
        vprn 1
            pim
                interface "loopback"
                exit
                interface "int-PE-4-CE-8"
                exit
            exit
            mvpn
                auto-discovery default
                c-mcast-signaling bgp
                provider-tunnel
    inclusive
        mldp
            no shutdown
        exit
    exit
exit
vrf-target unicast
exit
```

When PIM SSM is used, the configuration always shows RP static with no RP entries (this is enabled by default when PIM is provisioned). In order for the BGP routes to be accepted into the VRF, a route-target community is required (vrf-target). Although it is not mandatory for the mvpn target to be equal to the unicast target, Nokia recommends to use **vrf-target unicast** to avoid configuration mistakes and extra complexity.

The status of VPRN 1 on PE-1 is shown with the following output:

```
*A:PE-1# show router 1 mvpn

===============================================================================
MVPN 1 configuration data
===============================================================================
signaling          : Bgp                  auto-discovery     : Default
UMH Selection      : Highest-Ip           SA withdrawn       : Disabled
intersite-shared   : Enabled              Persist SA         : Disabled
vrf-import         : N/A
vrf-export         : N/A
vrf-target         : unicast
C-Mcast Import RT  : target:192.0.2.1:2

ipmsi              : ldp
i-pmsi P2MP AdmSt  : Up
i-pmsi Tunnel Name : mpls-if-73728
Mdt-type           : sender-receiver

BSR signalling     : none
Wildcard s-pmsi    : Disabled
Multistream-SPMSI  : Disabled
s-pmsi             : none
data-delay-interval: 3 seconds
enable-asm-mdt     : N/A

===============================================================================
*A:PE-1#
```

The following shows a debug of an Intra-AD BGP update message received by PE-1 that was sent by PE-2. The message contains the PMSI tunnel type to be used (LDP P2MP LSP), LSP identification (root node, opaque value) and the type of BGP update (Type: Intra-AD Len: 12 RD: 64496:102 Orig: 192.0.2.2):

```
11 2017/10/07 18:31:59.676 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 91
    Flag: 0x90 Type: 14 Len: 23 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.2
        Type: Intra-AD Len: 12 RD: 64496:102 Orig: 192.0.2.2
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64496:100
    Flag: 0xc0 Type: 22 Len: 22 PMSI:
        Tunnel-type LDP P2MP LSP (2)
        Flags: (0x0)[Type: None BM: 0 U: 0 Leaf: not required]
        MPLS Label 0
        Root-Node 192.0.2.2, LSP-ID 0x2001
```

The setup has four PEs, so every PE should see the Intra-AD routes from its peers; the following output shows the routes received in PE-1:

```
*A:PE-1# show router bgp routes mvpn-ipv4 type intra-ad
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
Flag  RouteType                   OriginatorIP           LocalPref   MED
      RD                          SourceAS               Path-Id     Label
      Nexthop                     SourceIP
      As-Path                     GroupIP
-------------------------------------------------------------------------------
u*>i  Intra-Ad                    192.0.2.2              100         0
      64496:102                   -                      None        -
      192.0.2.2                   -
      No As-Path                  -
u*>i  Intra-Ad                    192.0.2.3              100         0
      64496:103                   -                      None        -
      192.0.2.3                   -
      No As-Path                  -
u*>i  Intra-Ad                    192.0.2.4              100         0
      64496:104                   -                      None        -
      192.0.2.4                   -
      No As-Path                  -
-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:PE-1#
```

The detailed output of the Intra-AD received from PE-2 shows the Tunnel-Type LDP P2MP LSP (LSP-ID is 8193), the originator id (192.0.2.2), and the route-distinguisher (64496:102):

```
*A:PE-1# show router bgp routes mvpn-ipv4 type intra-ad detail
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
Original Attributes

Route Type     : Intra-Ad
Route Dist.    : 64496:102
Originator IP  : 192.0.2.2
Nexthop        : 192.0.2.2
Path Id        : None
From           : 192.0.2.2
Res. Nexthop   : 0.0.0.0
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : no-export target:64496:100
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.2
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h01m47s
VPRN Imported  :  1
-------------------------------------------------------------------------------
PMSI Tunnel Attributes :
Tunnel-type    : LDP P2MP LSP
Flags          : Type: RNVE(0) BM: 0 U: 0 Leaf: not required
MPLS Label     : 0
Root-Node      : 192.0.2.2              LSP-ID         : 8193
-------------------------------------------------------------------------------
--- snipped ---
-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:PE-1#
```

Because of the receiver-driven nature of mLDP, mLDP P2MP LSPs are set up unsolicited from the leaf PEs toward the head-end PE. The leaf PEs discover the head-end PE via I-PMSI/S-PMSI AD routes. The tunnel identifier carried in the PMSI attribute is used as the P2MP forwarding equivalence class (FEC) Element. The tunnel identifier consists of the address of the head-end PE, along with a generic LSP identifier value. The generic LSP identifier value is automatically generated by the head-end PE. The preceding show command displays the PMSI information with the detail of the root node (192.0.2.2) and the LSP-ID (8193). The PMSI was created after receiving the AD message from PE-2, where the following excerpt from the previous debug shows the same information (0x2001 in HEX is equal to 8193 in decimal).

```
Flag: 0xc0 Type: 22 Len: 22 PMSI:
        Tunnel-type LDP P2MP LSP (2)
        Flags: (0x0)[Type: None BM: 0 U: 0 Leaf: not required]
        MPLS Label 0
        Root-Node 192.0.2.2, LSP-ID 0x2001
```

Once the mLDP P2MP LSPs are created, the I-PMSI is instantiated in the core:

```
*A:PE-1# show router 1 pim neighbor

===============================================================================
PIM Neighbor ipv4
===============================================================================
Interface               Nbr DR Prty     Up Time       Expiry Time     Hold Time
   Nbr Address
-------------------------------------------------------------------------------
int-PE-1-CE-5           1               0d 00:02:28   0d 00:01:43     105
   172.16.15.2
mpls-if-73729           1               0d 00:02:18   never           65535
   192.0.2.2
mpls-if-73730           1               0d 00:02:08   never           65535
   192.0.2.3
mpls-if-73731           1               0d 00:01:58   never           65535
   192.0.2.4
-------------------------------------------------------------------------------
Neighbors : 4
===============================================================================
*A:PE-1#
```

```
*A:PE-1# show router 1 pim tunnel-interface

===============================================================================
PIM Interfaces ipv4
===============================================================================
Interface                        Originator Address   Adm  Opr  Transport Type
-------------------------------------------------------------------------------
mpls-if-73728                    192.0.2.1            Up   Up   Tx-IPMSI
mpls-if-73729                    192.0.2.2            Up   Up   Rx-IPMSI
mpls-if-73730                    192.0.2.3            Up   Up   Rx-IPMSI
mpls-if-73731                    192.0.2.4            Up   Up   Rx-IPMSI
-------------------------------------------------------------------------------
Interfaces : 4
```

```
===============================================================================
*A:PE-1#
```

Every PE has created an I-PMSI to the other PEs. Checking the mLDP P2MP LSPs that are originated, transit, or destination to PE-1:

```
*A:PE-1# show router ldp bindings active p2mp ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
===============================================================================
LDP Generic IPv4 P2MP Bindings (Active)
===============================================================================
P2MP-Id                                     Interface
RootAddr                                    Op              IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
-------------------------------------------------------------------------------
8193                                        73728
192.0.2.1                                   Push              --      262138
192.168.12.2                                1/1/1

8193                                        73728
192.0.2.1                                   Push              --      262138
192.168.13.2                                1/1/2

8193                                        73729
192.0.2.2                                   Pop             262138      --
  --                                          --

8193                                        73729
192.0.2.2                                   Swap            262138    262137
192.168.13.2                                1/1/2

8193                                        73730
192.0.2.3                                   Pop             262137      --
  --                                          --

8193                                        73730
192.0.2.3                                   Swap            262137    262137
192.168.12.2                                1/1/1

8193                                        73731
192.0.2.4                                   Pop             262136      --
  --                                          --

-------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 7
===============================================================================
--- snipped ---
```

```
*A:PE-1#
```

The two first entries in the output show the P2MP LSP where PE-1 is the root head-end (Push). The other two entries (Swap and Pop) correspond with transit and leaf for the P2MP LSPs originated by the other PEs. The command shows a P2MP-ID (8193) with an interface 73728 (matches with the **show router 1 pim tunnel interface** being the PIM interface created from PE-1) with two egress interfaces pointing to PE-2 and PE-3.

A similar command executed on PE-2 shows:

```
*A:PE-2# show router ldp bindings active p2mp ipv4
================================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
================================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
================================================================================
LDP Generic IPv4 P2MP Bindings (Active)
================================================================================
P2MP-Id                                     Interface
RootAddr                                    Op              IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
--------------------------------------------------------------------------------
8193                                        73729
192.0.2.1                                   Pop             262138      --
  --                                          --

8193                                        73729
192.0.2.1                                   Swap            262138    262136
192.168.24.2                                1/1/1

--- snipped ---

--------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 7
================================================================================
--- snipped ---
*A:PE-2#
```

On PE-2, the first entry shows that PE-2 is a leaf of the P2MP LSP tree created by PE-1 (ingress label is 262138 which was the egress label to reach PE-2 and is popped). However, the second entry shows that PE-2 is transit for the P2MP LSP going to PE-4 (ingress label 262138, egress label 262136 next hop PE-4).

The same command on PE-4 shows:

```
*A:PE-4# show router ldp bindings active p2mp ipv4
```

```
===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.4)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
===============================================================================
LDP Generic IPv4 P2MP Bindings (Active)
===============================================================================
P2MP-Id                                     Interface
RootAddr                                    Op              IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
-------------------------------------------------------------------------------
8193                                        73731
192.0.2.1                                   Pop             262136      --
  --                                          --

--- snipped ---
-------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 5
===============================================================================
--- snipped ---
*A:PE-4#
```

In the first entry, the root is PE-1 and the action is Pop, being the ingress label 262136, showing that this is another leaf for the P2MP LSP started on PE-1.

To complete the information, checking on PE-3, the first entry there is a Pop where the root is PE-1, and the ingress label is 262138:

```
*A:PE-3# show router ldp bindings active p2mp ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.3)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
===============================================================================
LDP Generic IPv4 P2MP Bindings (Active)
===============================================================================
P2MP-Id                                     Interface
RootAddr                                    Op              IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
-------------------------------------------------------------------------------
8193                                        73729
192.0.2.1                                   Pop             262138      --
  --                                          --
```

```
--- snipped ---

-------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 5
===============================================================================
--- snipped ---
*A:PE-3#
```

As a summary, each root PE has a P2MP LSP with three leaves (the other PEs) and they are also transit points to the P2MP LSPs created in the other PEs. As an additional check, an OAM ping can show the different leaves that a P2MP LSP has:

```
*A:PE-1# oam p2mp-lsp-ping ldp 8193 sender-addr 192.0.2.1 detail
P2MP identifier 8193: | 88 bytes MPLS payload

===============================================================================
Leaf Information
===============================================================================
From             RTT                    Return Code
-------------------------------------------------------------------------------
192.0.2.2        =1.16ms                EgressRtr(3)
192.0.2.3        =1.18ms                EgressRtr(3)
192.0.2.4        =1.84ms                EgressRtr(3)
===============================================================================

Total Leafs responded = 3
          round-trip min/avg/max  = 1.16 / 1.40 / 1.84 ms
Responses based on return code:
EgressRtr(3)=3
*A:PE-1#
```

An easy way to see the path that the LDP P2MP LSP follows for a specific leaf is the following command (such as LDP trace from PE-1 to PE-4):

```
*A:PE-1# oam ldp-treetrace prefix 192.0.2.4/32

ldp-treetrace for Prefix 192.0.2.4/32:

      192.168.24.2, ttl =   2 dst =      127.1.0.255 rc = EgressRtr status = Done
    Hops:      192.168.12.2

ldp-treetrace discovery state: Done
ldp-treetrace discovery status: ' OK '
Total number of discovered paths: 1
Total number of failed traces: 0

*A:PE-1#
```

The command shows that on PE-4, there is an active leaf of the P2MP LSP, and that there is an intermediate hop on PE-2.

Traffic Flow

The receiver H-8, connected to CE-8, wishes to join the group 232.1.1.1 with source 192.168.51.1 and sends an IGMPv3 report toward CE-8. CE-8 recognizes the report and sends a PIM join toward the source, therefore, it reaches PE-1 where the source is connected to through CE-5. The following output shows the debug seen on PE-4, where the PIM join is received from CE-8 and a BGP update Source Join is sent to all PEs (only the update sent to PE-1 is shown).

```
17 2017/10/07 18:38:05.446 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimJPProcessSG
pimJPProcessSG: (S,G)-> (192.168.51.2,232.1.1.1) type <S,G>, i/f int-PE-4-CE-8,
upNbr 172.16.48.1 isJoin 1 isRpt 0 holdTime 210"

18 2017/10/07 18:38:05.446 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimRtmFindRpfNexthop
Track (192.168.51.2,232.1.1.1) type <S,G> using 192.168.51.2"

19 2017/10/07 18:38:05.446 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimRtmAddSrcEntry
Added src entry for src 192.168.51.2"

20 2017/10/07 18:38:05.446 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimJPPrintFsmEvent
PIM JP Downstream: State NoInfo Event RxJoin StandbyEvent F, (S,G)
 (192.168.51.2,232.1.1.1) groupType <S,G>"

21 2017/10/07 18:38:05.446 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimJPPrintFsmEvent
PIM JP Upstream: State NotJoined Event JoinDesiredTrue StandbyEvent F, (S,G)
 (192.168.51.2,232.1.1.1) groupType <S,G>"

22 2017/10/07 18:38:05.446 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimSGUpJoinDesiredTrue
No upstream interface. pSG (192.168.51.2,232.1.1.1) rpfType 3"

23 2017/10/07 18:38:05.446 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimSGUpJoinDesiredTrue
No upstream interface SG (192.168.51.2,232.1.1.1) rpfType 3"

24 2017/10/07 18:38:05.446 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimRtmProcessNhresEvent
RTM-Nhres Event U-RTM NEW Src 192.168.51.2 SrcRtmUse UCAST"

25 2017/10/07 18:38:05.446 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimRtmProcessNhresEvent
Prefix 192.168.51.0/24 numNextHops 1 owner BGP_VPN metric 20 pref 170"

26 2017/10/07 18:38:05.446 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimRtmSrcResolveSGsInt
Trying to resolve SG (192.168.51.2,232.1.1.1)"

27 2017/10/07 18:38:05.446 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimRtmNotifyRpfChange
RPF Change to Source/RP 192.168.51.2  for SG (192.168.51.2,232.1.1.1) dynMLDP F via
 NH 192.0.2.1 IfIdx: 73731 RpfType: REMOTE Reason:  RTE_ADD old NH 0.0.0.0 IfIdx: 0
 RpfType: NONE mplsRpf F NextHops 1 reg 1/1 lfa 0/0"
```

```
28 2017/10/07 18:38:05.446 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimRtmNotifyRpfChange
SG (192.168.51.2,232.1.1.1) Source/RP 192.168.51.2 Ipmsi 73728 NhIf 0 new NhIf 73731"

29 2017/10/07 18:38:05.446 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimJPPrintFsmEvent
PIM JP Upstream: State Joined Event MribChange StandbyEvent F, (S,G)
 (192.168.51.2,232.1.1.1) groupType <S,G>"

30 2017/10/07 18:38:05.446 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimSGUpStateJMribChange
SG (192.168.51.2,232.1.1.1), type <S,G> oldMribNhopIp 0.0.0.0 oldRpfNbrIp 0.0.0.0,
 oldRpfType NONE oldRpfIf 0 rptMribNhopIp 0.0.0.0, rptRpfNbrIp 0.0.0.0 rtmReason 48
 isSGExtNet : no"

31 2017/10/07 18:38:05.446 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimSGUpStateJMribChange
SG (192.168.51.2,232.1.1.1), type <S,G> newMribNhopIp 192.0.2.1 newRpfNbrIp 192.0.2.1
 newRpfType REMOTE newRpfIf 73731"

32 2017/10/07 18:38:05.446 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimAddToJPTxPdu
pimAddToJPTxPdu: (S,G)-> (192.168.51.2,232.1.1.1), type <S,G>, txPendFlag J isStandby
 F"

33 2017/10/07 18:38:05.446 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimRtmUpdateSGMetric
SG metric 4294967295 pref 2147483647, new metric 20 pref 170"

--- snipped ---

36 2017/10/07 18:38:05.446 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 76
    Flag: 0x90 Type: 14 Len: 33 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.4
        Type: Source-Join Len:22 RD: 64496:101 SrcAS: 64496
                                 Src: 192.168.51.2 Grp: 232.1.1.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:192.0.2.1:2
"
```

The following debug shows that PE-1 receives the BGP update Source Join with source 192.168.1.1 and group 232.1.1.1 and sends a PIM join toward CE-5:

```
19 2017/10/07 18:38:05.446 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.4
"Peer 1: 192.0.2.4: UPDATE
Peer 1: 192.0.2.4 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 76
    Flag: 0x90 Type: 14 Len: 33 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.4
        Type: Source-Join Len:22 RD: 64496:101 SrcAS: 64496
                                 Src: 192.168.51.2 Grp: 232.1.1.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:192.0.2.1:2
"

20 2017/10/07 18:38:05.447 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimProcessMvpnRouteMsg
originator 0.0.0.0: add rtType SOURCE_TREE_JOIN nextHop 192.0.2.4
source 192.168.51.2 group 232.1.1.1"

21 2017/10/07 18:38:05.447 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimJPProcessSG
pimJPProcessSG: (S,G)-> (192.168.51.2,232.1.1.1) type <S,G>, i/f mpls-if-73728,
upNbr 192.0.2.1 isJoin 1 isRpt 0 holdTime 65535"

22 2017/10/07 18:38:05.447 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimRtmFindRpfNexthop
Track (192.168.51.2,232.1.1.1) type <S,G> using 192.168.51.2"

23 2017/10/07 18:38:05.447 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimRtmAddSrcEntry
Added src entry for src 192.168.51.2"

24 2017/10/07 18:38:05.447 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimJPPrintFsmEvent
PIM JP Downstream: State NoInfo Event RxJoin StandbyEvent F, (S,G)
 (192.168.51.2,232.1.1.1) groupType <S,G>"

25 2017/10/07 18:38:05.447 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimJPPrintFsmEvent
PIM JP Upstream: State NotJoined Event JoinDesiredTrue StandbyEvent F, (S,G)
 (192.168.51.2,232.1.1.1) groupType <S,G>"

26 2017/10/07 18:38:05.447 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimAddToJPTxPdu
pimAddToJPTxPdu: (S,G)-> (192.168.51.2,232.1.1.1), type <S,G>,
txPendFlag J isStandby F"

27 2017/10/07 18:38:05.447 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimRtmProcessNhresEvent
RTM-Nhres Event U-RTM NEW Src 192.168.51.2 SrcRtmUse UCAST"

28 2017/10/07 18:38:05.447 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimRtmProcessNhresEvent
Prefix 192.168.51.0/24 numNextHops 1 owner BGP metric 0 pref 170"

--- snipped ---
```

```
37 2017/10/07 18:38:05.447 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimSGEncodeGroupSet
Encoding  Join for source 192.168.51.2"

38 2017/10/07 18:38:05.447 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimSGEncodeGroupSet
num joined srcs 1, num pruned srcs 0"

39 2017/10/07 18:38:05.447 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimSendJoinPrunePdu
sending JP PDU with 1 groups, if 5 adj 172.16.15.2"

40 2017/10/07 18:38:05.447 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimSendJoinPrunePdu
if 5, adj 172.16.15.2. Nothing to send"
```

The BGP update source join received by PE-1 is displayed with the command:

```
*A:PE-1# show router bgp routes mvpn-ipv4 type source-join
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
Flag  RouteType                   OriginatorIP           LocalPref   MED
      RD                          SourceAS               Path-Id     Label
      Nexthop                     SourceIP
      As-Path                     GroupIP
-------------------------------------------------------------------------------
u*>i  Source-Join                 -                      100         0
      64496:101                   64496                  None        -
      192.0.2.4                   192.168.51.2
      No As-Path                  232.1.1.1
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-1#
```

To verify the traffic: on PE-1 there is a group 232.1.1.1 with source 192.168.51.2, the Reverse Path Forwarding (RPF) is CE-5, the multicast traffic is flowing from CE-5 to PE-1 using int-PE-1-CE-5 and the outgoing interface is using the PMSI mLDP mpls-if-73728.

```
*A:PE-1# show router 1 pim group detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 232.1.1.1
```

```
Source Address     : 192.168.51.2
RP Address         : 0
Advt Router        : 172.16.15.2
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 172.16.15.2
MRIB Src Flags     : remote
Keepalive Timer    : Not Running
Up Time            : 0d 00:00:41        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:19
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 172.16.15.2
Incoming Intf      : int-PE-1-CE-5
Outgoing Intf List : mpls-if-73728

Curr Fwding Rate   : 1042.6 kbps
Forwarded Packets  : 3582               Discarded Packets  : 0
Forwarded Octets   : 5365836            RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-1#
```

On PE-4, the same (S,G) arrives in the incoming interface mpls-if-73731 and the outgoing interface is int-PE-4-CE-8.

```
*A:PE-4# show router 1 pim group detail
===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 232.1.1.1
Source Address     : 192.168.51.2
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 192.0.2.1
MRIB Src Flags     : remote
Keepalive Timer    : Not Running
Up Time            : 0d 00:00:44        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:16
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 192.0.2.1
Incoming Intf      : mpls-if-73731
Outgoing Intf List : int-PE-4-CE-8
Curr Fwding Rate   : 1042.6 kbps
```

```
Forwarded Packets   : 3785             Discarded Packets  : 0
Forwarded Octets    : 5669930          RPF Mismatches     : 0
Spt threshold       : 0 kbps           ECMP opt threshold : 7
Admin bandwidth     : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-4#
```

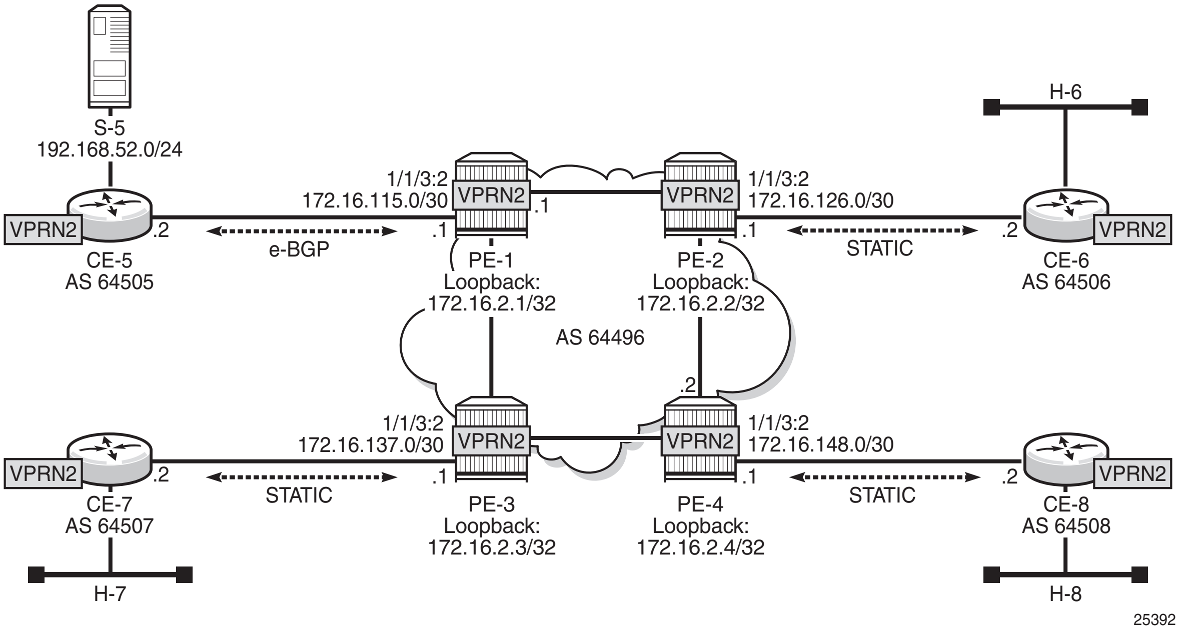
When the receiver is not interested in the channel group any more, the receiver H-8 sends an IGMPv3 leave, PE-4 sends a PIM prune translated to a BGP MP_UNREACH NLRI to all PEs. As mentioned before, rapid withdrawals are sent without waiting for the MRAI (for simplicity, only one BGP update is shown in the output debug).

```
41 2017/10/07 18:39:15.413 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimJPProcessSG
pimJPProcessSG: (S,G)-> (192.168.51.2,232.1.1.1) type <S,G>, i/f int-PE-4-CE-8,
upNbr 172.16.48.1 isJoin 0 isRpt 0 holdTime 210"

42 2017/10/07 18:39:15.413 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimJPPrintFsmEvent
PIM JP Downstream: State Joined Event RxPrune StandbyEvent F,
(S,G) (192.168.51.2,232.1.1.1) groupType <S,G>"

43 2017/10/07 18:39:15.413 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimJPPrintFsmEvent
PIM JP Downstream: State PrunePending Event PrunePendTimerExp StandbyEvent F,
(S,G) (192.168.51.2,232.1.1.1) groupType <S,G>"

44 2017/10/07 18:39:15.413 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimJPPrintFsmEvent
PIM JP Upstream: State Joined Event JoinDesiredFalse StandbyEvent F,
(S,G) (192.168.51.2,232.1.1.1) groupType <S,G>"

45 2017/10/07 18:39:15.413 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimAddToJPTxPdu
pimAddToJPTxPdu: (S,G)-> (192.168.51.2,232.1.1.1), type <S,G>,
txPendFlag P isStandby F"

46 2017/10/07 18:39:15.413 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimRtmStopRpfNexthop
Stop tracking (192.168.51.2,232.1.1.1) type <S,G> with 192.168.51.2
pRtmNhop 0x179196078"

47 2017/10/07 18:39:15.413 UTC MINOR: DEBUG #2001 vprn1 PIM[vprn 2 vprn1]
"PIM[vprn 2 vprn1]: pimRtmDelSrcEntry
Deleted src entry for src 192.168.51.2"

48 2017/10/07 18:39:15.413 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0    Total Path Attr Length = 31
    Flag: 0x90 Type: 15 Len: 27 Multiprotocol Unreachable NLRI:
        Address Family MVPN_IPV4 Type: Source-Join Len:22 RD: 64496:101
        SrcAS: 64496 Src: 192.168.51.2 Grp: 232.1.1.1"
```

PMSI using RSVP-TE

Figure 165 shows the details of the topology for VPRN 2.

Figure 165 **VPRN 2 Topology used for RSVP-TE P2MP**



Unicast

For the sake of simplicity, check **Steps 1** to **6** in PMSI using mLDP for VPRN 2 creation information. The same steps are repeated for RSVP, check Figure 165 for details. The result is the configuration in all the PEs, taking as an example PE-1:

```
# on PE-1
configure
    service
        vprn 2 customer 1 create
            description "P2MP RSVP"
            autonomous-system 64496
            route-distinguisher 64496:201
            vrf-target target:64496:200
            interface "loopback" create
                address 172.16.2.1/30
                loopback
            exit
            interface "int-PE-1-CE-5" create
                address 172.16.115.1/30
                sap 1/1/3:2 create
                exit
            exit

            bgp
                group "EXTERNAL"
                    type external
                    peer-as 64505
                    neighbor 172.16.115.2
                    exit
                exit
                no shutdown
            exit
            spoke-sdp 12 create
            exit
            spoke-sdp 13 create
            exit
            spoke-sdp 14 create
            exit
            no shutdown
        exit
```

Because RSVP is the signaling protocol to establish the P2MP LSPs, RSVP is configured on the interfaces. In addition, to use P2MP RSVP, an LSP template is needed. The template defines the characteristics of the LSP to be created, for example, make-before-break, bandwidth, administrative groups, CSPF, specific paths, etc. A basic template is used here. TE parameters specified in the template are commonly used in each RSVP PATH message for each of the branches of the P2MP RSVP LSP. The template is used in the mvpn context within the VPRN configuration (see Auto-Discovery and RSVP PMSI Establishment). The resignal timer for P2MP is configured to the minimum value of sixty minutes (60 — 10080 minutes):

```
# on PE-1
configure
    router
        mpls
            p2mp-resignal-timer 60
            interface system
            exit
            interface int-PE-1-PE-2
            exit
            interface int-PE-1-PE-3
            exit
            path EMPTY
                no shutdown
            exit
            lsp-template VRF2 p2mp
                default-path EMPTY
                cspf
                fast-reroute facility
                exit
                no shutdown
            exit
            no shutdown
        exit
    exit
```

```
*A:PE-1# configure router rsvp no shutdown
```

Auto-Discovery and RSVP PMSI Establishment

The MP-BGP based auto-discovery is implemented with a new address family defined in RFC 4760 MP_REACH_NLRI/MP_UNREACH_NLRI attributes, with AFI 1 (IPv4) or 2 (IPv6) SAFI 5 (temporary value assigned by IANA). This is the mechanism by which each PE advertises the presence of an MVPN to other PEs. This can be achieved using PIM (like in Rosen MVPN) or using BGP. With the default parameter, BGP is automatically chosen because the PMSIs are RSVP and PIM is not an option in this case. Any PE that is a member of an MVPN will advertise to the other PEs using a BGP multi-protocol network layer reachability information (NLRI) update that is sent to all PEs within the AS. This update will contain an Intra-AS I-PMSI auto-discovery route type, also known as an Intra-AD. These use an address family mvpn-ipv4, so each PE must be configured to originate and accept such updates (this was done earlier when configuring the families).

At this step (auto-discovery), the information about the PMSI is exchanged, but the PMSI is not instantiated.

As each PE contains a CE which will be part of the multicast VRF, it is necessary to enable PIM on each interface containing the attachment circuit toward a CE, and to configure the I-PMSI multicast tunnel for the VRF. In order for the BGP routes to be accepted into the VRF a route-target community is required (vrf-target). Although it is not mandatory for the MVPN vrf-target to be equal to the unicast target, Nokia recommends to use vrf-target unicast to avoid configuration mistakes and extra complexity.

On each PE, the multicast configuration in the VPRN instance is as follows:

```
# on PE-1
configure service
        vprn 2
            pim
                interface "loopback"
                exit
                interface "int-PE-1-CE-5"
                exit
            exit
            mvpn
                auto-discovery default
                c-mcast-signaling bgp
                provider-tunnel
                    inclusive
                        rsvp
                            lsp-template "VRF2"
                            no shutdown
                        exit
                    exit
                exit
                vrf-target unicast
                exit
            exit
```

The status of VPRN 2 on PE-1 is shown with the following output:

```
*A:PE-1# show router 2 mvpn
===============================================================================
MVPN 2 configuration data
===============================================================================
signaling          : Bgp                  auto-discovery     : Default
UMH Selection      : Highest-Ip           SA withdrawn       : Disabled
intersite-shared   : Enabled              Persist SA         : Disabled
vrf-import         : N/A
vrf-export         : N/A
vrf-target         : unicast
C-Mcast Import RT  : target:192.0.2.1:3

ipmsi              : rsvp VRF2
i-pmsi P2MP AdmSt  : Up
i-pmsi Tunnel Name : VRF2-2-73732
enable-bfd-root    : false                enable-bfd-leaf    : false
Mdt-type           : sender-receiver

BSR signalling     : none
Wildcard s-pmsi    : Disabled
Multistream-SPMSI  : Disabled
s-pmsi             : none
data-delay-interval: 3 seconds
enable-asm-mdt     : N/A
===============================================================================
*A:PE-1#
```

The following shows a debug of an Intra-AD BGP update message received by PE-1 that was sent by PE-4. The message contains the PMSI tunnel-type to be used (RSVP P2MP LSP), the P2MP LSP ID (encoded as extended tunnel ID and P2MP-ID carried in the RSVP Session object), and the type of BGP update (Type: Intra-AD Len: 12 RD: 64496:204 Orig: 192.0.2.4):

```
29 2017/10/07 18:47:40.709 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.4
"Peer 1: 192.0.2.4: UPDATE
Peer 1: 192.0.2.4 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 86
    Flag: 0x90 Type: 14 Len: 23 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.4
        Type: Intra-AD Len: 12 RD: 64496:204 Orig: 192.0.2.4
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64496:200
    Flag: 0xc0 Type: 22 Len: 17 PMSI:
        Tunnel-type RSVP-TE P2MP LSP (1)
        Flags: (0x0)[Type: None BM: 0 U: 0 Leaf: not required]
        MPLS Label 0
        P2MP-ID 0x2, Tunnel-ID: 61442, Extended-Tunnel-ID 192.0.2.4"
```

The setup has four PEs, so every PE should see the others peer Intra-AD route; the following output shows the routes received in PE-1:

```
*A:PE-1# show router bgp routes mvpn-ipv4 type intra-ad
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
Flag  RouteType                   OriginatorIP           LocalPref   MED
      RD                          SourceAS               Path-Id     Label
      Nexthop                     SourceIP
      As-Path                     GroupIP
-------------------------------------------------------------------------------
u*>i  Intra-Ad                    192.0.2.2              100         0
      64496:202                   -                      None        -
      192.0.2.2                   -
      No As-Path                  -
u*>i  Intra-Ad                    192.0.2.3              100         0
      64496:203                   -                      None        -
      192.0.2.3                   -
      No As-Path                  -
u*>i  Intra-Ad                    192.0.2.4              100         0
      64496:204                   -                      None        -
      192.0.2.4                   -
      No As-Path                  -
-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:PE-1#
```

The detailed output of the Intra-AD received from PE-4 shows the tunnel-type RSVP-TE P2MP LSP (P2MP-ID is 2), the originator id (192.0.2.4), and the route-distinguisher (64496:204):

```
*A:PE-1# show router bgp routes mvpn-ipv4 type intra-ad originator-ip 192.0.2.4 detail
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
```

```
===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
Original Attributes

Route Type     : Intra-Ad
Route Dist.    : 64496:204
Originator IP  : 192.0.2.4
Nexthop        : 192.0.2.4
Path Id        : None
From           : 192.0.2.4
Res. Nexthop   : 0.0.0.0
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : no-export target:64496:200
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.4
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h01m26s
VPRN Imported  :  2
-------------------------------------------------------------------------------
PMSI Tunnel Attributes :
Tunnel-type    : RSVP-TE P2MP LSP
Flags          : Type: RNVE(0) BM: 0 U: 0 Leaf: not required
MPLS Label     : 0
P2MP-ID        : 2                      Tunnel-ID      : 61442
Extended-Tunne*: 192.0.2.4
-------------------------------------------------------------------------------

Modified Attributes

--- snipped ---

-------------------------------------------------------------------------------
Routes : 1
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-1#
```

For the I-PMSI, the head-end PE firstly discovers all the leaf PEs via I-PMSI A-D routes, it then signals the P2MP LSP to all the leaf PEs using RSVP-TE (subsequently adding or removing S2L (source to leaf) paths as PEs are added or removed from the MVPN).

As in the mLDP case, the demarcation of the domains is in the PE. The PE router participates in both the customer multicast domain and the provider’s multicast domain. The customer’s CEs are limited to a multicast adjacency with the multicast instance on the PE created to support that specific customer’s IP-VPN. This way, customers are isolated from the provider’s core multicast domain and other customer multicast domains while the provider’s core P routers only participate in the provider’s multicast domain and are isolated from all customers’s multicast domains. C-trees to P-tunnels bindings are also discovered using BGP routes, instead of PIM join TLVs. MVPN c-multicast routing information is exchanged between PEs by using c-multicast routes that are carried using MCAST-VPN NLRIs.

Once the RSVP-TE P2MP LSPs are created, the I-PMSI is instantiated in the core:

```
*A:PE-1# show router 2 pim neighbor

===============================================================================
PIM Neighbor ipv4
===============================================================================
Interface               Nbr DR Prty     Up Time       Expiry Time     Hold Time
   Nbr Address
-------------------------------------------------------------------------------
int-PE-1-CE-5           1               0d 00:01:23   0d 00:01:24     105
   172.16.115.2
mpls-if-73733           1               0d 00:02:03   never           65535
   192.0.2.2
mpls-if-73734           1               0d 00:01:48   never           65535
   192.0.2.3
mpls-if-73735           1               0d 00:01:38   never           65535
   192.0.2.4
-------------------------------------------------------------------------------
Neighbors : 4
===============================================================================
*A:PE-1#
```

```
*A:PE-1# show router 2 pim tunnel-interface

===============================================================================
PIM Interfaces ipv4
===============================================================================
Interface                        Originator Address   Adm  Opr  Transport Type
-------------------------------------------------------------------------------
mpls-if-73732                   192.0.2.1            Up   Up   Tx-IPMSI
mpls-if-73733                   192.0.2.2            Up   Up   Rx-IPMSI
mpls-if-73734                   192.0.2.3            Up   Up   Rx-IPMSI
mpls-if-73735                   192.0.2.4            Up   Up   Rx-IPMSI
-------------------------------------------------------------------------------
Interfaces : 4
===============================================================================
*A:PE-1#
```

The following command displays the PMSIs created on a PE, taking PE-3 as an example:

```
*A:PE-3# show router 2 pim tunnel-interface
```

```
===============================================================================
PIM Interfaces ipv4
===============================================================================
Interface                        Originator Address   Adm  Opr  Transport Type
-------------------------------------------------------------------------------
mpls-if-73732                    192.0.2.3            Up   Up   Tx-IPMSI
mpls-if-73733                    192.0.2.1            Up   Up   Rx-IPMSI
mpls-if-73734                    192.0.2.2            Up   Up   Rx-IPMSI
mpls-if-73735                    192.0.2.4            Up   Up   Rx-IPMSI
-------------------------------------------------------------------------------
Interfaces : 4
===============================================================================
*A:PE-3#

*A:PE-3# tools dump router 2 mvpn provider-tunnels

===============================================================================
MVPN 2 Inclusive Provider Tunnels Originating
===============================================================================
ipmsi (RSVP)                            P2MP-ID     Tunl-ID     Ext-Tunl-ID
-------------------------------------------------------------------------------
VRF2-2-73732                            2           61441       192.0.2.3

===============================================================================
MVPN 2 Selective Provider Tunnels Originating
===============================================================================
spmsi (RSVP)                            P2MP-ID     Tunl-ID     Ext-Tunl-ID
-------------------------------------------------------------------------------

No Tunnels Found
-------------------------------------------------------------------------------

===============================================================================
MVPN 2 Inclusive Provider Tunnels Terminating
===============================================================================
ipmsi (RSVP)                            P2MP-ID     Tunl-ID     Ext-Tunl-ID
-------------------------------------------------------------------------------
mpls-if-73733                           2           61441       192.0.2.1
mpls-if-73734                           2           61441       192.0.2.2
mpls-if-73735                           2           61442       192.0.2.4

===============================================================================
MVPN 2 Selective Provider Tunnels Terminating
===============================================================================
spmsi (RSVP)                            P2MP-ID     Tunl-ID     Ext-Tunl-ID
-------------------------------------------------------------------------------

No Tunnels Found
-------------------------------------------------------------------------------
*A:PE-3#
```

Every PE has created an I-PMSI to the other PEs. As an example, PE-1 has established an LSP with name VRF2-2-73732 with PE-2, PE-3 and PE-4 as leaves. The S2L path is empty because the template did not have any S2L path configured for simplicity.

```
*A:PE-1# show router mpls p2mp-lsp detail

===============================================================================
MPLS P2MP LSPs (Originating) (Detail)
===============================================================================
Legend :
    + - Inherited
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name        : VRF2-2-73732
LSP Type        : P2mpAutoLsp                LSP Tunnel ID        : 61441
LSP Index       : 61441                      TTM Tunnel Id        : 61441
From            : 192.0.2.1
Adm State       : Up                         Oper State           : Up
LSP Up Time     : 0d 00:02:59                LSP Down Time        : 0d 00:00:00
Transitions     : 1                          Path Changes         : 1
Retry Limit     : 0                          Retry Timer          : 30 sec
Signaling       : RSVP                       Resv. Style          : SE
Hop Limit       : 255                        Negotiated MTU       : n/a
Adaptive        : Enabled                    ClassType            : 0
FastReroute     : Enabled                    Oper FR              : Enabled
FR Method       : Facility                   FR Hop Limit         : 16
FR Node Protect : Disabled                   FR Prop Adm Grp      : Disabled
FR Object       : Enabled
CSPF            : Enabled                    ADSPEC               : Disabled
Metric          : Disabled                   Use TE metric        : Disabled
Load Bal Wt     : N/A                        ClassForwarding      : Disabled
Include Grps    :                            Exclude Grps         :
None                                            None
Least Fill      : Disabled

Revert Timer    : Disabled                   Next Revert In       : N/A
Auto BW         : Disabled
LdpOverRsvp     : Enabled
VprnAutoBind    : Enabled
IGP Shortcut    : Enabled                    BGP Shortcut         : Enabled
IGP LFA         : Disabled                   IGP Rel Metric       : Disabled
BGPTransTun     : Enabled
Oper Metric     : Disabled
Prop Adm Grp    : Disabled

P2MPInstance    : 2                          P2MP-Inst-type       : Primary
S2L Cfg Counter : 3                          S2L Oper Counter     : 3
S2l-Name        : EMPTY                      To                   : 192.0.2.2
S2l-Name        : EMPTY                      To                   : 192.0.2.3
S2l-Name        : EMPTY                      To                   : 192.0.2.4
===============================================================================
*A:PE-1#
```

Checking the RSVP-TE P2MP LSPs that are originated, transit, or destination to PE-1, the show command allows filtering by type, in this case showing the originated LSPs only:

```
*A:PE-1# show router mpls p2mp-info type originate

===============================================================================
MPLS P2MP LSPs (Originate)
===============================================================================
-------------------------------------------------------------------------------
S2L VRF2-2-73732::EMPTY
-------------------------------------------------------------------------------
Source IP Address    : 192.0.2.1              Tunnel ID     : 61441
P2MP ID              : 2                      Lsp ID        : 58880
S2L Name             : VRF2-2-73732::EMPTY    To            : 192.0.2.2
Out Interface        : 1/1/1                  Out Label     : 262138
Num. of S2ls         : 2
-------------------------------------------------------------------------------
S2L VRF2-2-73732::EMPTY
-------------------------------------------------------------------------------
Source IP Address    : 192.0.2.1              Tunnel ID     : 61441
P2MP ID              : 2                      Lsp ID        : 58880
S2L Name             : VRF2-2-73732::EMPTY    To            : 192.0.2.3
Out Interface        : 1/1/2                  Out Label     : 262136
Num. of S2ls         : 1
-------------------------------------------------------------------------------
S2L VRF2-2-73732::EMPTY
-------------------------------------------------------------------------------
Source IP Address    : 192.0.2.1              Tunnel ID     : 61441
P2MP ID              : 2                      Lsp ID        : 58880
S2L Name             : VRF2-2-73732::EMPTY    To            : 192.0.2.4
Out Interface        : 1/1/1                  Out Label     : 262138
Num. of S2ls         : 2
-------------------------------------------------------------------------------
P2MP Cross-connect instances : 3
===============================================================================
*A:PE-1#
```

Following the path of the S2L from PE-1 to PE-4 (third entry S2L VRF2-2-73732), the outgoing interface is 1/1/1 that connects PE-1 to PE-2, so the LSP goes to PE-4 via PE-2. The return path need not be via PE-2; it may be via PE-3.

```
*A:PE-2# show router mpls p2mp-info type transit

===============================================================================
MPLS P2MP LSPs (Transit)
===============================================================================
-------------------------------------------------------------------------------
S2L VRF2-2-73732::EMPTY
-------------------------------------------------------------------------------
Source IP Address    : 192.0.2.1              Tunnel ID     : 61441
P2MP ID              : 2                      Lsp ID        : 37376
S2L Name             : VRF2-2-73732::EMPTY    To            : 192.0.2.4
Out Interface        : 1/1/1                  Out Label     : 262131
Num. of S2ls         : 1
-------------------------------------------------------------------------------
S2L VRF2-2-73732::EMPTY
-------------------------------------------------------------------------------
Source IP Address    : 192.0.2.4              Tunnel ID     : 61442
P2MP ID              : 2                      Lsp ID        : 53248
S2L Name             : VRF2-2-73732::EMPTY    To            : 192.0.2.1
Out Interface        : 1/1/2                  Out Label     : 262131
Num. of S2ls         : 1
-------------------------------------------------------------------------------
P2MP Cross-connect instances : 2
```

```
===============================================================================
*A:PE-2#
```

As transit, PE-2 shows that there is an LSP coming from PE-1 (VRF2-2-73732) and the outgoing interface is 1/1/1 that connects PE-2 with PE-4.

Using the same command with a different filter on PE-4, 3 P2MP LSPs are terminated, one from each remote PE (PE-1, PE-2 and PE-3). On PE-4, an S2L VRF2-2-73732 from 192.0.2.1 and P2MP ID = 2 is traced.

```
*A:PE-4# show router mpls p2mp-info type terminate

===============================================================================
MPLS P2MP LSPs (Terminate)
===============================================================================
-------------------------------------------------------------------------------
S2L VRF2-2-73732::EMPTY
-------------------------------------------------------------------------------
Source IP Address    : 192.0.2.1            Tunnel ID     : 61441
P2MP ID              : 2                    Lsp ID        : 58880
S2L Name             : VRF2-2-73732::EMPTY   To            : 192.0.2.4
In Interface         : 1/1/2                In Label      : 262131
Num. of S2ls         : 1
-------------------------------------------------------------------------------
S2L VRF2-2-73732::EMPTY
-------------------------------------------------------------------------------
Source IP Address    : 192.0.2.2            Tunnel ID     : 61441
P2MP ID              : 2                    Lsp ID        : 56320
S2L Name             : VRF2-2-73732::EMPTY   To            : 192.0.2.4
In Interface         : 1/1/2                In Label      : 262136
Num. of S2ls         : 2
-------------------------------------------------------------------------------
S2L VRF2-2-73732::EMPTY
-------------------------------------------------------------------------------
Source IP Address    : 192.0.2.3            Tunnel ID     : 61441
P2MP ID              : 2                    Lsp ID        : 37888
S2L Name             : VRF2-2-73732::EMPTY   To            : 192.0.2.4
In Interface         : 1/1/1                In Label      : 262132
Num. of S2ls         : 1
-------------------------------------------------------------------------------
P2MP Cross-connect instances : 3
===============================================================================
*A:PE-4#
```

The following output shows the P2MP LSP on PE-1 with more detail:

```
*A:PE-1# show router mpls p2mp-lsp VRF2-2-73732"VRF2-2-73732" p2mp-instance 2"2" s2l
EMPTY detail

===============================================================================
MPLS LSP VRF2-2-73732 S2L EMPTY (Detail)
===============================================================================
Legend :
    @ - Detour Available                          # - Detour In Use
    b - Bandwidth Protected                       n - Node Protected
    S - Strict                                    L - Loose
    A - ABR
    s - Soft Preemption
===============================================================================
-------------------------------------------------------------------------------
LSP VRF2-2-73732 S2L EMPTY
-------------------------------------------------------------------------------
LSP Name         : VRF2-2-73732         S2l LSP ID           : 58880
P2MP ID          : 2                    S2l Grp Id           : 1
Adm State        : Up                   Oper State           : Up
S2l State:       : Active                                    :
S2L Name         : EMPTY                To                   : 192.0.2.2
S2l Admin        : Up                   S2l Oper             : Up
OutInterface     : 1/1/1                Out Label            : 262138
S2L Up Time      : 0d 00:03:42          S2L Dn Time          : 0d 00:00:00
RetryAttempt     : 0                    NextRetryIn          : 0 sec
S2L Trans        : 1                    CSPF Queries         : 1
Failure Code     : noError              Failure Node         : n/a
Inter-area       : False
ExplicitHops     :
    No Hops Specified
Actual Hops      :
    192.168.12.1 (192.0.2.1) @                   Record Label        : N/A
 -> 192.168.12.2 (192.0.2.2)                     Record Label        : 262138
ComputedHops     :
    192.168.12.1(S)
 -> 192.168.12.2(S)
LastResignal     : n/a
-------------------------------------------------------------------------------
LSP VRF2-2-73732 S2L EMPTY
-------------------------------------------------------------------------------
LSP Name         : VRF2-2-73732         S2l LSP ID           : 58880
P2MP ID          : 2                    S2l Grp Id           : 2
Adm State        : Up                   Oper State           : Up
S2l State:       : Active                                    :
S2L Name         : EMPTY                To                   : 192.0.2.3
S2l Admin        : Up                   S2l Oper             : Up
OutInterface     : 1/1/2                Out Label            : 262136
S2L Up Time      : 0d 00:03:27          S2L Dn Time          : 0d 00:00:00
RetryAttempt     : 0                    NextRetryIn          : 0 sec
S2L Trans        : 1                    CSPF Queries         : 1
Failure Code     : noError              Failure Node         : n/a
Inter-area       : False
ExplicitHops     :
    No Hops Specified
Actual Hops      :
    192.168.13.1 (192.0.2.1) @                   Record Label        : N/A
 -> 192.168.13.2 (192.0.2.3)                     Record Label        : 262136
ComputedHops     :
    192.168.13.1(S)
 -> 192.168.13.2(S)
LastResignal     : n/a
-------------------------------------------------------------------------------
LSP VRF2-2-73732 S2L EMPTY
-------------------------------------------------------------------------------
LSP Name         : VRF2-2-73732         S2l LSP ID           : 58880
P2MP ID          : 2                    S2l Grp Id           : 3
Adm State        : Up                   Oper State           : Up
S2l State:       : Active                                    :
S2L Name         : EMPTY                To                   : 192.0.2.4
```

```
S2l Admin        : Up                   S2l Oper            : Up
OutInterface     : 1/1/1                Out Label           : 262138
S2L Up Time      : 0d 00:03:17          S2L Dn Time         : 0d 00:00:00
RetryAttempt     : 0                    NextRetryIn         : 0 sec
S2L Trans        : 1                    CSPF Queries        : 1
Failure Code     : noError              Failure Node        : n/a
Inter-area       : False
ExplicitHops     :
    No Hops Specified
Actual Hops      :
    192.168.12.1 (192.0.2.1) @                   Record Label        : N/A
 -> 192.168.12.2 (192.0.2.2) @                   Record Label        : 262138
 -> 192.168.24.2 (192.0.2.4)                     Record Label        : 262131
ComputedHops     :
    192.168.12.1(S)
 -> 192.168.12.2(S)
 -> 192.168.24.2(S)
LastResignal     : n/a
===============================================================================
*A:PE-1#
```

The last entry, VRF2-2-73732, provides the details of the S2L traced earlier, displaying the different hops (PE-1, PE-2, and PE-4), the fast reroute protection (link protection is supported only) and the labels used (262138 from PE-1 to PE-2, 262131 from PE-2 to PE-4). On PE-1, although only one has been shown, both links PE-1 to PE-3 and PE-1 to PE-2 are fast reroute protected.

If any of the protected links between PE-1 and PE-2 or PE-3 are broken, fast reroute will be initiated. The protected bypass hops are displayed with the following command:

```
*A:PE-1# show router mpls bypass-tunnel protected-lsp p2mp detail

===============================================================================
MPLS Bypass Tunnels (Detail)
===============================================================================
-------------------------------------------------------------------------------
bypass-link192.168.12.2-61442
-------------------------------------------------------------------------------
To             : 192.168.24.1        State               : Up
Out I/F        : 1/1/2               Out Label           : 262138
Up Time        : 0d 00:03:55         Active Time         : n/a
Reserved BW    : 0 Kbps              Protected LSP Count : 3
Type           : P2mp                Bypass Path Cost    : 30
Setup Priority : 7                   Hold Priority       : 0
Class Type     : 0
Exclude Node   : None                Inter-Area          : False
Computed Hops  :
    192.168.13.1(S)                  Egress Admin Groups : None
 -> 192.168.13.2(S)                  Egress Admin Groups : None
 -> 192.168.34.2(S)                  Egress Admin Groups : None
 -> 192.168.24.1(S)                  Egress Admin Groups : None
Actual Hops    :
    192.168.13.1 (192.0.2.1)         Record Label        : N/A
 -> 192.168.13.2 (192.0.2.3)         Record Label        : 262138
 -> 192.168.34.2 (192.0.2.4)         Record Label        : 262138
 -> 192.168.24.1 (192.0.2.2)         Record Label        : 262137

Protected LSPs -
LSP Name       : VRF2-2-73732::EMPTY
From           : 192.0.2.1           To                  : 192.0.2.2
Avoid Node/Hop : 192.168.12.2        Downstream Label    : 262138
Bandwidth      : 0 Kbps

LSP Name       : VRF2-2-73732::EMPTY
From           : 192.0.2.3           To                  : 192.0.2.2
Avoid Node/Hop : 192.168.12.2        Downstream Label    : 262134
Bandwidth      : 0 Kbps

LSP Name       : VRF2-2-73732::EMPTY
From           : 192.0.2.1           To                  : 192.0.2.4
Avoid Node/Hop : 192.168.12.2        Downstream Label    : 262138
Bandwidth      : 0 Kbps

-------------------------------------------------------------------------------
bypass-link192.168.13.2-61443
-------------------------------------------------------------------------------
To             : 192.168.34.1        State               : Up
Out I/F        : 1/1/1               Out Label           : 262136
Up Time        : 0d 00:03:40         Active Time         : n/a
Reserved BW    : 0 Kbps              Protected LSP Count : 1
Type           : P2mp                Bypass Path Cost    : 30
Setup Priority : 7                   Hold Priority       : 0
Class Type     : 0
Exclude Node   : None                Inter-Area          : False
Computed Hops  :
    192.168.12.1(S)                  Egress Admin Groups : None
 -> 192.168.12.2(S)                  Egress Admin Groups : None
 -> 192.168.24.2(S)                  Egress Admin Groups : None
 -> 192.168.34.1(S)                  Egress Admin Groups : None
Actual Hops    :
    192.168.12.1 (192.0.2.1)         Record Label        : N/A
 -> 192.168.12.2 (192.0.2.2)         Record Label        : 262136
 -> 192.168.24.2 (192.0.2.4)         Record Label        : 262135
 -> 192.168.34.1 (192.0.2.3)         Record Label        : 262134

Protected LSPs -
LSP Name       : VRF2-2-73732::EMPTY
From           : 192.0.2.1           To                  : 192.0.2.3
Avoid Node/Hop : 192.168.13.2        Downstream Label    : 262136
Bandwidth      : 0 Kbps

===============================================================================
*A:PE-1#
```

Traffic Flow

The receiver H-8, connected to CE-8, wishes to join the group 232.2.2.2 with source 192.168.52.1 and so sends an IGMPv3 report toward CE-8. CE-8 recognizes the report and sends a PIM join toward the source, therefore, it reaches PE-1 where the source is connected to through CE-5. The following output shows the debug seen on PE-4, where the PIM join is received from CE-8 and a BGP update Source Join is sent to all PEs (only the update sent to PE-1 is shown).

```
1 2017/10/07 20:30:34.032 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimJPProcessSG
pimJPProcessSG: (S,G)-> (192.168.52.2,232.2.2.2) type <S,G>,
i/f int-PE-4-CE-8, upNbr 172.16.148.1 isJoin 1 isRpt 0 holdTime 210"

2 2017/10/07 20:30:34.032 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimRtmFindRpfNexthop
Track (192.168.52.2,232.2.2.2) type <S,G> using 192.168.52.2"

3 2017/10/07 20:30:34.032 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimRtmAddSrcEntry
Added src entry for src 192.168.52.2"

4 2017/10/07 20:30:34.032 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimJPPrintFsmEvent
PIM JP Downstream: State NoInfo Event RxJoin StandbyEvent F,
(S,G) (192.168.52.2,232.2.2.2) groupType <S,G>"

5 2017/10/07 20:30:34.032 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimJPPrintFsmEvent
PIM JP Upstream: State NotJoined Event JoinDesiredTrue StandbyEvent F,
(S,G) (192.168.52.2,232.2.2.2) groupType <S,G>"

6 2017/10/07 20:30:34.032 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimSGUpJoinDesiredTrue
No upstream interface. pSG (192.168.52.2,232.2.2.2) rpfType 3"

7 2017/10/07 20:30:34.032 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimSGUpJoinDesiredTrue
No upstream interface SG (192.168.52.2,232.2.2.2) rpfType 3"

8 2017/10/07 20:30:34.032 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimRtmProcessNhresEvent
RTM-Nhres Event U-RTM NEW Src 192.168.52.2 SrcRtmUse UCAST"

9 2017/10/07 20:30:34.032 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimRtmProcessNhresEvent
Prefix 192.168.52.0/24 numNextHops 1 owner BGP_VPN metric 20 pref 170"

10 2017/10/07 20:30:34.032 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimRtmSrcResolveSGsInt
Trying to resolve SG (192.168.52.2,232.2.2.2)"

11 2017/10/07 20:30:34.032 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimRtmNotifyRpfChange
RPF Change to Source/RP 192.168.52.2  for SG (192.168.52.2,232.2.2.2) dynMLDP F via
 NH 192.0.2.1 IfIdx: 73735 RpfType: REMOTE Reason:  RTE_ADD old NH 0.0.0.0
IfIdx: 0 RpfType: NONE mplsRpf F NextHops 1 reg 1/1 lfa 0/0"
```

```
12 2017/10/07 20:30:34.032 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimRtmNotifyRpfChange
SG (192.168.52.2,232.2.2.2) Source/RP 192.168.52.2 Ipmsi 73732 NhIf 0 new NhIf 73735"

13 2017/10/07 20:30:34.032 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimJPPrintFsmEvent
PIM JP Upstream: State Joined Event MribChange StandbyEvent F,
(S,G) (192.168.52.2,232.2.2.2) groupType <S,G>"

14 2017/10/07 20:30:34.032 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimSGUpStateJMribChange
SG (192.168.52.2,232.2.2.2), type <S,G> oldMribNhopIp 0.0.0.0 oldRpfNbrIp 0.0.0.0,
 oldRpfType NONE oldRpfIf 0 rptMribNhopIp 0.0.0.0, rptRpfNbrIp 0.0.0.0 rtmReason 48
 isSGExtNet : no"

15 2017/10/07 20:30:34.032 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimSGUpStateJMribChange
SG (192.168.52.2,232.2.2.2), type <S,G> newMribNhopIp 192.0.2.1
newRpfNbrIp 192.0.2.1 newRpfType REMOTE newRpfIf 73735"

16 2017/10/07 20:30:34.032 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimAddToJPTxPdu
pimAddToJPTxPdu: (S,G)-> (192.168.52.2,232.2.2.2), type <S,G>, txPendFlag J
isStandby F"

17 2017/10/07 20:30:34.032 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimRtmUpdateSGMetric
SG metric 4294967295 pref 2147483647, new metric 20 pref 170"

--- snipped ---

20 2017/10/07 20:30:34.033 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 76
    Flag: 0x90 Type: 14 Len: 33 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.4
        Type: Source-Join Len:22 RD: 64496:201 SrcAS: 64496
                          Src: 192.168.52.2 Grp: 232.2.2.2
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:192.0.2.1:3
"
```

The following debug shows that PE-1 receives the BGP update Source Join with source 192.168.52.1 and group 232.2.2.2 and sends a PIM join toward CE-5:

```
1 2017/10/07 20:30:34.034 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.4
"Peer 1: 192.0.2.4: UPDATE
Peer 1: 192.0.2.4 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 76
    Flag: 0x90 Type: 14 Len: 33 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.4
        Type: Source-Join Len:22 RD: 64496:201 SrcAS: 64496
                          Src: 192.168.52.2 Grp: 232.2.2.2
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:192.0.2.1:3
"

2 2017/10/07 20:30:34.034 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimProcessMvpnRouteMsg
originator 0.0.0.0: add rtType SOURCE_TREE_JOIN nextHop 192.0.2.4 source 192.168.52.2
group 232.2.2.2"

3 2017/10/07 20:30:34.034 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimJPProcessSG
pimJPProcessSG: (S,G)-> (192.168.52.2,232.2.2.2) type <S,G>,
i/f mpls-if-73732, upNbr 192.0.2.1 isJoin 1 isRpt 0 holdTime 65535"

4 2017/10/07 20:30:34.034 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimRtmFindRpfNexthop
Track (192.168.52.2,232.2.2.2) type <S,G> using 192.168.52.2"

5 2017/10/07 20:30:34.034 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimRtmAddSrcEntry
Added src entry for src 192.168.52.2"

6 2017/10/07 20:30:34.034 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimJPPrintFsmEvent
PIM JP Downstream: State NoInfo Event RxJoin StandbyEvent F,
(S,G) (192.168.52.2,232.2.2.2) groupType <S,G>"

7 2017/10/07 20:30:34.034 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimJPPrintFsmEvent
PIM JP Upstream: State NotJoined Event JoinDesiredTrue StandbyEvent F, (
S,G) (192.168.52.2,232.2.2.2) groupType <S,G>"

8 2017/10/07 20:30:34.034 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimAddToJPTxPdu
pimAddToJPTxPdu: (S,G)-> (192.168.52.2,232.2.2.2), type <S,G>,
txPendFlag J isStandby F"

9 2017/10/07 20:30:34.034 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimRtmProcessNhresEvent
RTM-Nhres Event U-RTM NEW Src 192.168.52.2 SrcRtmUse UCAST"

10 2017/10/07 20:30:34.034 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimRtmProcessNhresEvent
Prefix 192.168.52.0/24 numNextHops 1 owner BGP metric 0 pref 170"

--- snipped ---
```

```
19 2017/10/07 20:30:34.034 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimSGEncodeGroupSet
Encoding  Join for source 192.168.52.2"

20 2017/10/07 20:30:34.034 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimSGEncodeGroupSet
num joined srcs 1, num pruned srcs 0"

21 2017/10/07 20:30:34.034 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimSendJoinPrunePdu
sending JP PDU with 1 groups, if 7 adj 172.16.115.2"
```

The BGP update source join received by PE-1 is displayed with the following command:

```
*A:PE-1# show router bgp routes mvpn-ipv4 type source-join
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
Flag  RouteType                   OriginatorIP           LocalPref   MED
      RD                          SourceAS               Path-Id     Label
      Nexthop                     SourceIP
      As-Path                     GroupIP
-------------------------------------------------------------------------------
u*>i  Source-Join                 -                      100         0
      64496:201                   64496                  None        -
      192.0.2.4                   192.168.52.2
      No As-Path                  232.2.2.2
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-1#
```

To verify the traffic: on PE-1, there is a group 232.2.2.2 with source 192.168.52.1, the RPF is CE-5, and the multicast traffic is flowing from CE-5 to PE-1 using int-PE-1-CE-5 and the outgoing interface is using the PMSI RSVP mpls-if-73732.

```
*A:PE-1# show router 2 pim group detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 232.2.2.2
Source Address     : 192.168.52.2
RP Address         : 0
Advt Router        : 172.16.115.2
Flags              :                    Type               : (S,G)
```

```
Mode                : sparse
MRIB Next Hop       : 172.16.115.2
MRIB Src Flags      : remote
Keepalive Timer     : Not Running
Up Time             : 0d 00:00:36        Resolved By        : rtable-u

Up JP State         : Joined             Up JP Expiry       : 0d 00:00:24
Up JP Rpt           : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State      : No Info
Reg From Anycast RP: No

Rpf Neighbor        : 172.16.115.2
Incoming Intf       : int-PE-1-CE-5
Outgoing Intf List : mpls-if-73732

Curr Fwding Rate    : 1018.6 kbps
Forwarded Packets   : 3022               Discarded Packets  : 0
Forwarded Octets    : 4526956            RPF Mismatches     : 0
Spt threshold       : 0 kbps             ECMP opt threshold : 7
Admin bandwidth     : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-1#
```

On PE-4, the same (S,G) arrives in the incoming interface mpls-if-73734 and the outgoing interface is int-PE-4-CE-8.

```
*A:PE-4# show router 2 pim group detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address       : 232.2.2.2
Source Address      : 192.168.52.2
RP Address          : 0
Advt Router         : 192.0.2.1
Flags               :                    Type               : (S,G)
Mode                : sparse
MRIB Next Hop       : 192.0.2.1
MRIB Src Flags      : remote
Keepalive Timer     : Not Running
Up Time             : 0d 00:00:41        Resolved By        : rtable-u

Up JP State         : Joined             Up JP Expiry       : 0d 00:00:18
Up JP Rpt           : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State      : No Info
Reg From Anycast RP: No

Rpf Neighbor        : 192.0.2.1
Incoming Intf       : mpls-if-73735
Outgoing Intf List : int-PE-4-CE-8

Curr Fwding Rate    : 1018.6 kbps
Forwarded Packets   : 3476               Discarded Packets  : 0
Forwarded Octets    : 5207048            RPF Mismatches     : 0
```

```
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-4#
```

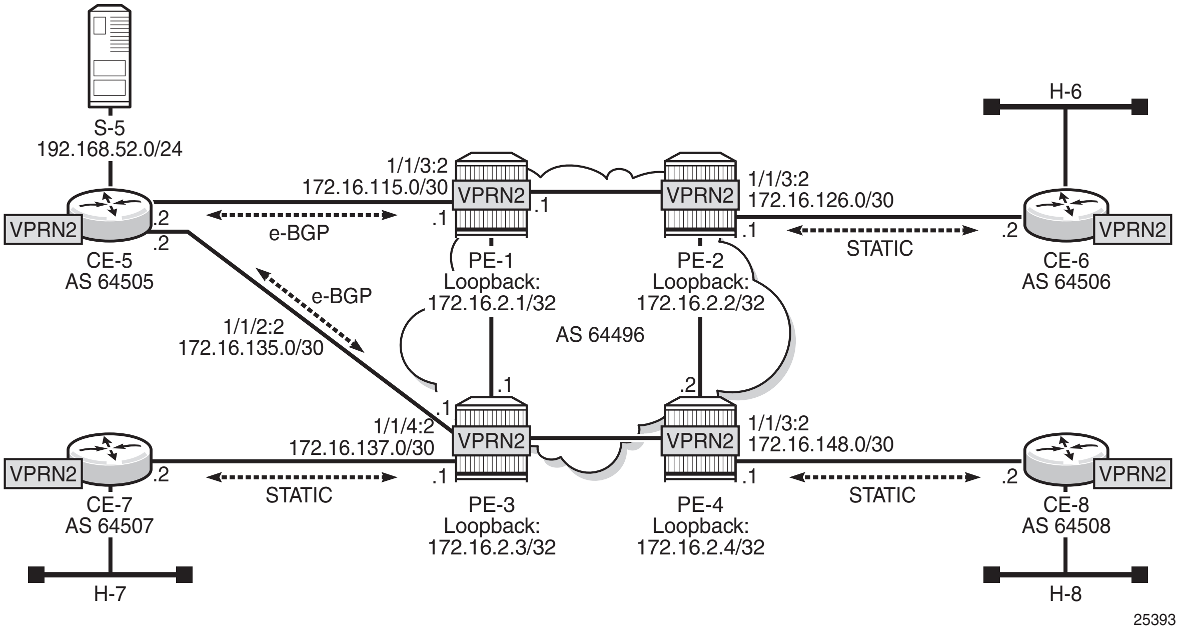
When the receiver is not interested in the channel group anymore, the receiver H-8 sends an IGMPv3 leave, PE-4 sends a PIM prune translated to a BGP MP_UNREACH NLRI to all PEs. As mentioned before, rapid withdrawals are sent without waiting for the MRAI (for simplicity, only one BGP update is shown in the output debug).

```
21 2017/10/07 20:36:55.424 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimJPProcessSG
pimJPProcessSG: (S,G)-> (192.168.52.2,232.2.2.2) type <S,G>,
i/f int-PE-4-CE-8, upNbr 172.16.148.1 isJoin 0 isRpt 0 holdTime 210"

22 2017/10/07 20:36:55.424 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimJPPrintFsmEvent
PIM JP Downstream: State Joined Event RxPrune StandbyEvent F,
(S,G) (192.168.52.2,232.2.2.2) groupType <S,G>"

23 2017/10/07 20:36:55.424 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimJPPrintFsmEvent
PIM JP Downstream: State PrunePending Event PrunePendTimerExp StandbyEvent F,
(S,G) (192.168.52.2,232.2.2.2) groupType <S,G>"

24 2017/10/07 20:36:55.424 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimJPPrintFsmEvent
PIM JP Upstream: State Joined Event JoinDesiredFalse StandbyEvent F,
(S,G) (192.168.52.2,232.2.2.2) groupType <S,G>"

25 2017/10/07 20:36:55.424 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimAddToJPTxPdu
pimAddToJPTxPdu: (S,G)-> (192.168.52.2,232.2.2.2),
type <S,G>, txPendFlag P isStandby F"

26 2017/10/07 20:36:55.424 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimRtmStopRpfNexthop
Stop tracking (192.168.52.2,232.2.2.2)
type <S,G> with 192.168.52.2 pRtmNhop 0x179195f48"

27 2017/10/07 20:36:55.424 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimRtmDelSrcEntry
Deleted src entry for src 192.168.52.2"

28 2017/10/07 20:36:55.424 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 31
    Flag: 0x90 Type: 15 Len: 27 Multiprotocol Unreachable NLRI:
        Address Family MVPN_IPV4
        Type: Source-Join Len:22 RD: 64496:201 SrcAS: 64496
                                 Src: 192.168.52.2 Grp: 232.2.2.2
"
```

MVPN Source Redundancy

So far, the multicast traffic has been streamed toward router CE-5 from a single source. For security, the source can be redundant (two sources attached to different CEs that connect to a pair of PEs). To simulate the redundancy, CE-5 has been connected to both PE-1 and PE-3, using VPRN 2, and equal cost multi-path (ECMP) is configured with the value of 2 in all PEs. With this configuration, any PE is able to reach the source through PE-1 and PE-2. The (S,G) is the same as the one used in P2MP RSVP TE (192.168.52.1, 232.2.2.2). Figure 166 shows the VPRN 2 topology with the source redundancy.

Figure 166 **VPRN 2 Topology used for MVPN Source Redundancy**



The configuration change with respect to the previous section (P2MP RSVP-TE PMSIs) is an additional interface created in both CE-5 and PE-3 (int-CE-5-PE-3 on CE-5 and int-PE-3-CE-5 on PE-3), the addition of these interfaces to PIM and also the creation an e-BGP session between the two routers. The following is the additional configuration on PE-3 (CE-5 configuration changes are not displayed for brevity).

```
# on PE-3
configure
    service
        vprn 2
            interface "int-PE-3-CE-5" create
                address 172.16.35.1/30
                sap 1/1/4:2 create
                exit
            exit
            bgp
                group "EXTERNAL"
                    type external
                    peer-as 64505
                    neighbor 172.16.35.2
                    exit
                exit
                no shutdown
            exit
            pim
                interface "int-PE-3-CE-5"
                exit
            exit
        exit
```

Checking the routes on PE-4, the source is reachable through PE-1 and PE-2 as ECMP is set to 2. If the configuration of the VPRN is provisioned with **auto-bind-tunnel resolution-filter rsvp resolution filter,** instead of static spoke-SDPs, the command **ignore-nh-metric** is also needed.

```
*A:PE-4# configure service vprn 2 ecmp 2
```

```
*A:PE-4# show router 2 route-table

===============================================================================
Route Table (Service: 2)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age         Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
--- snipped ---
192.168.52.0/24                               Remote  BGP VPN   00h01m42s   170
       192.0.2.1 (tunneled)                                         0
192.168.52.0/24                               Remote  BGP VPN   00h01m42s   170
       192.0.2.3 (tunneled)                                         0
-------------------------------------------------------------------------------
No. of Routes: 11
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-4#
```

When PE-4 receives a c-join/prune, PE-4 needs to find the **upstream multicast hop** (UMH) for the (S,G). This is the upstream multihop selection and is configurable. The values are highest-ip, hash-based, tunnel-status, and unicast-rt-pref

```
*A:PE-4# configure service vprn 2 mvpn umh-selection
  - no umh-selection
  - umh-selection {highest-ip|hash-based|tunnel-status|unicast-rt-pref}
```

The default is highest-ip, which is the selection of the highest /32 IP addresses (in this setup, PE-3 is preferred versus PE-1). A BGP c-join is sent with the route target equal to the VRF import extended community distributed by PE-3 for the subnet of the source (see following PE-4 debug).

```
17 2017/10/07 20:35:31.112 UTC MINOR: DEBUG #2001 vprn2 PIM[vprn 3 vprn2]
"PIM[vprn 3 vprn2]: pimSGUpStateJMribChange
SG (192.168.52.2,232.2.2.2), type <S,G> newMribNhopIp 192.0.2.3
newRpfNbrIp 192.0.2.3 newRpfType REMOTE newRpfIf 73733"

--- snipped ---

20 2017/10/07 20:35:31.112 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 76
    Flag: 0x90 Type: 14 Len: 33 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.4
        Type: Source-Join Len:22 RD: 64496:203 SrcAS: 64496
                                 Src: 192.168.52.2 Grp: 232.2.2.2
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:192.0.2.3:3
"
```

The second option is hash-based, where the UMH is selected (both PEs are potentially possible UMHs) after hashing the source and group addresses of the stream. For this example, PE-3 is also preferred.

The third option, tunnel-status, is based on the status of the P2MP RSVP tunnel (not available in mLDP or PIM). The roots PE-1 and PE-3 are sending BFD messages to the leaf PE-4 (in fact this is UFD, unidirectional forwarding detection). The c-join from PE-4 for the (S,G) is sent to both PE-1 and PE-3, and in return the traffic is forwarded from both PE-1 and PE-3 for the c-group onto the I-PMSI; therefore PE-4 receives two copies of the c-(S,G) stream. By configuration, the stream from the primary PE-1 is selected by PE-4 to be forwarded to receiver H-8. If BFD messages are no longer received over the primary P2MP LSP, then the stream from the standby PE-3 is selected and forwarded to the receiver.

The configuration on PE-1 and PE-3 is similar and is as follows (only PE-3 is shown):

```
# on PE-3
configure
    service
        vprn 2
```

```
            mvpn
                auto-discovery default
                c-mcast-signaling bgp
                umh-selection tunnel-status
                provider-tunnel
                    inclusive
                        rsvp
                            lsp-template "VRF2"
                            enable-bfd-root 100
                            no shutdown
                        exit
                    exit
                exit
                vrf-target unicast
                exit
            exit
```

PE-1 and PE-3 are root. On PE-4, BFD is configured as leaf and the primary PE (PE-1) and backup PE (PE-3) are also provisioned:

```
# on PE-4
configure
    service
        vprn 2
            mvpn
                auto-discovery default
                c-mcast-signaling bgp
                umh-selection tunnel-status
                umh-pe-backup
                    umh-pe 192.0.2.1 standby 192.0.2.3
                exit
                provider-tunnel
                    inclusive
                        rsvp
                            lsp-template "VRF2"
                            enable-bfd-leaf
                            no shutdown
                        exit
                    exit
                exit
                vrf-target unicast
                exit
            exit
```

This BFD (UFD) configuration on the root establishes a session with the leaf. The root only transmits BFD packets; it doesn’t receive any.

```
*A:PE-1# show router 2 bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
```

```
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
mpls-if-73740                                        Up          330          0
  127.0.0.0                                           3          100          0
  pim                                           central          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:PE-1#
```

On PE-4, two BFD sessions are received, one from each root (note that BFD packets are only received):

```
*A:PE-4# show router 2 bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
mpls-if-73743                                        Up            0        189
  192.0.2.3                                           3         1000        100
  pim                                           central          N/A        N/A
mpls-if-73744                                        Up            0        188
  192.0.2.1                                           3         1000        100
  pim                                           central          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 2
===============================================================================
*A:PE-4#
```

PE-4 delivers the multicast traffic from the primary configured UMH, PE-1. If, as an example of a failure condition, PE-1 goes down (reboot), PE-4 will switch to the PE-3 P2MP LSP.

MDT AFI SAFI for Rosen MVPN

In Rosen MVPN up to version 6, the default MDT is PIM sparse mode only, and there is no auto-discovery mechanism available. In SR-OS Release 7.0, and later, it is possible to configure PIM SSM with auto-discovery, using AFI 1 and SAFI 5. Rosen MVPN version 7 allows use of MDT PIM SM or SSM, and auto-discovery based on AFI 1 and SAFI 66 to distribute the default MDT. Rosen MVPN version 9 adds a new MDT NLRI. SR OS has added the capability and support of MDT SAFI as specified in RFC 6037.

MDT SAFI is used to discover PEs in a specific MVPN, so that PIM SSM can be used for default MDT. The basic idea is the same as MVPN BGP auto-discovery, but it uses a different BGP SAFI. BGP messages in which AFI=1 and SAFI=66 are "MDT-SAFI" messages. The NLRI format is 8-byte-RD:IPv4-address followed by the MDT group address. The IPv4 address identifies the PE that originated this route and the RD identifies a VRF in that PE. The group address must be an IPv4 multicast group address and is used to build the P-tunnels.

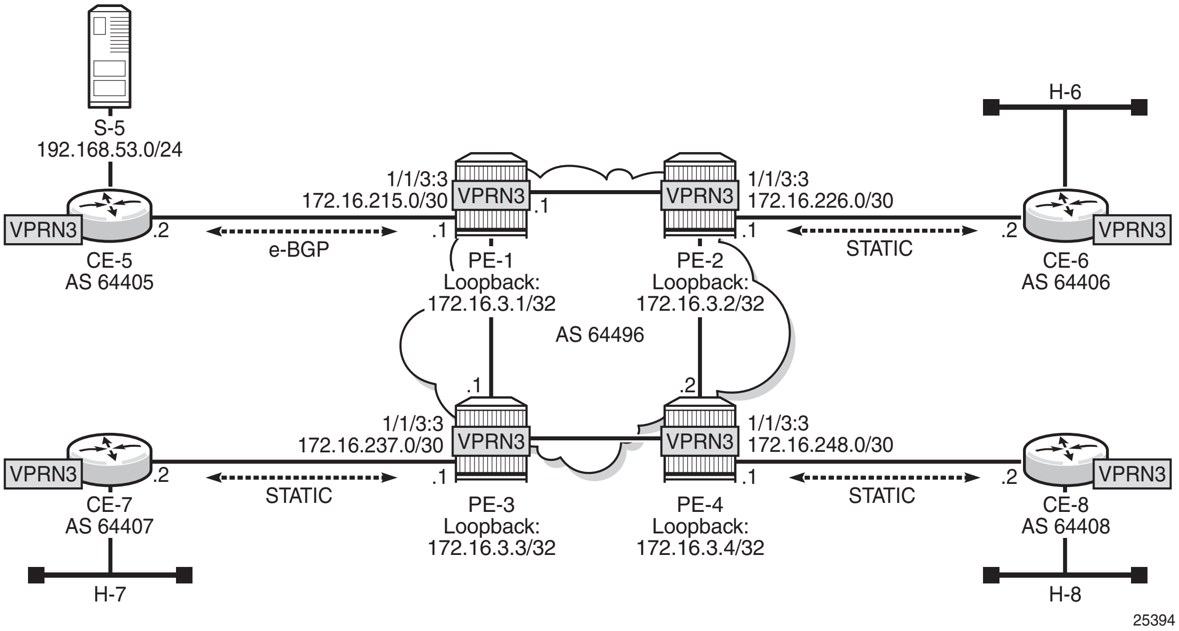
All PEs attached to a given MVPN must specify the same group-address. MDT-SAFI routes do not carry RTs and the group address is used to associate a received MDT-SAFI route with a VRF.

MDT SAFI can only be used when the implicit provider tunnel is PIM GRE based with a specific IPv4 group address.

For additional information on the use of PIM PMSIs, see NG-MVPN Configuration with PIM.

Figure 167 shows the topology of VPRN 3.

Figure 167 **VPRN 3 Topology used for MDT SAFI**



In this scenario, there is no MPLS based PMSI, there is PIM in the core for the control plane and the data traffic is GRE encapsulated. PIM needs to be configured in the base router on interface system and on the other interfaces pointing to other PEs. PIM is used for c-signaling. In addition, auto-discovery is provisioned to use mdt-safi and a PIM SSM inclusive PMSI with address 239.1.1.1 as the default MDT. The configuration is as follows on PE-4:

```
configure
    router
        pim
            interface "system"
            exit
            interface "int-PE-4-PE-2"
            exit
            interface "int-PE-4-PE-3"
            exit
        exit
    exit
    service
        vprn 3 customer 1 create
            description "PIM SSM / MDT SAFI"
            autonomous-system 64496
            route-distinguisher 64496:304
            vrf-target target:64496:300
            interface "loopback" create
                address 172.16.3.4/32
                loopback
            exit
            interface "int-PE-4-CE-8" create
                address 172.16.248.1/30
                sap 1/1/3:3 create
                exit
            exit
            pim
                interface "loopback"
                exit
                interface "int-PE-4-CE-8"
                exit
            exit
            mvpn
                auto-discovery mdt-safi
                provider-tunnel
                    inclusive
                        pim ssm 239.1.1.1
                        exit
                    exit
                exit
                vrf-target unicast
                exit
            exit
            spoke-sdp 341 create
            exit
            spoke-sdp 342 create
            exit
            spoke-sdp 343 create
            exit
            no shutdown
```

The following debug output shows a BGP update with MDT AFI SAFI on PE-4:

```
11 2017/10/07 20:42:28.669 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 62
    Flag: 0x90 Type: 14 Len: 26 Multiprotocol Reachable NLRI:
        Address Family MDT-SAFI
        NextHop len 4 NextHop 192.0.2.4
        [MDT-SAFI] Addr 192.0.2.4, Group 239.1.1.1, RD 64496:304
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64496:300
"
```

The following output shows the MDT-SAFI routes that have been learned at PE-4:

```
*A:PE-4# show router bgp routes mdt-safi
===============================================================================
 BGP Router ID:192.0.2.4        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MDT-SAFI Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop                     Group-Addr                         Label
      As-Path
-------------------------------------------------------------------------------
u*>i  64496:301:192.0.2.1                                100         0
      192.0.2.1                   239.1.1.1                          -
      No As-Path
u*>i  64496:302:192.0.2.2                                100         0
      192.0.2.2                   239.1.1.1                          -
      No As-Path
u*>i  64496:303:192.0.2.3                                100         0
      192.0.2.3                   239.1.1.1                          -
      No As-Path
-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:PE-4#
```

Conclusion

This chapter provides information to configure multicast within a VPRN with next generation multicast VPN techniques. Specifically, the use of MPLS I-PMSIs (mLDP and P2MP RSVP-TE), MVPN source redundancy, and the complete set of features needed to interoperate with Rosen MVPN in live deployments are covered.

NG-MVPN Configuration with PIM

This chapter provides information about multicast in a VPRN service.

Topics in this chapter include:

- Applicability
- Summary
- Overview
- Configuration
- Conclusion

Applicability

Initially, this chapter was written for SR OS Release 7.0.R5. The configuration in the current edition is based on SR OS Release 15.0.R2. There are no prerequisites for this configuration.

Summary

Multicast VPN (MVPN) architectures describe a set of VRFs that support the transport of multicast traffic across a provider network.

RFC 6037 (herein referred to as Rosen MVPN) describes the use of Multicast Distribution Trees (MDTs) established between PEs within a VRF. Each VRF required its own tree. Customer edge routers form Protocol Independent Multicast (PIM) adjacencies with the PE, and PE-PE PIM adjacencies are formed across the multicast tree. PIM signaling and data streams are transported across the MDT. There were a number of limitations with the Rosen MVPN implementation including, but not limited to:

- Rosen MVPN requires a set of MDTs per VPN, which requires a PIM state per MDT. There is no option to aggregate MDT across multiple VPNs
- Customer signaling, PE discovery and Data MDT signaling are all PIM-based. There is no mechanism available to decouple these. Thus there is an incongruency between unicast and multicast VPNs using Rosen MVPN.

- There is no mechanism for using MPLS to encapsulate multicast traffic in the VPN. GRE is the only encapsulation method available in Rosen MVPN.
- Rosen MVPN multicast trees are signaled using PIM only. NG MVPN allows the use of mLDP, RSVP P2MP LSPs.
- PE to PE protocol exchanges for Rosen MVPN is achieved using PIM only. NG MVPN allows for the use of BGP signaling as per unicast Layer 3 VPNs.

Next Generation MVPN addresses these limitations by extending the idea of the per-VRF tree, by introducing the idea of Provider Multicast Service Interfaces (PMSI). These are equivalent to the default MDTs of Rosen MVPN in that they support control plane traffic (customer multicast signaling), and the data MDTs which carry multicast data traffic streams between PEs within a multicast VRF.

Next Generation MVPN allows the decoupling of the mechanism required to create a multicast VPN, such as PE auto-discovery (which PEs are members of which VPN), PMSI signaling (creation of tunnels between PEs) and customer multicast signaling (multicast signaling —IGMP/PIM — received from customer edge routers). Two types of PMSI exist:

- Inclusive (I-PMSI): contains all the PEs for an MVPN.
- Selective (S-PMSI): contains only a subset of PEs of an MVPN.

Knowledge of MPLS-VPN RFC 4364, *BGP/MPLS IP Virtual Private Networks (VPNs)*, architecture and functionality, as well as an understanding of multicast protocols, is assumed throughout.

This chapter provides configuration details required to implement the parts of Next Generation MVPN shown in Table 16.

Table 16 **Next Generation MVPN Components**

Provider Multicast Domain				**Customer Multicast Domain**		
I-PMSI	**Auto-discovery**	**C-Mcast**	**S-PMSI Creation**	**PE-based RP**	**Anycast RP on PE**	**PIM SSM**
PIM ASM	PIM	PIM join/leave	PIM SSM with S-PMSI join TLV	X	X	X
PIM ASM	BGP A/D	PIM join/leave	PIM SSM with S-PMSI join TLV			X

The first section of this chapter describes the common configuration required for each PE within the provider multicast domain regardless of the MVPN PE auto-discovery or customer signaling methods. This includes IGP and VPRN service configuration.

Following the common configuration, specific MVPN configuration required for the configuration for the provider multicast domain using PIM Any Source Multicast (ASM) with auto-discovery based on PIM or BGP auto-discovery (A/D), PIM used for the customer multicast signaling and PIM Source Specific Multicast (SSM) used for the S-PMSI creation are described. The customer domain configuration covers the following three cases:

1. PIM ASM with the Rendezvous Point (RP) in the provider PE
2. PIM ASM using anycast RP on the provider RPs
3. PIM SSM

Other possible options, not covered in this section but are described in the 7450 ESS/7750 SR/7950 XRS Multicast Routing Protocols Guide:

- The use of PIM SSM for the provider multicast I-PMSI.
- The use of BGP for the customer multicast signaling in the provider multicast domain.
- The provider S-PMSI creation through BGP S-PMSI A/D.
- The use of the customer RP based in the customer CE.

The use of mLDP and RSVP p2mp LSPs for the I/S-PMSI was not available in release 7.0.

The Multicast in a VPRN II example in NG-MVPN Configuration with MPLS introduces features that were not supported in Release 7.0.R5. It provides configuration details to implement:

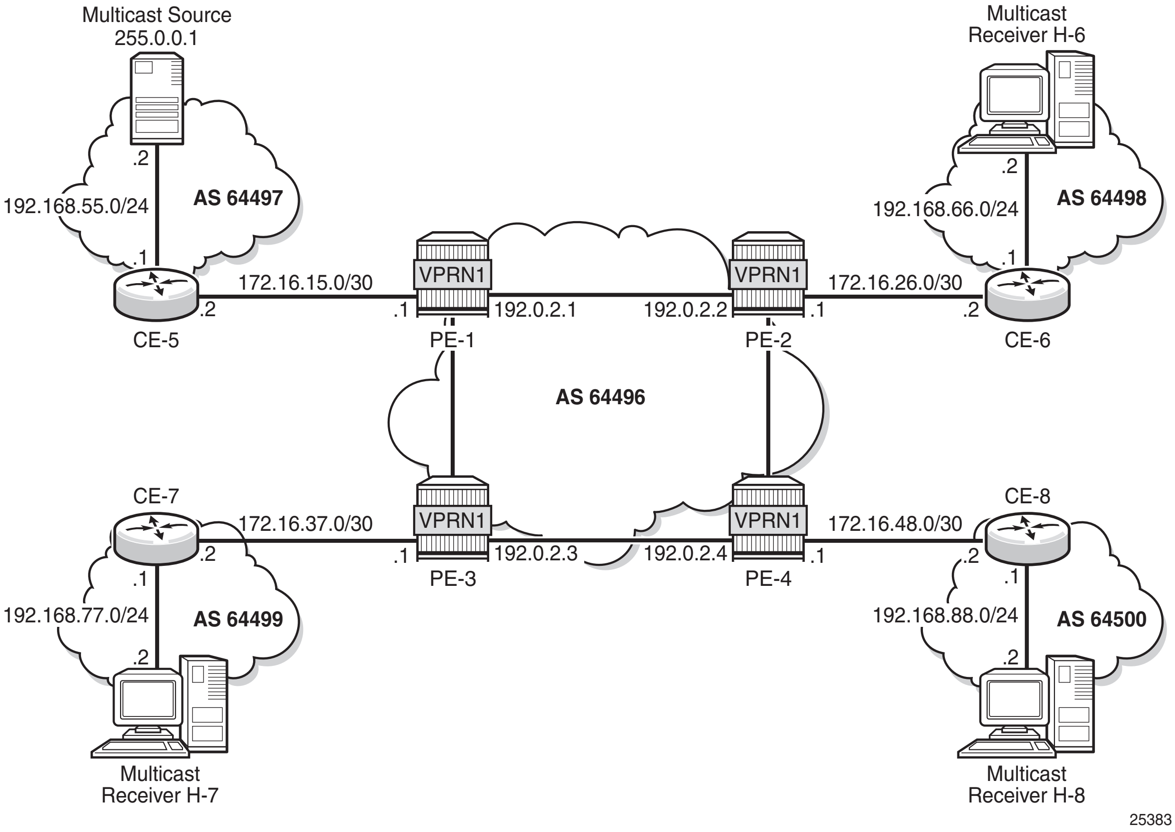
- Multicast LDP (mLDP) and RSVP-TE Point to Multi-point (P2MP) for building customer trees (C-trees) which are using MPLS instead of PIM techniques.
- MVPN source redundancy
- MDT AFI/SAFI (to fully interoperate with Cisco networks).

References

- IETF
  - RFC 6513, Multicast in MPLS/BGP IP VPNs.
  - RFC 6514, BGP Encodings and Procedures for Multicast in MPLS/ BGP IP VPNs.
- 7450 ESS/7750 SR/7950 XRS Layer 3 Services Guide

Overview

Figure 168 **Network Topology**



The network topology is displayed in Figure 168. The setup consists of four SR 7750s acting as Provider Edge (PE) routers within a single Autonomous System (AS).

- Full mesh IS-IS or OSPF in each AS
- LDP on all interfaces in each AS (RSVP could also be used)
- MP-iBGP sessions between the PE routers in each AS (Route Reflectors (RRs) could also be used).
- Layer 3-VPN on all PEs with identical route targets, in the form AS-number: *vprn-service-id*

Connected to each PE is a single SR OS router acting as a Customer Edge (CE) router. CE-5 has a multicast source connected, and PE-6, PE-7, and PE-8 each have a single receiver connected which will receive the multicast streams from the source. In this document, each receiver is both IGMPv2 and IGMPv3 capable. If the customer domain multicast signaling plane uses Source Specific Multicasting (SSM), then an IGMPv3 receiver is configured; if Any Source Multicasting (ASM) is used, the receiver is IGMPv2 capable.

If the receiver is IGMPv3 capable, it will issue IGMPv3 reports that will include a list of required source addresses. The receiver will join the 232.0.0.1 multicast group.

If the receiver is only IGMPv2 capable, then it will issue IGMPv2 reports which do not specify a source of the group. In this case, a Rendezvous Point is required within the PIM control plane of the multicast VRF which is source-aware. In this case, the receiver will join the 225.0.0.1 multicast group.

When the receiver wishes to become a member of any group, the source address of the group must be known to the CE. As a result, the source address must be IP reachable by each CE, so it is advertised by CE-5 to the PEs with attachment circuits in VPRN1 using BGP.

Static routes are then configured on the receiver CEs to achieve IP reachability to the source address of multicast groups. In the case of PIM ASM, any RP that is configured must also be reachable from the CE.

Multicast VPN Overview

Multicast traffic from the source is streamed toward router CE-5. Receivers connected to PE-2, PE-3 and PE-4 are interested in joining this multicast group.

All CEs are PIM enabled routers, which form a PIM adjacency with their nearest PE. The PIM adjacencies between PEs across the Provider network are achieved using I-PMSIs. I-PMSIs carry PIM control messages between PEs. Data plane traffic is transported across the I-PMSI until a configured bandwidth threshold is reached. A Selective PMSI is then signaled that carries data plane traffic. This threshold can be as low as 1kb/second and must be explicitly configured along with the S-PMSI multicast group. An S-PMSI per customer group per VPRN is configured. If no S-PMSI and threshold is configured, data traffic will continue to be forwarded across the provider network within the I-PMSI.

Configuration

The configuration is divided into the following sections:

- Provider Common Configuration
  - PE Global Configuration
  - PE VPRN Configuration
- PE VPRN Multicast Configuration
  - Auto-Discovery within Provider Domain using PIM
  - PIM Autodiscovery: Customer Signaling using PIM
- PIM Any Source Multicasting with RP at the provider PE
- PIM Any Source Multicasting with Anycast RP at the provider PE
- PIM Source Specific Multicasting
  - BGP Autodiscovery: PE VPRN Multicast Configuration
  - Data Path Using Selective-PMSIs

Provider Common Configuration

This section describes the common configuration required for each PE within the Provider multicast domain, regardless of the MVPN PE auto-discovery or customer signaling methods. This includes IGP and VPRN service configuration.

The configuration tasks can be summarized as follows:

- PE global configuration. This includes configuration of the Interior Gateway Protocol (IGP) (IS-IS or OSPF); configuration of link layer LDP between PEs; configuration of iBGP between PEs, to facilitate VPRN route learning; configuration of PIM.
- VPRN configuration on PEs. This includes configuration of basic VPRN parameters (route-distinguisher, route target communities); configuration of attachment circuits toward CEs; configuration of VRF routing protocol and any policies toward CE.
- VRF PIM and MVPN parameters — I-PMSI
- CE configuration.

PE Global Configuration

Step 1. On each of the PE routers, configure the appropriate router interfaces, OSPF (or IS-IS) and link layer LDP. For clarity in the following configuration steps, only the configuration for PE-1 is shown. PE-2, PE-3, and PE-4 are similar.

```
*A:PE-1# configure
    router
        interface "int-PE-1-PE-2"
            address 192.168.12.1/30
            port 1/1/1
        exit
        interface "int-PE-1-PE-3"
            address 192.168.13.1/30
            port 1/1/2
        exit
        interface "system"
            address 192.0.2.1/32
        exit
        autonomous-system 64496
        ospf
            area 0.0.0.0
                interface "system"
                exit
                interface "int-PE-1-PE-2"
                    interface-type point-to-point
                exit
                interface "int-PE-1-PE-3"
                    interface-type point-to-point
                exit
            exit
            no shutdown
        exit
        ldp
            interface-parameters
                interface "int-PE-1-PE-2" dual-stack
                    ipv4
                        no shutdown
                    exit
                exit
                interface "int-PE-1-PE-3" dual-stack
                    ipv4
                        no shutdown
                    exit
                exit
            exit
        exit
```

Step 2. Verify that OSPF adjacencies are formed and that LDP peer sessions are formed.

```
*A:PE-1# show router ospf neighbor

===============================================================================
Rtr Base OSPFv2 Instance 0 Neighbors
===============================================================================
Interface-Name                   Rtr Id          State      Pri  RetxQ   TTL
   Area-Id
-------------------------------------------------------------------------------
int-PE-1-PE-2                    192.0.2.2       Full       1    1       36
   0.0.0.0
int-PE-1-PE-3                    192.0.2.3       Full       1    1       34
   0.0.0.0
-------------------------------------------------------------------------------
No. of Neighbors: 2
===============================================================================
*A:PE-1#
```

```
*A:PE-1# show router ldp session ipv4

==============================================================================
LDP IPv4 Sessions
==============================================================================
Peer LDP Id         Adj Type  State         Msg Sent  Msg Recv  Up Time
------------------------------------------------------------------------------
192.0.2.2:0         Link      Established   176       178       0d 00:07:36
192.0.2.3:0         Link      Established   175       176       0d 00:07:30
------------------------------------------------------------------------------
No. of IPv4 Sessions: 2
==============================================================================
*A:PE-1#
```

Step 3. Configure BGP between the PEs for VPRN routing.

```
*A:PE-1# configure
    router
        bgp
            group "INTERNAL"
                family vpn-ipv4
                type internal
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
            exit
            no shutdown
        exit
```

Step 4. Verify that BGP sessions are established for address family VPN-IPv4.

```
*A:PE-1# show router bgp summary all

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
```

```
192.0.2.2
Def. Instance  64496        4    0 00h00m30s 0/0/0 (VpnIPv4)
                            4    0
192.0.2.3
Def. Instance  64496        3    0 00h00m25s 0/0/0 (VpnIPv4)
                            3    0
192.0.2.4
Def. Instance  64496        3    0 00h00m18s 0/0/0 (VpnIPv4)
                            3    0

-------------------------------------------------------------------------------
*A:PE-1#
```

Step 5. Enable PIM on all network interfaces, including the system interface. This allows the signaling of PMSIs that transport PIM signaling within each VRF.

Step 6. Each I-PMSI will be signaled using PIM ASM, so a rendezvous point (RP) is required within the global PIM configuration. A static RP is used and PE-1 is selected. All PEs must be configured with this RP address.

```
*A:PE-1# configure
    router
        pim
            interface "system"
            exit
            interface "int-PE-1-PE-2"
            exit
            interface "int-PE-1-PE-3"
            exit
            rp
                static
                    address 192.0.2.1
                        group-prefix 239.255.0.0/16
                    exit
                exit
            exit
```

Step 7. The following command shows the PIM neighbor relationships.

```
*A:PE-1# show router pim neighbor

===============================================================================
PIM Neighbor ipv4
===============================================================================
Interface              Nbr DR Prty     Up Time       Expiry Time     Hold Time
   Nbr Address
-------------------------------------------------------------------------------
int-PE-1-PE-2          1               0d 00:00:32   0d 00:01:44     105
   192.168.12.2
int-PE-1-PE-3          1               0d 00:00:25   0d 00:01:21     105
   192.168.13.2
-------------------------------------------------------------------------------
Neighbors : 2
===============================================================================
*A:PE-1#
```

PE VPRN Configuration

A VPRN (VPRN 1) is created on each PE. This will be the multicast VPRN. PE-1 is the PE containing the attachment circuit toward CE-5. CE-5 is the CE nearest the source. PE-2, PE-3, and PE-4 contain attachment circuits toward CE-6, CE-7, and CE-8 respectively. CE-6 has receiving host H-6 attached; CE-7 has receiving host H-7, and CE-8 receiving host H-8.

Step 1. Create VPRN 1 on each PE, containing a route-distinguisher and vrf-target of 64496:1. The autonomous system number is 64496. Use **auto-bind-tunnel resolution-filter ldp** for next hop tunnel route resolution.

```
*A:PE-1# configure
    service
        vprn 1 customer 1 create
            autonomous-system 64496
            route-distinguisher 64496:1
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64496:1
```

Step 2. Create an attachment circuit interface on PE-1 toward CE-5.

```
*A:PE-1# configure
    service
        vprn 1
            interface "int-PE-1-CE-5" create
                address 172.16.15.1/30
                sap 1/1/3 create
                exit
            exit
```

Step 3. The source address of the multicast stream will need to be reachable by all routers (PEs and CEs) within the VPN. This will be advertised within BGP from the CE to the PE. Create a BGP peering relationship within VPRN 1 on PE-1 with CE-5.

```
*A:PE-1# configure
    service
        vprn 1
            bgp
                group "EXTERNAL"
                    type external
                    peer-as 64497
                    neighbor 172.16.15.2
                    exit
                exit
                no shutdown
            exit
            no shutdown
```

Step 4. On CE-5, create a VPRN to support the connection of the source to the CE and to connect the CE to the PE. Two attachment circuits are required, as well as a BGP peering relationship with the PE. This uses a default address family of **ipv4**.

(A pair of IES services could also be used to provide the attachment circuits.)

```
*A:CE-5# configure
    service
        vprn 1 customer 1 create
            autonomous-system 64497
            route-distinguisher 64497:1
            interface "int-CE-5-PE-1" create
                address 172.16.15.2/30
                sap 1/1/1 create
                exit
            exit
            interface "int-CE-5-S-5" create
                address 192.168.55.1/24
                sap 1/1/3 create
                exit
            exit
            bgp
                group "EXTERNAL"
                    type external
                    peer-as 64496
                    neighbor 172.16.15.1
                    exit
                exit
                no shutdown
            exit
            no shutdown
```

Step 5. The following BGP summaries shows that the PE-CE BGP peer relationship between CE-5 and PE-1 is established for address family IPv4:

```
*A:CE-5# show router 1 bgp summary all

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
172.16.15.1
Svc: 1         64496        3    0 00h00m06s 0/0/0 (IPv4)
                            3    0
-------------------------------------------------------------------------------
*A:CE-5#


*A:PE-1# show router 1 bgp summary

===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
BGP Admin State         : Up          BGP Oper State              : Up
Total Peer Groups       : 1           Total Peers                 : 1
Total BGP Paths         : 5           Total Path Memory           : 944
Total IPv4 Remote Rts   : 0           Total IPv4 Rem. Active Rts  : 0
---snip---

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
                   AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
172.16.15.2
                64497       4    0 00h00m53s 0/0/0 (IPv4)
                            4    0
-------------------------------------------------------------------------------
*A:PE-1#
```

Step 6. In order for the CE connecting to the source to be advertised within BGP, a route policy is required. The subnet containing the multicast source is 192.168.55.0/24, so a prefix-list can be used to define a match, and then used within a route policy to inject into BGP.

```
*A:CE-5# configure
    router
        policy-options
            begin
            prefix-list "SOURCE-PREFIX"
                prefix 192.168.55.0/24 exact
            exit
            policy-statement "EXPORT-SOURCE-PREFIX-TO-BGP"
                entry 10
                    from
                        prefix-list "SOURCE-PREFIX"
                    exit
                    to
                        protocol bgp
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

Step 7. Apply this policy as an export policy within the BGP context.

```
*A:CE-5# configure
    service
        vprn 1
            bgp
                export "EXPORT-SOURCE-PREFIX-TO-BGP"
            exit
```

This results in the 192.168.55.0/24 subnet being seen in the BGP RIB_OUT on CE-5.

```
*A:CE-5# show router 1 bgp routes 192.168.55.0/24 hunt
===============================================================================
 BGP Router ID:192.0.2.5        AS:64497       Local AS:64497
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
---snip---
-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
Network        : 192.168.55.0/24
Nexthop        : 172.16.15.2
Path Id        : None
To             : 172.16.15.1
Res. Nexthop   : n/a
Local Pref.    : n/a                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : No Community Members
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.1
Origin         : IGP
AS-Path        : 64497
Route Tag      : 0
Neighbor-AS    : 64497
Orig Validation: NotFound
Source Class   : 0                      Dest Class     : 0

-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:CE-5#
```

It is also seen in the PE-1 VRF 1 FIB:

```
*A:PE-1# show router 1 route-table

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                   Metric
-------------------------------------------------------------------------------
172.16.15.0/30                                Local   Local     00h05m11s  0
       int-PE-1-CE-5                                               0
```

```
172.16.26.0/30                                Remote  BGP VPN   00h04m57s  170
       192.0.2.2 (tunneled)                                         0
172.16.37.0/30                                Remote  BGP VPN   00h04m30s  170
       192.0.2.3 (tunneled)                                         0
172.16.48.0/30                                Remote  BGP VPN   00h04m22s  170
       192.0.2.4 (tunneled)                                         0
192.168.55.0/24                               Remote  BGP       00h02m08s  170
       172.16.15.2                                                  0
-------------------------------------------------------------------------------
No. of Routes: 5
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

This prefix will also be automatically advertised within the BGP VPRN to all other PEs, and will be installed in VRF 1.

For example, on PE-2:

```
*A:PE-2# show router 1 route-table

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.15.0/30                                Remote  BGP VPN   00h05m05s  170
       192.0.2.1 (tunneled)                                         0
172.16.26.0/30                                Local   Local     00h05m08s  0
       int-PE-2-CE-6                                                0
172.16.37.0/30                                Remote  BGP VPN   00h04m56s  170
       192.0.2.3 (tunneled)                                         0
172.16.48.0/30                                Remote  BGP VPN   00h04m51s  170
       192.0.2.4 (tunneled)                                         0
192.168.55.0/24                               Remote  BGP VPN   00h02m07s  170
       192.0.2.1 (tunneled)                                         0
-------------------------------------------------------------------------------
No. of Routes: 5
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-2#
```

Each CE containing the multicast receivers must be able to reach the source. The following output shows the VPRN configuration of CE-6 containing an interface toward PE-2 and an interface toward receiving host H-6. A static route will suffice and is configured with next hop of the PE-2 PE-CE interface.

```
*A:CE-6# configure
    service
```

```
vprn 1 customer 1 create
    route-distinguisher 64498:1
    interface "int-CE-6-H-6" create
        address 192.168.66.1/24
        sap 1/1/2 create
        exit
    exit
    interface "int-CE-6-PE-2" create
        address 172.16.26.2/30
        sap 1/1/1 create
        exit
    exit
    static-route-entry 192.168.55.0/24
        next-hop 172.16.26.1
            no shutdown
        exit
    exit
    no shutdown
```

PE VPRN Multicast Configuration

This section gives details of the VPRN configuration that allows the support of multicasting.

Sub-sections include:

1. Auto-discovery — This is the mechanism by which each PE advertises the presence of an MVPN to other PEs. This can be achieved using PIM or using BGP. This section covers PIM auto-discovery (auto-discovery using BGP is shown later).
2. Customer domain signaling — This discusses the mechanism of transporting customer signaling.
3. Data plane connectivity — This is the signaling of S-PMSIs within the provider domain to carry each individual customer multicast stream.

This chapter describes the PIM and BGP auto-discovery mechanisms in detail. For each of these, there is an example of customer domain signaling. For completion, a single example of S-PMSI creation is also shown.

Auto-Discovery within Provider Domain Using PIM

Each PE advertises its membership of a multicast VPN using PIM through the configuration of an Inclusive PMSI (I-PMSI). This is a multicast group that is common to each VPRN. The configuration for PE 1 and 2 is as follows:

```
*A:PE-1# configure
```

```
    service
        vprn 1
            mvpn
                provider-tunnel
                    inclusive
                        pim asm 239.255.255.1
                        exit
                    exit
                exit
            exit

*A:PE-2# configure
    service
        vprn 1
            mvpn
                provider-tunnel
                    inclusive
                        pim asm 239.255.255.1
                        exit
                    exit
                exit
            exit
```

The multicast group address used for the PMSI must be the same on all PEs for this VPRN instance.

Verify that PIM in the Global Routing Table (GRT) has signaled the I-PMSIs.

For the PE acting as the RP for global PIM:

```
*A:PE-1# show router pim group

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
239.255.255.1               (*,G)                                       3
   *                           192.0.2.1
239.255.255.1               (S,G)               spt      system         3
   192.0.2.1                   192.0.2.1
239.255.255.1               (S,G)               spt      int-PE-1-PE-2  3
   192.0.2.2                   192.0.2.1
239.255.255.1               (S,G)               spt      int-PE-1-PE-3  3
   192.0.2.3                   192.0.2.1
239.255.255.1               (S,G)               spt      int-PE-1-PE-2  2
   192.0.2.4                   192.0.2.1
-------------------------------------------------------------------------------
Groups : 5
===============================================================================
*A:PE-1#
```

This shows an incoming (S,G) join from all other PEs within the multicast VRF, plus an outgoing (*,G) join to the same PEs.

PE-3 will have the following PIM groups:

```
*A:PE-3# show router pim group

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
239.255.255.1               (*,G)                        int-PE-3-PE-1  1
   *                           192.0.2.1
239.255.255.1               (S,G)               spt      system         2
   192.0.2.3                   192.0.2.1
-------------------------------------------------------------------------------
Groups : 2
===============================================================================
*A:PE-3#
```

This shows an (S,G) join toward the RP at 192.0.2.1, plus a (*,G) join from the RP. These represent the outgoing and incoming PIM interfaces for the VRF.

This results in a series of PIM neighbors through the I-PMSIs within the VRF, which are maintained using PIM hellos.

```
*A:PE-1# show router 1 pim neighbor

===============================================================================
PIM Neighbor ipv4
===============================================================================
Interface               Nbr DR Prty     Up Time       Expiry Time     Hold Time
   Nbr Address
-------------------------------------------------------------------------------
int-PE-1-CE-5           1               0d 00:01:03   0d 00:01:43     105
   172.16.15.2
1-mt-239.255.255.1      1               0d 00:01:24   0d 00:01:25     105
   192.0.2.2
1-mt-239.255.255.1      1               0d 00:01:18   0d 00:01:31     105
   192.0.2.3
1-mt-239.255.255.1      1               0d 00:01:11   0d 00:01:37     105
   192.0.2.4
-------------------------------------------------------------------------------
Neighbors : 4
===============================================================================
*A:PE-1#
```

PIM Auto-Discovery: Customer Signaling using PIM

Consider now how the signaling plane of the customer domain is dealt with at the provider domain.

The customer domain configuration covers the following three cases:

1. PIM ASM with the RP in the provider PE.
2. PIM ASM using anycast RP on the provider RPs.
3. PIM SSM.

PIM Any Source Multicasting with RP at the Provider PE

Each PE connects to a CE which will be part of the multicast VRF, so it is necessary to enable PIM on each interface containing an attachment circuit toward a CE, and to configure the I-PMSI multicast tunnel for the VRF.

There is a requirement for an RP, because customer multicast signaling will be PIM-ASM.

The RP for the customer multicast will be on PE-2. In order to facilitate this, a loopback interface is created (called RP within the VPRN 1 context of PE-2, and will be advertised to all PEs. It must also be a PIM enabled interface.

The additional configuration for the RP on PE-2 is the following:

```
*A:PE-2# configure
    service
        vprn 1
            interface "RP" create
                address 10.2.3.5/32
                loopback
            exit
            pim
                interface "RP"
                exit
                rp
                    static
                        address 10.2.3.5
                            group-prefix 225.0.0.0/8
                        exit
                    exit
                exit
                no shutdown
            exit
```

The RP must also be configured on each of the PEs and CEs.

On PE-3, the PIM configuration in VPRN 1 is as follows:

```
*A:PE-3# configure
    service
        vprn 1
            pim
                interface "int-PE-3-CE-7"
                exit
                rp
                    static
                        address 10.2.3.5
                            group-prefix 225.0.0.0/8
                        exit
                    exit
                exit
            exit
```

The configuration on the other nodes is similar; only the interfaces are different.

Customer Edge Router Multicast Configuration

Each CE router will have a PIM neighbor peer relationship with its nearest PE.

The CE router (CE-5) containing the source will have PIM enabled on the interface connected to the source. It will also have a static RP entry, as the incoming sources need to be registered with the RP.

```
*A:CE-5# configure
    service
        vprn 1
            pim
                interface "int-CE-5-PE-1"
                exit
                interface "int-CE-5-S-5"
                exit
                rp
                    static
                        address 10.2.3.5
                            group-prefix 225.0.0.0/8
                        exit
                    exit
                exit
            exit
```

The CE containing the receivers will have IGMP enabled on the interface connected to the receivers. Once again, there needs to be an RP configured, because the router needs to issue PIM joins to the RP. The additional configuration in VPRN 1 on CE-6 is as follows:

```
*A:CE-6# configure
    service
        vprn 1
```

```
            static-route-entry 10.0.0.0/8
                next-hop 172.16.26.1
                    no shutdown
                exit
            exit
            static-route-entry 192.168.55.0/24
                next-hop 172.16.26.1
                    no shutdown
                exit
            exit
            igmp
                interface "int-CE-6-H-6"
                exit
            exit
            pim
                interface "int-CE-6-PE-2"
                exit
                rp
                    static
                        address 10.2.3.5
                            group-prefix 225.0.0.0/8
                        exit
                    exit
                exit
            exit
```

Traffic Flow

The source sends a multicast stream using group address 225.0.0.1 toward CE-5. As the group matches the group address in the static RP configuration, the router sends a register join toward the RP. At this time, no receivers are interested in the group, so there are no entries in the Outgoing Interface List (OIL), and the number of outgoing interfaces (OIFs) is zero.

The PIM status of CE-5 within VPN 1 is as follows:

```
*A:CE-5# show router 1 pim group

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
225.0.0.1                   (S,G)                    int-CE-5-S-5   0
   192.168.55.2                10.2.3.5
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:CE-5#
```

The receiver H-6 connected to CE-6, wishes to join the group 225.0.0.1, and sends an IGMPv2 report toward CE-6. CE-6 recognizes the report, which contains no source.

```
*A:CE-6# show router 1 igmp group
===============================================================================
IGMP Interface Groups
===============================================================================

(*,225.0.0.1)                                             UpTime: 0d 00:00:05
    Fwd List  : int-CE-6-H-6
-------------------------------------------------------------------------------
Entries : 1
===============================================================================
IGMP Host Groups
===============================================================================
No Matching Entries
===============================================================================
IGMP SAP Groups
===============================================================================
No Matching Entries
===============================================================================
*A:CE-6#
```

CE-6 is not aware of the source of the group so initiates a (*,G) PIM join toward the RP.

At the RP, the following (*,G) join is received:

```
*A:PE-2# show router 1 pim group 225.0.0.1 type starg detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 225.0.0.1
Source Address     : *
RP Address         : 10.2.3.5
Advt Router        : 192.0.2.2
Flags              :                    Type               : (*,G)
Mode               : sparse
MRIB Next Hop      :
MRIB Src Flags     : self
Keepalive Timer    : Not Running
Up Time            : 0d 00:00:20        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:39
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Rpf Neighbor       :
Incoming Intf      :
Outgoing Intf List : int-PE-2-CE-6

Curr Fwding Rate   : 0.0 kbps
Forwarded Packets  : 0                  Discarded Packets  : 0
Forwarded Octets   : 0                  RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
```

```
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-2#
```

The RP can now forward traffic from itself toward CE-6, as the outgoing interface is seen as int-PE-2-CE-6.

CE-6 is now able to determine the source from the traffic stream, so it initiates a Reverse Path Forwarding (RPF) lookup of the source address in the route table, and issues an (S,G) PIM join toward the source.

The join is propagated across the provider network, from PE-2 toward PE-1 which is the resolved RPF next hop for the source.

```
*A:PE-1# show router 1 pim group detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 225.0.0.1
Source Address     : 192.168.55.2
RP Address         : 10.2.3.5
Advt Router        : 172.16.15.2
Flags              : spt                  Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 172.16.15.2
MRIB Src Flags     : remote
Keepalive Timer    : Not Running
Up Time            : 0d 00:01:15       Resolved By        : rtable-u

Up JP State        : Joined            Up JP Expiry       : 0d 00:00:44
Up JP Rpt          : Not Joined StarG  Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 172.16.15.2
Incoming Intf      : int-PE-1-CE-5
Outgoing Intf List : 1-mt-239.255.255.1

Curr Fwding Rate   : 7524.9 kbps
Forwarded Packets  : 1531063           Discarded Packets  : 0
Forwarded Octets   : 70428898          RPF Mismatches     : 0
Spt threshold      : 0 kbps            ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-1#
```

The outgoing interface is the I-PMSI: 1-mt-239.255.255.1.

The join is received by CE-5, which contains the subnet of the source.

CE-5 now recognizes the multicast group as a valid stream. This becomes the root of the shortest path tree for the group.

```
*A:CE-5# show router 1 pim group

================================================================================
Legend:  A = Active   S = Standby
================================================================================
PIM Groups ipv4
================================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
--------------------------------------------------------------------------------
225.0.0.1                   (S,G)               spt      int-CE-5-S-5   1
   192.168.55.2                10.2.3.5
--------------------------------------------------------------------------------
Groups : 1
================================================================================
*A:CE-5#
```

For completion, consider a second receiver H-7 interested in group 225.0.0.1. The IGMPv2 report is translated into a (*,G) PIM join at CE-7 toward the RP.

```
*A:CE-7# show router 1 pim group type starg
================================================================================
Legend:  A = Active   S = Standby
================================================================================
PIM Groups ipv4
================================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
--------------------------------------------------------------------------------
225.0.0.1                   (*,G)                        int-CE-7-PE-3  1
   *                           10.2.3.5
--------------------------------------------------------------------------------
Groups : 1
================================================================================
*A:CE-7#
```

At the RP (PE-2), there is now a second interface in the OIL.

```
*A:PE-2# show router 1 pim group 225.0.0.1 type starg detail
================================================================================
PIM Source Group ipv4
================================================================================
Group Address      : 225.0.0.1
Source Address     : *
RP Address         : 10.2.3.5
Advt Router        : 192.0.2.2
Flags              :                    Type              : (*,G)
Mode               : sparse
MRIB Next Hop      :
MRIB Src Flags     : self
Keepalive Timer    : Not Running
Up Time            : 0d 00:02:05        Resolved By       : rtable-u
Up JP State        : Joined             Up JP Expiry       : 0d 00:00:55
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Rpf Neighbor       :
Incoming Intf      :
Outgoing Intf List : int-PE-2-CE-6, 1-mt-239.255.255.1

Curr Fwding Rate   : 0.0 kbps
Forwarded Packets  : 0                  Discarded Packets  : 0
Forwarded Octets   : 0                  RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-2#
```

The second interface is the I-PMSI, which is the multicast tunnel toward all other PEs. At PE-3, the (*,G) join has the I-PMSI as an incoming interface, and the PE-CE interface as the outgoing interface.

```
*A:PE-3# show router 1 pim group type starg detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 225.0.0.1
Source Address     : *
RP Address         : 10.2.3.5
Advt Router        : 192.0.2.2
Flags              :                    Type               : (*,G)
Mode               : sparse
MRIB Next Hop      : 192.0.2.2
MRIB Src Flags     : remote
Keepalive Timer    : Not Running
Up Time            : 0d 00:00:32        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:28
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Rpf Neighbor       : 192.0.2.2
Incoming Intf      : 1-mt-239.255.255.1
Outgoing Intf List : int-PE-3-CE-7

Curr Fwding Rate   : 0.0 kbps
Forwarded Packets  : 168                Discarded Packets  : 0
Forwarded Octets   : 7728               RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-3#
```

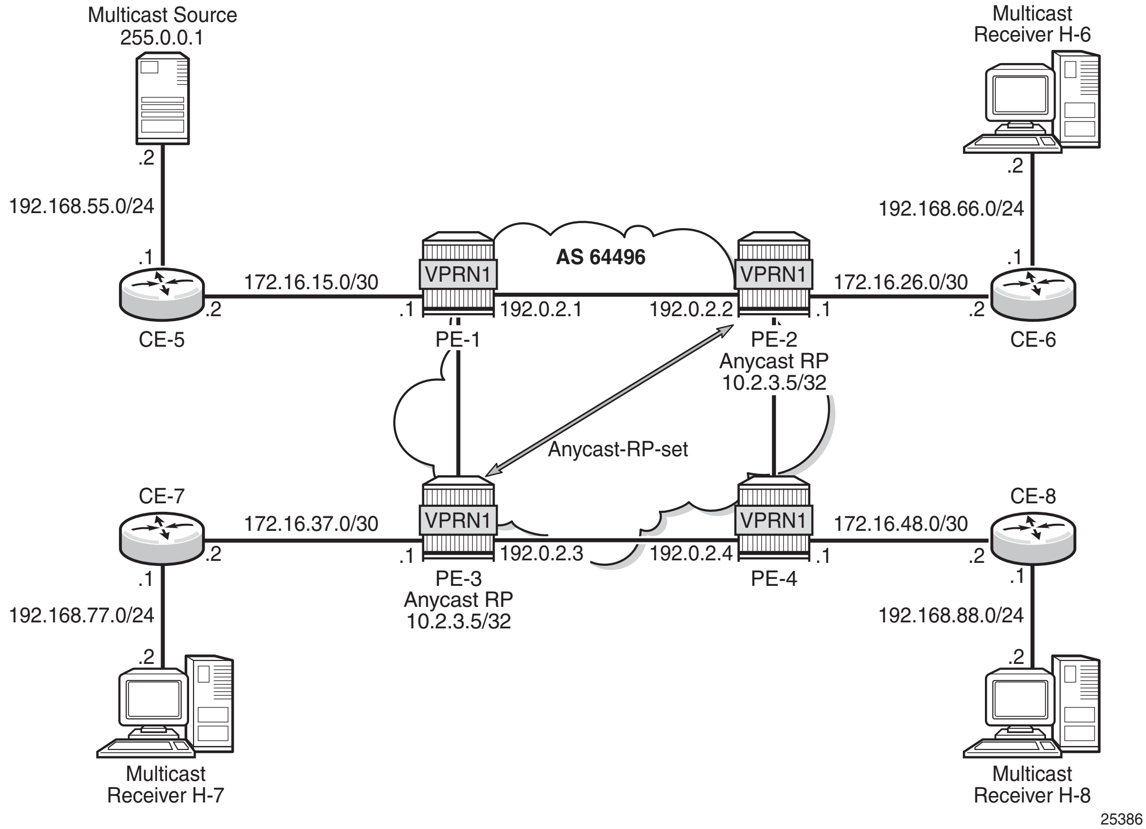
Once again, when the CE receives traffic from the group, it can use the source address in the packet to initiate an (S,G) join toward the source to join the Shortest Path Tree (SPT).

```
*A:CE-7# show router 1 pim group type sg detail
===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 225.0.0.1
Source Address     : 192.168.55.2
RP Address         : 10.2.3.5
Advt Router        :
Flags              : spt                 Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 172.16.37.1
MRIB Src Flags     : remote
Keepalive Timer Exp: 0d 00:02:40
Up Time            : 0d 00:00:52         Resolved By        : rtable-u

Up JP State        : Joined              Up JP Expiry       : 0d 00:00:08
Up JP Rpt          : Not Pruned          Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 172.16.37.1
Incoming Intf      : int-CE-7-PE-3
Outgoing Intf List : int-CE-7-H-7

Curr Fwding Rate   : 5481.9 kbps
Forwarded Packets  : 791128              Discarded Packets  : 0
Forwarded Octets   : 36391888            RPF Mismatches     : 0
Spt threshold      : 0 kbps              ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:CE-7#
```

PIM Any Source Multicasting with Anycast RP at the Provider PE

The example topology for anycast RP is shown in Figure 169. The setup consists of four SR OS routers acting as Provider Edge (PE) routers within a single Autonomous System (AS).

Figure 169 **Example Topology for Anycast RP**



Connected to each PE is a single SR OS router acting as a Customer Edge (CE) router. CE-5 has a single multicast source connected, and PEs 2 to 4 each have a single receiver connected which will receive the multicast stream from the source. In this section, each receiver is IGMPv2 capable, and will issue IGMPv2 reports. An RP is required by the C-signaling plane to resolve each (*,G) group state into an (S,G) state. In this case, two RPs are chosen to form an Anycast set to resolve each (*,G) group into an (S,G) state.

Multicast traffic from the source group 225.0.0.1 is streamed toward router CE-5. Receivers connected to PE-2, PE-3 and PE-4 are interested in joining this multicast group.

Anycast RP - PE VPRN Configuration

As previously stated, there is a requirement for an RP, as customer multicast signaling will be PIM-ASM and IGMPv2.

In this case, an anycast RP will be used. This is configured on PE-2 and PE-3, and an anycast set is created.

As each PE contains a CE which will be part of the multicast VRF, it is necessary to enable PIM on each interface containing an attachment circuit toward a CE, and to configure the I-PMSI multicast tunnel for the VRF.

The following output shows the VPRN configuration for PE-2 containing the RP and anycast RP configuration. The loopback interface lo1 is used for inter-RP communication:

```
*A:PE-2# configure
    service
        vprn 1
---snip---
            interface "RP" create
                address 10.2.3.5/32
                loopback
            exit
            interface "lo1" create
                address 10.0.0.2/32
                loopback
            exit
            pim
                interface "int-PE-2-CE-6"
                exit
                interface "RP"
                exit
                interface "lo1"
                exit
                rp
                    static
                        address 10.2.3.5
                            group-prefix 225.0.0.0/8
                        exit
                    exit
                    anycast 10.2.3.5
                        rp-set-peer 10.0.0.2
                        rp-set-peer 10.0.0.3
                    exit
                exit
                no shutdown
            exit
```

Similarly, the VPRN configuration for PE-3 is:

```
*A:PE-3# configure
    service
        vprn 1
---snip---
            interface "RP" create
                address 10.2.3.5/32
                loopback
            exit
            interface "lo1" create
                address 10.0.0.3/32
                loopback
            exit
            pim
                interface "int-PE-3-CE-7"
                exit
                interface "RP"
                exit
                interface "lo1"
                exit
                rp
                    static
                        address 10.2.3.5
                            group-prefix 225.0.0.0/8
                        exit
                    exit
                    anycast 10.2.3.5
                        rp-set-peer 10.0.0.2
                        rp-set-peer 10.0.0.3
                    exit
                exit
                no shutdown
            exit
```

As previously stated, there is a requirement for an RP, as customer multicast signaling will be PIM-ASM and IGMPv2.

In this case, an anycast RP will be used. This is configured on PE-2 and PE-3, and an anycast set is created.

The anycast address will be 10.2.3.5/32 and is created as an interface called **RP** on both PE-2 and PE-3.

An additional loopback interface, called “lo1” is created on each VPRN on PEs containing the anycast address. These are used as source addresses for communication between the routers within the RP set. These addresses will be automatically advertised to all PEs as VPN-IPv4 addresses, and will be installed in the VRF 1 forwarding table of all PEs containing VPRN 1.

Note: All routers containing RP must have their own loopback address included in the RP set as well as all peer routers.

The multicast group address used for the Inclusive PMSI is chosen to be 239.255.255.1 and must be the same on all PEs for this VPRN instance. This is analogous to the MDT within the Rosen MVPN implementation.

```
*A:PE-2# configure
    service
        vprn 1
            mvpn
                provider-tunnel
                    inclusive
                        pim asm 239.255.255.1
                        exit
                    exit
                exit
            exit
```

Verify that PIM in the global routing table (GRT) has signaled the I-PMSIs.

For the PE acting as the RP for global PIM:

```
*A:PE-1# show router pim group

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
239.255.255.1               (*,G)                                       3
   *                           192.0.2.1
239.255.255.1               (S,G)               spt      system         3
   192.0.2.1                   192.0.2.1
239.255.255.1               (S,G)               spt      int-PE-1-PE-2  3
   192.0.2.2                   192.0.2.1
239.255.255.1               (S,G)               spt      int-PE-1-PE-3  3
   192.0.2.3                   192.0.2.1
239.255.255.1               (S,G)               spt      int-PE-1-PE-2  2
   192.0.2.4                   192.0.2.1
-------------------------------------------------------------------------------
Groups : 5
===============================================================================
*A:PE-1#
```

PE-3 will have:

```
*A:PE-3# show router pim group

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
239.255.255.1               (*,G)                        int-PE-3-PE-1  1
   *                           192.0.2.1
239.255.255.1               (S,G)               spt      system         2
   192.0.2.3                   192.0.2.1
-------------------------------------------------------------------------------
Groups : 2
===============================================================================
*A:PE-3#
```

This shows a (S,G) join toward the RP at 192.0.2.1, plus a (*,G) join from RP. These represent the outgoing and incoming PIM interfaces for the VRF.

This results in a series of PIM neighbors through the I-PMSIs within the VRF, which are maintained using PIM hellos.

```
*A:PE-1# show router 1 pim neighbor

===============================================================================
PIM Neighbor ipv4
===============================================================================
Interface               Nbr DR Prty     Up Time       Expiry Time     Hold Time
   Nbr Address
-------------------------------------------------------------------------------
int-PE-1-CE-5           1               0d 00:08:01   0d 00:01:15     105
   172.16.15.2
1-mt-239.255.255.1      1               0d 00:08:22   0d 00:01:27     105
   192.0.2.2
1-mt-239.255.255.1      1               0d 00:08:15   0d 00:01:33     105
   192.0.2.3
1-mt-239.255.255.1      1               0d 00:08:09   0d 00:01:39     105
   192.0.2.4
-------------------------------------------------------------------------------
Neighbors : 4
===============================================================================
*A:PE-1#
```

Verify PIM RP set on PE-2 (similar for PE-3):

```
*A:PE-2# show router 1 pim anycast

===============================================================================
PIM Anycast RP Entries ipv4
===============================================================================
Anycast RP                              Anycast RP Peer
-------------------------------------------------------------------------------
10.2.3.5                                10.0.0.2
                                        10.0.0.3
-------------------------------------------------------------------------------
PIM Anycast RP Entries : 2
===============================================================================
*A:PE-2#
```

Anycast RP — Customer Edge Router Multicast Configuration

Each CE router will have a PIM neighbor peer relationship with its nearest PE.

The CE router (CE-5) containing the source will have PIM enabled on the interface connected to the source.

```
*A:CE-5# configure
    service
        vprn 1
            pim
                interface "int-CE-5-PE-1"
                exit
                interface "int-CE-5-S-5"
                exit
                rp
                    static
                        address 10.2.3.5
                            group-prefix 225.0.0.0/8
                        exit
                    exit
                exit
            exit
```

The CE containing the receivers will have IGMP enabled on the interface connected to the receivers.

```
*A:CE-6# configure
    service
        vprn 1
            igmp
                interface "int-CE-6-H-6"
                exit
            exit
```

Traffic Flow

Figure 170 **IGMP and PIM Control Messaging Schematic**

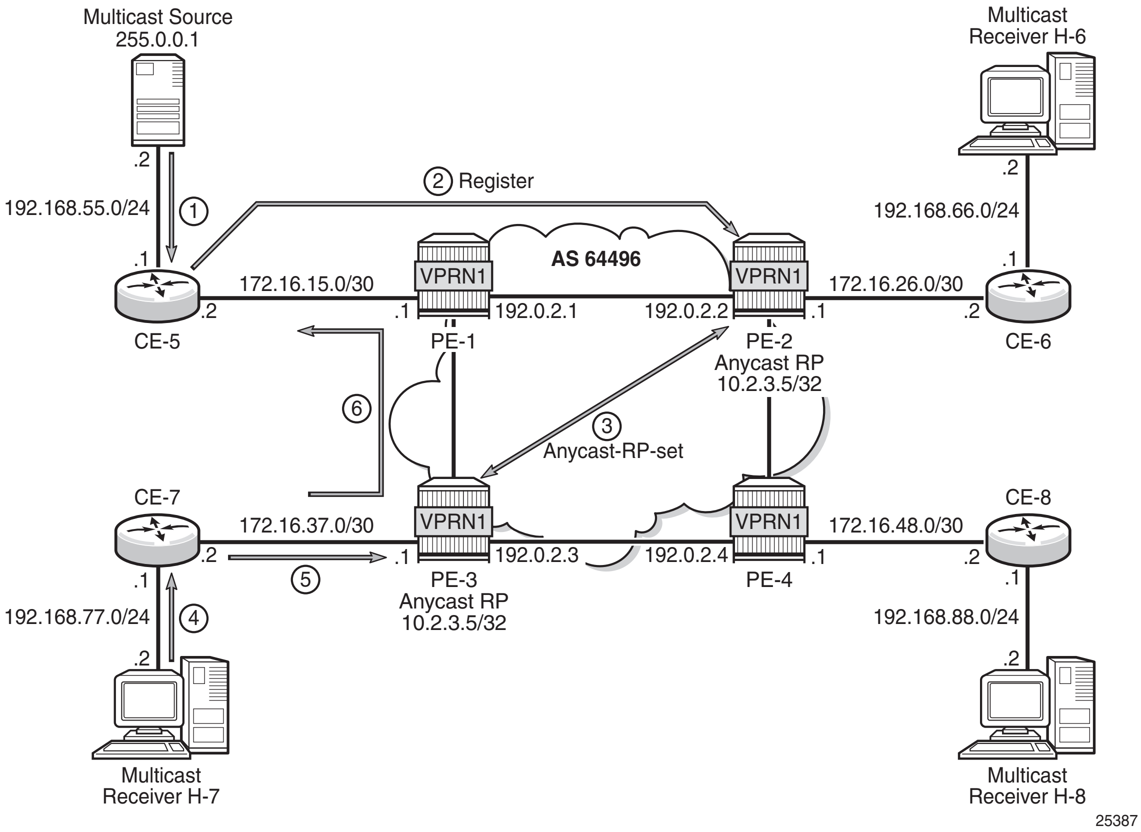


Figure 170 shows the sequence of IGMP and PIM control messaging.

1. The source multicasts a stream with group address 225.0.0.1 toward CE-5.
2. CE-5 matches the group with the group address prefix in the static RP configuration and sends a register message toward the RP.

```
*A:CE-5# show router 1 pim group detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 225.0.0.1
Source Address     : 192.168.55.2
RP Address         : 10.2.3.5
Advt Router        : 192.0.2.5
Flags              :                    Type              : (S,G)
Mode               : sparse
MRIB Next Hop      : 192.168.55.2
MRIB Src Flags     : direct
Keepalive Timer Exp: 0d 00:03:07
```

```
Up Time            : 0d 00:00:23        Resolved By        : rtable-u

Up JP State        : Not Joined         Up JP Expiry       : 0d 00:00:00
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : Pruned             Register Stop Exp  : 0d 00:00:32
Reg From Anycast RP: No

Rpf Neighbor       : 192.168.55.2
Incoming Intf      : int-CE-5-S-5
Outgoing Intf List :
Outgoing Sap List  :
Outgoing Host List :

Curr Fwding Rate   : 19716.7 kbps
Forwarded Packets  : 1221836            Discarded Packets  : 0
Forwarded Octets   : 56204456           RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:CE-5#
```

The register message is sent to the nearest RP, the RP with the lowest IGP cost.

When the register is sent through PE-1, it is PE-1 that determines which RP will receive the message.

```
*A:PE-1# show router 1 route-table 10.2.3.5/32

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.2.3.5/32                                   Remote  BGP VPN   00h00m59s  170
       192.0.2.2 (tunneled)                                         0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

The PE which will receive the register is 192.0.2.2 (PE-2). The PIM group status on PE-2 is:

```
*A:PE-2# show router 1 pim group

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
225.0.0.1                   (S,G)                        1-mt-239.255.* 0
   192.168.55.2                10.2.3.5
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-2#
```

This shows that RP is aware of the (S,G) status of the group 225.0.0.1, and becomes a root of a shared tree for this group. The Outgoing Interface List (OIL) is empty.

3. PE-2 will now send a register message to all other RPs within the anycast set, in this case to PE-3 (which has VPRN 1 containing address 10.0.0.3).

   The PIM status of the group 225.0.0.1 on PE-3 is:

```
*A:PE-3# show router 1 pim group

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
225.0.0.1                   (S,G)                        1-mt-239.255.* 0
   192.168.55.2                10.2.3.5
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-3#
```

Now both PEs within the RP set for VPRN have an (S,G) state for 225.0.0.1.

4. The receiver H-7, wishes to join the group 225.0.0.1, and sends in an IGMPv2 report toward CE-7. CE-7 recognizes the report, but has no PIM state for this group.
5. CE-7 sends a PIM join toward the RP, in this case the nearest RP will be PE-3.

   PE-3 already has (S,G) state for this group, so will forward traffic toward receiver H-7.
6. CE-7 does a Reverse Path Forwarding (RPF) lookup of the source address in the route table, and issues a PIM join toward the source.

   The join is propagated across the provider network toward PE-1, which is the resolved RPF next hop for the source.

```
*A:CE-7# show router 1 pim group type sg detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 225.0.0.1
Source Address     : 192.168.55.2
RP Address         : 10.2.3.5
Advt Router        :
Flags              : spt                Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 172.16.37.1
MRIB Src Flags     : remote
Keepalive Timer Exp: 0d 00:03:10
Up Time            : 0d 00:00:22        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:38
Up JP Rpt          : Not Pruned         Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 172.16.37.1
Incoming Intf      : int-CE-7-PE-3
Outgoing Intf List : int-CE-7-H-7

Curr Fwding Rate   : 6803.4 kbps
Forwarded Packets  : 406012             Discarded Packets  : 0
Forwarded Octets   : 18676552           RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:CE-7#
```

The join is received by CE-5, which contains the subnet of the source.

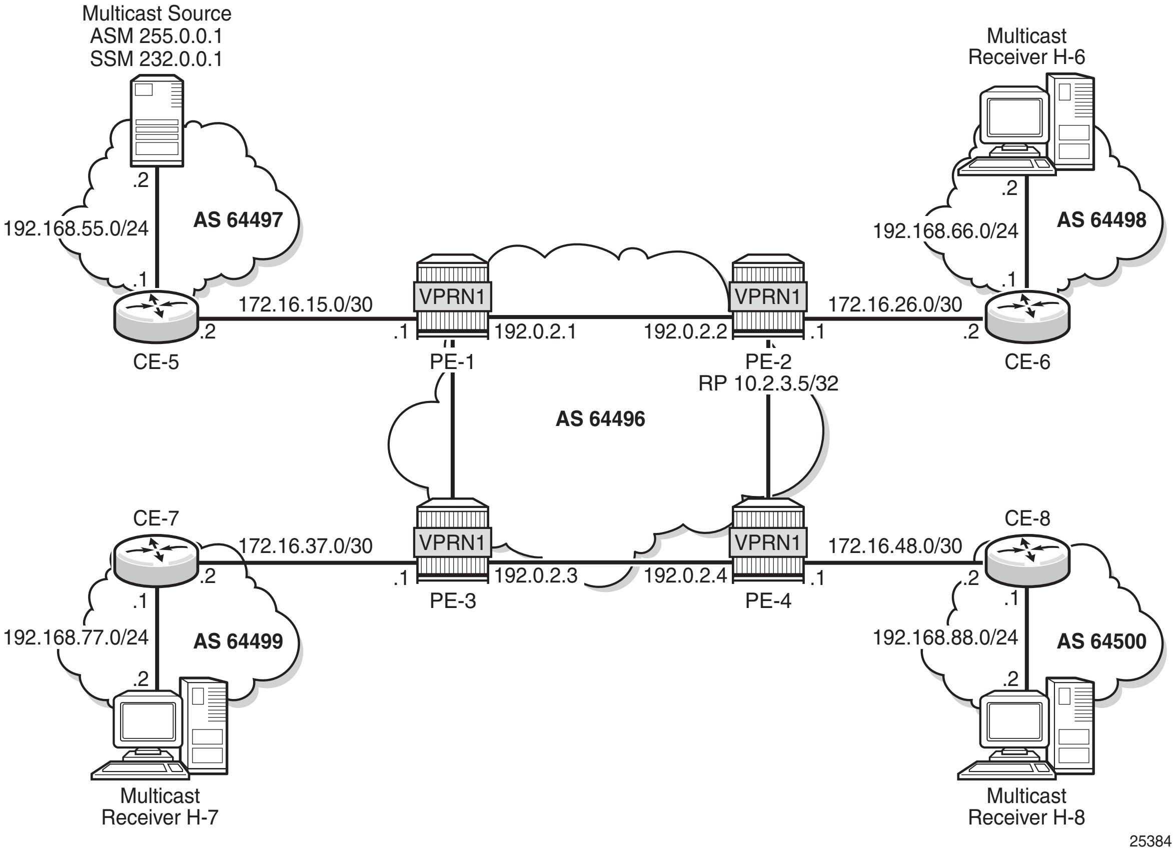
CE-5 now recognizes the multicast group as a valid stream. CE-5 becomes the root of the shortest path tree for the group.

```
*A:CE-5# show router 1 pim group

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
225.0.0.1                  (S,G)                spt      int-CE-5-S-5   1
   192.168.55.2                10.2.3.5
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:CE-5#
```

PIM Source-Specific Multicasting

There is no requirement for an RP, because customer multicast signaling will be PIM-SSM. The multicast group address used for the PMSI must be the same on all PEs for this VPRN instance.

Figure 171 **PIM SSM in Customer Signaling Plane**



The following output shows the VPRN configuration for PIM and MVPN for PE-1.

```
*A:PE-1# configure
    service
        vprn 1
            pim
                interface "int-PE-1-CE-5"
            exit
            mvpn
                provider-tunnel
                    inclusive
                        pim asm 239.255.255.1
                        exit
                    exit
                exit
            exit
```

There is a similar configuration required for each of the other PEs. Verify that PIM in the GRT has signaled the I-PMSIs.

For the PE acting as the RP for global PIM:

```
*A:PE-1# show router pim group

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
239.255.255.1               (*,G)                                       3
   *                           192.0.2.1
239.255.255.1               (S,G)               spt      system         3
   192.0.2.1                   192.0.2.1
239.255.255.1               (S,G)               spt      int-PE-1-PE-2  3
   192.0.2.2                   192.0.2.1
239.255.255.1               (S,G)               spt      int-PE-1-PE-3  3
   192.0.2.3                   192.0.2.1
239.255.255.1               (S,G)               spt      int-PE-1-PE-2  2
   192.0.2.4                   192.0.2.1
-------------------------------------------------------------------------------
Groups : 5
===============================================================================
*A:PE-1#
```

PE-3 will have:

```
*A:PE-3# show router pim group

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
239.255.255.1               (*,G)                        int-PE-3-PE-1  1
   *                           192.0.2.1
239.255.255.1               (S,G)               spt      system         2
   192.0.2.3                   192.0.2.1
-------------------------------------------------------------------------------
Groups : 2
===============================================================================
*A:PE-3#
```

This shows a (S,G) join toward the RP at 192.0.2.1, plus a (*,G) join from RP. These represent the outgoing and incoming PIM interfaces for the VRF.

This results in a series of PIM neighbors through the I-PMSIs within the VRF, which are maintained using PIM hellos.

```
*A:PE-1# show router 1 pim neighbor

===============================================================================
PIM Neighbor ipv4
===============================================================================
Interface               Nbr DR Prty     Up Time       Expiry Time     Hold Time
   Nbr Address
-------------------------------------------------------------------------------
int-PE-1-CE-5           1               0d 00:08:01   0d 00:01:15     105
   172.16.15.2
1-mt-239.255.255.1      1               0d 00:08:22   0d 00:01:27     105
   192.0.2.2
1-mt-239.255.255.1      1               0d 00:08:15   0d 00:01:33     105
   192.0.2.3
1-mt-239.255.255.1      1               0d 00:08:09   0d 00:01:39     105
   192.0.2.4
-------------------------------------------------------------------------------
Neighbors : 4
===============================================================================
*A:PE-1#
```

PIM SSM — Customer Edge Router Multicast Configuration

Each CE router will have a PIM neighbor peer relationship with its nearest PE.

The CE router (CE-5) containing the source will have PIM enabled on the interface connected to the source.

```
*A:CE-5# configure
    service
        vprn 1
            pim
                interface "int-CE-5-PE-1"
                exit
                interface "int-CE-5-S-5"
                exit
            exit
```

The CE containing the receivers will have IGMP enabled on the interface connected to the receivers and PIM on the interface facing the PE.

```
*A:CE-6# configure
    service
        vprn 1
            static-route-entry 192.168.55.0/24
                next-hop 172.16.26.1
                    no shutdown
                exit
            exit
            igmp
```

```
                interface "int-CE-6-H-6"
                exit
            exit
            pim
                interface "int-CE-6-PE-2"
                exit
            exit
```

Traffic Flow

The source multicasts a stream with group address 232.0.0.1 toward CE-5. When there is no receiver interested in the group at this time, there are no outgoing interfaces, so the Outgoing Interface List (OIL) is empty.

```
*A:CE-5# show router 1 pim group

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
232.0.0.1                   (S,G)                        int-CE-5-S-5   0
   192.168.55.2
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:CE-5#
```

The receiver H-6, wishes to join the group 232.0.0.1, and so sends in an IGMPv3 report toward CE-6. CE-6 recognizes the report, which contains the source 192.168.55.2 in the include filter list.

```
*A:CE-6# show router 1 igmp group
===============================================================================
IGMP Interface Groups
===============================================================================

(192.168.55.2,232.0.0.1)                                  UpTime: 0d 00:00:05
    Fwd List  : int-CE-6-H-6
-------------------------------------------------------------------------------
Entries : 1
===============================================================================
IGMP Host Groups
===============================================================================
No Matching Entries
===============================================================================
IGMP SAP Groups
===============================================================================
No Matching Entries
===============================================================================
*A:CE-6#
```

CE-6 does a RPF lookup of the source address in the route table, and issues a PIM join toward the source.

The join is propagated across the provider network, toward PE-1 which is the resolved RPF next hop for the source.

```
*A:PE-1# show router 1 pim group detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 232.0.0.1
Source Address     : 192.168.55.2
RP Address         : 0
Advt Router        : 172.16.15.2
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 172.16.15.2
MRIB Src Flags     : remote
Keepalive Timer    : Not Running
Up Time            : 0d 00:00:12        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:47
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 172.16.15.2
Incoming Intf      : int-PE-1-CE-5
Outgoing Intf List : 1-mt-239.255.255.1

Curr Fwding Rate   : 7854.4 kbps
Forwarded Packets  : 255250             Discarded Packets  : 0
Forwarded Octets   : 11741500           RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-1#
```

The outgoing interface is the I-PMSI: 1-mt-239.255.255.1.

The join is received by CE-5, which contains the subnet of the source.

CE-5 now recognizes the multicast group as a valid stream. CE-5 becomes the root of the shortest path tree for the group.

```
*A:CE-5# show router 1 pim group

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
232.0.0.1                   (S,G)                        int-CE-5-S-5   1
   192.168.55.2
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:CE-5#
```

PE BGP Auto-Discovery

Discovery of Multicast-enabled Virtual Private Networks (MVPNs) can also be achieved using BGP. To this end, any PE that is a member of a multicast VPN will advertise this using a BGP multi-protocol Network Layer Reachability Information (NLRI) update that is sent to all PEs within the AS. This update will contain an intra-AS I-PMSI Auto-Discovery route type, also known as an Intra-AD. These use a dedicated address family — **mvpn-ipv4** — so each PE must be configured to originate and accept such updates. The following needs to be modified in the BGP context for all PE nodes:

```
configure
    router
        bgp
            group "INTERNAL"
                family vpn-ipv4 mvpn-ipv4
                exit all
```

This is achieved in the GRT within the BGP context.

This allows each BGP speaker to advertise its capabilities within a BGP Open message.

The following BGP summary on PE-1 shows that BGP sessions are established between the PEs for address families VPN-IPv4 and MVPN-IPv4 in the base routing instance:

```
*A:PE-1# show router bgp summary all

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.2
```

```
Def. Instance  64496      226    0 00h05m04s 1/1/2 (VpnIPv4)
                           28    0           0/0/0 (MvpnIPv4)
192.0.2.3
Def. Instance  64496      226    0 00h05m03s 1/1/2 (VpnIPv4)
                           28    0           0/0/0 (MvpnIPv4)
192.0.2.4
Def. Instance  64496      227    0 00h05m02s 1/1/2 (VpnIPv4)
                           27    0           0/0/0 (MvpnIPv4)
172.16.15.2
Svc: 1         64497      217    0 01h44m14s 4/1/4 (IPv4)
                          218    0
-------------------------------------------------------------------------------
*A:PE-1#
```

BGP Auto-Discovery — PE VPRN Multicast Configuration

Each PE contains a CE which will be part of the multicast VRF, so it is necessary to enable PIM on each interface containing an attachment circuit toward a CE, and to configure the I-PMSI multicast tunnel for the VRF.

In order for the BGP routes to be accepted into the VRF, a route-target community is required (vrf-target). This is configured in the **configure service vprn 1 mvpn** context and, in this case, is set to the same value as the unicast vrf-target, the vrf-target community as the **configure service vprn 1 vrf-target** context.

On each PE, the MVPN context of the VPRN instance is configured as follows:

```
*A:PE-2# configure
    service
        vprn 1
            mvpn
                auto-discovery default
                provider-tunnel
                    inclusive
                        pim asm 239.255.255.1
                        exit
                    exit
                exit
                vrf-target unicast
                exit
            exit
```

The multicast group address used for the PMSI must be the same on all PEs for this VPRN instance.

The presence of auto-discovery will initiate BGP updates between the PEs that contain an MVPN, such as Intra-AD MVPN routes, are generated and advertised to each peer

```
*A:PE-1# show router bgp routes mvpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
Flag  RouteType                   OriginatorIP           LocalPref   MED
      RD                          SourceAS                           Label
      Nexthop                     SourceIP
      As-Path                     GroupIP
-------------------------------------------------------------------------------
u*>i  Intra-Ad                    192.0.2.2              100         0
      64496:1                     -                                  -
      192.0.2.2                   -
      No As-Path                  -
u*>i  Intra-Ad                    192.0.2.3              100         0
      64496:1                     -                                  -
      192.0.2.3                   -
      No As-Path                  -
u*>i  Intra-Ad                    192.0.2.4              100         0
      64496:1                     -                                  -
      192.0.2.4                   -
      No As-Path                  -
-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:PE-1#
```

This shows that PE-1 has received an Intra-AD route from each of the other PEs, each of which has multicast VPRN 1 configured.

Examining the intra-AD routes received from PE-2 shows that the route-target community matches the unicast VRF-target (64496:1), and also that the PMSI tree has a multicast group address of 239.255.255.1, which matches the I-PMSI group configuration on PE-1.

```
*A:PE-1# show router bgp routes mvpn-ipv4 type intra-ad originator-ip 192.0.2.2
        detail
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
Original Attributes

Route Type     : Intra-Ad
Route Dist.    : 64496:1
```

```
Originator IP  : 192.0.2.2
Nexthop        : 192.0.2.2
From           : 192.0.2.2
Res. Nexthop   : 0.0.0.0
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : no-export target:64496:1
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.2
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h00m40s
VPRN Imported  :  1
-------------------------------------------------------------------------------
PMSI Tunnel Attribute :
Tunnel-type    : PIM-SM Tree
Flags          : Type: RNVE(0) BM: 0 U: 0 Leaf: not required
MPLS Label     : 0
Sender         : 192.0.2.2              P-Group        : 239.255.255.1
-------------------------------------------------------------------------------
---snip---

-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-1#
```

Verify that PIM in the GRT has signaled the I-PMSIs.

For the PE acting as the RP for global PIM:

```
*A:PE-1# show router pim group

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address                Type               Spt Bit  Inc Intf        No.Oifs
   Source Address               RP                 State    Inc Intf(S)
-------------------------------------------------------------------------------
239.255.255.1               (*,G)                                        3
   *                            192.0.2.1
239.255.255.1               (S,G)               spt      system          3
   192.0.2.1                    192.0.2.1
239.255.255.1               (S,G)               spt      int-PE-1-PE-2   3
   192.0.2.2                    192.0.2.1
239.255.255.1               (S,G)               spt      int-PE-1-PE-3   3
   192.0.2.3                    192.0.2.1
```

```
239.255.255.1               (S,G)               spt      int-PE-1-PE-2  2
   192.0.2.4                   192.0.2.1
-------------------------------------------------------------------------------
Groups : 5
===============================================================================
*A:PE-1#
```

This shows an incoming (S,G) join from all other PEs within the multicast VRF, plus an outgoing (*,G) join to the same PEs.

PE-3 will have the following PIM groups:

```
*A:PE-3# show router pim group

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
239.255.255.1               (*,G)                        int-PE-3-PE-1  1
   *                           192.0.2.1
239.255.255.1               (S,G)               spt      system         2
   192.0.2.3                   192.0.2.1
-------------------------------------------------------------------------------
Groups : 2
===============================================================================
*A:PE-3#
```

This shows a (S,G) join toward the RP at 192.0.2.1, plus a (*,G) join from RP. These represent the outgoing and incoming PIM interfaces for the VRF.

This results in a series of PIM neighbors through the I-PMSIs within the VRF. The neighbors were discovered using BGP (rather than with PIM as per Rosen MVPN), therefore, there are no PIM hellos exchanged.

```
*A:PE-1# show router 1 pim neighbor

===============================================================================
PIM Neighbor ipv4
===============================================================================
Interface               Nbr DR Prty     Up Time       Expiry Time     Hold Time
   Nbr Address
-------------------------------------------------------------------------------
int-PE-1-CE-5           1               0d 00:16:24   0d 00:01:22     105
   172.16.15.2
1-mt-239.255.255.1      1               0d 00:01:04   never           65535
   192.0.2.2
1-mt-239.255.255.1      1               0d 00:01:03   never           65535
   192.0.2.3
1-mt-239.255.255.1      1               0d 00:00:53   never           65535
   192.0.2.4
-------------------------------------------------------------------------------
```

```
Neighbors : 4
===============================================================================
*A:PE-1#
```

BGP Auto-Discovery — Customer Signaling Domain

The customer signaling is independent from the provider PE discovery mechanism, therefore, all of the customer signaling techniques described when using PIM for auto-discovery within provider domain are also applicable when using BGP for auto-discovery, namely

- PIM Any Source Multicasting with RP at the provider PE
- PIM Any Source Multicasting with Anycast RP at the provider PE
- PIM Source Specific Multicasting

Data Path Using Selective PMSI

When a configurable data threshold for a multicast group has been exceeded, multicast traffic across the provider network can be switched to a Selective PMSI (S-PMSI).

This has to be configured as a separate group and must contain a threshold which, if exceeded, will see a new PMSI signaled by the PE nearest the source, and traffic switched onto the S-PMSI.

```
*A:PE-1# configure
    service
        vprn
            mvpn
                provider-tunnel
                    inclusive
                        pim asm 239.255.255.1
                        exit
                    exit
                    selective
                        data-threshold 232.0.0.0/8 1
                        pim-ssm 232.255.1.0/24
                    exit
                exit
            exit
```

This shows that when the traffic threshold for multicast groups covered by the range 232.0.0.0/8 exceeds 1 kb/s between a pair of PEs, then an S-PMSI is signaled between the PEs. This is a separate multicast tunnel over which traffic in that group now flows.

```
*A:PE-1# show router 1 pim s-pmsi detail

===============================================================================
PIM Selective provider tunnels
===============================================================================
Md Source Address  : 192.0.2.1          Md Group Address   : 232.255.1.0
Number of VPN SGs  : 1                  Uptime             : 0d 00:00:16
MT IfIndex         : 16389

VPN Group Address  : 232.0.0.1
VPN Source Address : 192.168.55.2
State              : TX Joined          Mdt Threshold      : 1
Join Timer         : 0d 00:01:02        Holddown Timer     : 0d 00:00:44
===============================================================================
PIM Selective provider tunnels Interfaces : 1
===============================================================================
*A:PE-1#
```

In this example, the (S,G) group is (192.168.55.2, 232.0.0.1). When the data rate has exceeded the configured MDT threshold of 1 kb/s, a new provider tunnel with a group address of 232.255.1.0 has been signaled and now carries the multicast stream.

The TX Joined state indicates that the S-PMSI has been sourced at this PE — PE-1.

Comparing this to PE-3, where a receiver is connected through a CE indicates that it has received a join to connect the S-PMSI.

```
*A:PE-3# show router 1 pim s-pmsi detail
===============================================================================
PIM Selective provider tunnels
===============================================================================
Md Source Address  : 192.0.2.1          Md Group Address   : 232.255.1.0
Number of VPN SGs  : 1                  Uptime             : 0d 00:00:24
MT IfIndex         : 24576              Egress Fwding Rate : 7790.4 kbps

VPN Group Address  : 232.0.0.1
VPN Source Address : 192.168.55.2
State              : RX Joined
Expiry Timer       : 0d 00:02:36
===============================================================================
PIM Selective provider tunnels Interfaces : 1
===============================================================================
*A:PE-3#
```

Conclusion

This chapter provides configuration on how to configure multicast within a VPRN with next generation multicast VPN techniques. Specifically, discovery of multicast VPNs using PIM and BGP auto-discovery mechanisms are described with a number of ASM and SSM signaling techniques within the customer domain.

NG-MVPN Sender-Only, Receiver-Only

This chapter provides information about next generation multicast virtual private network (NG-MVPN) sender-only and receiver-only configurations.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

The sender-only/receiver-only feature as described in this chapter is supported in SR OS release 11.0.R1, and later. The CLI in this edition is based on SR OS release 15.0.R4.

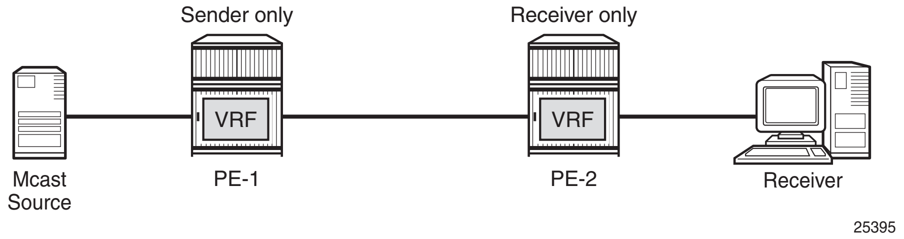
Knowledge of the Nokia multicast and Layer 3 VPNs concepts are assumed throughout this document.

Overview

This example covers a basic technology overview, the network topology, and configuration examples which are used for the Multicast VPN (MVPN) sender-only, receiver-only feature.

By default, if multiple PE nodes form a peering relationship within a common MVPN instance, then each PE node originates a multicast tree locally towards the remaining PE nodes that are members of this MVPN instance. This behavior creates a full mesh of Inclusive-Provider Multicast Service Interfaces (I-PMSIs) across all PE nodes in the MVPN.

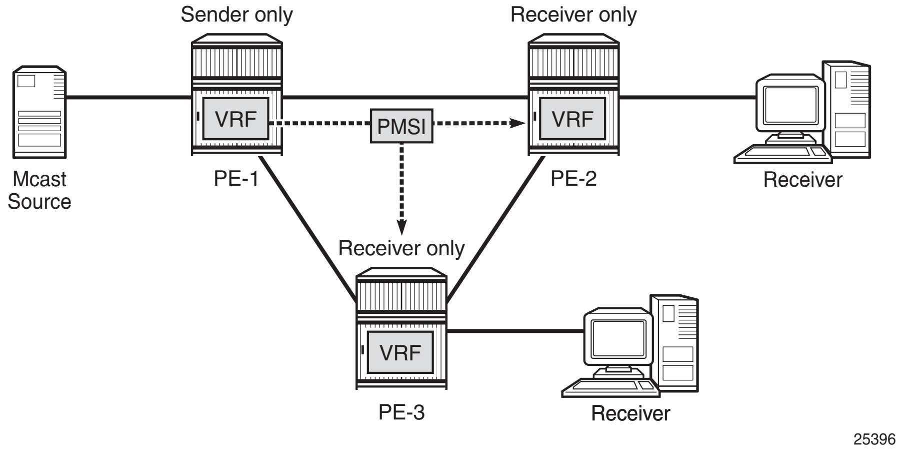
Figure 172 **Default PMSI Hierarchy**



It is often the case that an MVPN has many sites with multicast receivers, but only a few sites that host either both receivers and sources, or sources only.

The MVPN sender-only/receiver-only feature optimizes control and data plane resources by preventing unnecessary I-PMSI meshing when a PE hosts multicast sources only, or multicast receivers only, for an MVPN. An example of such an optimization is presented in Figure 173.

Figure 173 **Optimized PMSI Structure**



The general rules to follow are:

- For PE nodes that host only multicast sources for a given MVPN, operators can now block these PEs, through configuration, from joining I-PMSIs from the other PEs in this MVPN.
- For PE nodes that host only multicast receivers for a given MVPN, operators can now block these PEs, through configuration, to set-up a local I-PMSI to the other PEs in this MVPN.

MVPN sender-only/receiver-only is supported with next generation-MVPN for both IPv4 and IPv6 customer multicast using:

- IPv4 RSVP-TE provider tunnels
- IPv4 LDP provider tunnels

Extra attention should be given to the Bootstrap Router/Rendezvous Point (BSR/RP) placement when sender-only/receiver-only is enabled:

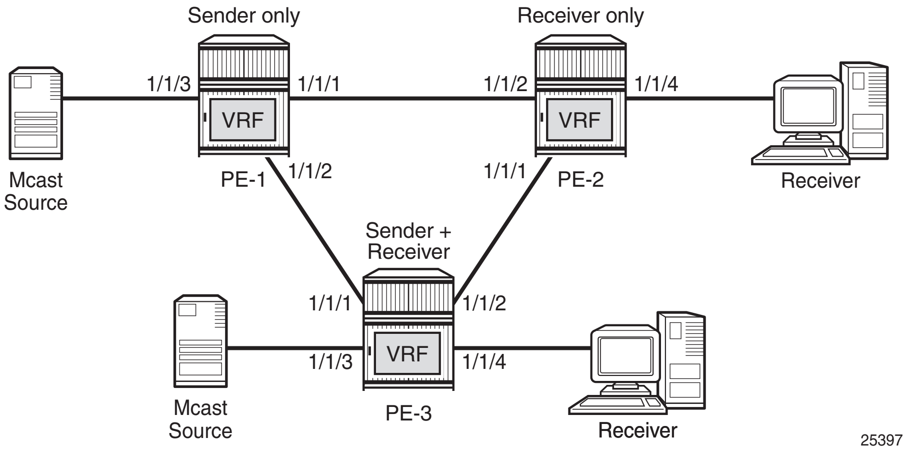
- The RP should be at the sender-receiver or sender-only site so that (*,G) traffic can be sent over the tunnel
- The BSR should be deployed at the sender-receiver site.
- The BSR can be at a sender-only site if the RPs are at the same site.

Note: (*,G) refers to an individual multicast stream indicating any source (*) and the multicast group (G) used by the stream.

Configuration

The example topology is shown in Figure 174.

Figure 174 **Example Topology**



To configure the sender-only/receiver-only feature, the following configuration command is used:

```
*A:PE>config>service>vprn>mvpn# mdt-type
  - mdt-type {sender-only|receiver-only|sender-receiver}
  - no mdt-type
```

sender-receiver is the default option and is visible using the **info detail** command.

This command restricts the MVPN instance to a specific role and provides an option to configure either a sender-only or receiver-only mode per PE node per service.

Parameters:

sender-only — MVPN has only senders connected to PE node.

receiver-only — MVPN has only receivers connected to PE node.

sender-receiver — MVPN has both sender and receivers connected to PE node.

Considerations:

- Two general approaches for building MVPNs will be covered in detail in this example:
  - Point-to-multipoint (P2MP) RSVP MVPNs
  - Multicast LDP (mLDP) MVPNs
- IPv4 and IPv6 multicast streaming are used for every MVPN at the same time.
- Basic principles of an MVPN including I-PMSI, S-PMSI, mLDP and P2MP RSVP are covered in the NG-MVPN Configuration with PIM and chapters of this guide.

PIM SSM is used for IPv4/IPv6 Customer (C)-multicast groups.

Initial Configuration

Step 1. The PE routers already have the following configuration:

- Interfaces (IPv4/IPv6)
- IGP (IS-IS or OSPF/OSPFv3)
- LDP (IPv4 only suffices)
- MPLS/RSVP
- BGP

Step 2. The MPLS/RSVP configuration on PE-1 is as follows. An P2MP LSP template is created with an empty path, without explicit hops.

```
# on PE-1
```

```
configure
    router
        mpls
            interface system
            exit
            interface "int-PE-1-PE-2"
            exit
            interface "int-PE-1-PE-3"
            exit
            no shutdown
            path EMPTY
                no shutdown
            exit
            lsp-template MVPN-P2MP-LSP p2mp
                default-path EMPTY
                cspf
                no shutdown
            exit
        exit
        rsvp no shutdown
    exit
exit
```

Step 3. The BGP configuration on PE-1 is as follows. No route reflector is used.

```
# on PE-1
configure
    router
        bgp
            min-route-advertisement 1
            enable-peer-tracking
            rapid-withdrawal
            rapid-update mvpn-ipv4 mvpn-ipv6
            group INTERNAL
                family vpn-ipv4 vpn-ipv6 mvpn-ipv4 mvpn-ipv6
                type internal
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
            exit
        exit
    exit
exit
```

RSVP-Based MVPN Configuration

Step 1. Configure a basic MVPN using P2MP RSVP as a transport protocol for C-multicast groups.

For this setup, PE-2 and PE-3 are configured to receive the following multicast groups:

- IPv4 group 232.0.0.1 source 172.16.1.1
- IPv6 group FF3E::8000:1 source 2001:DB8:1::1

Step 2. Configure the MDT type for the MVPN.

Based on the example topology, PE-1 is configured as **sender-only** for the MVPN.

```
# on PE-1
configure
    service
        vprn 1 customer 1 create
            description "RSVP-based MVPN"
            ecmp 2
            autonomous-system 64500
            route-distinguisher 64500:101
            auto-bind-tunnel
                resolution-filter
                    ldp
                    rsvp
                exit
                resolution filter
            exit
            vrf-target target:64500:1
            interface "int-PE-1-S-1" create
                description "to multicast source"
                address 172.16.1.2/30
                ipv6
                    address 2001:db8:1::2/126
                exit
                sap 1/1/3 create
                exit
            exit
            pim
                no ipv6-multicast-disable
                apply-to all
            exit
            mvpn
                auto-discovery default
                c-mcast-signaling bgp
                mdt-type sender-only
                provider-tunnel
                    inclusive
                        rsvp
                            lsp-template "MVPN-P2MP-LSP"
                            no shutdown
                        exit
                    exit
                exit
                vrf-target unicast
                exit
            exit
            service-name "RSVP-based MVPN"
            no shutdown
        exit
```

Based on the example topology, PE-2 is configured as **receiver-only** for the MVPN. PE-2 has static joins for the IPv4 and IPv6 multicast groups:

- group 232.0.0.1,source 172.16.1.1
- group FF3E::8000:1, source 2001:DB8:1::1

```
# on PE-2
configure
    service
        vprn 1 customer 1 create
            description "RSVP-based MVPN"
            ecmp 2
            autonomous-system 64500
            route-distinguisher 64500:102
            ignore-nh-metric
            auto-bind-tunnel
                resolution-filter
                    ldp
                    rsvp
                exit
                resolution filter
            exit
            vrf-target target:64500:1
            interface "int-PE-2-H-2" create
                description "to receiver Host-2"
                address 172.16.2.2/30
                ipv6
                    address 2001:db8:2::2/126
                exit
                sap 1/1/4 create
                exit
            exit
            igmp
                interface "int-PE-2-H-2"
                    static
                        group 232.0.0.1
                            source 172.16.1.1
                        exit
                    exit
                    no shutdown
                exit
                no shutdown
            exit
            mld
                interface "int-PE-2-H-2"
                    static
                        group ff3e::8000:1
                            source 2001:db8:1::1
                        exit
                    exit
                    no shutdown
                exit
                no shutdown
            exit
            pim
                no ipv6-multicast-disable
            exit
            mvpn
```

```
                auto-discovery default
                c-mcast-signaling bgp
                mdt-type receiver-only
                provider-tunnel
                    inclusive
                        rsvp
                            lsp-template "MVPN-P2MP-LSP"
                            no shutdown
                        exit
                    exit
                exit
                vrf-target unicast
                exit
            exit
            service-name "RSVP-based MVPN"
            no shutdown
        exit
    exit
exit
```

Based on the example topology, PE-3 is configured as **sender-receiver** (default) for the MVPN. PE-1 has also static joins for the IPv4 and IPv6 multicast groups:

- group 232.0.0.1,source 172.16.1.1
- group FF3E::8000:1, source 2001:DB8:1::1

The interface to the local source for PE-3 is not configured in this example. PE-3 acts as a receiver, not as a sender. Nonetheless, it is configured as sender-receiver and that has its consequences for the I-PMSIs that will be established.

```
# on PE-3
configure
    service
        vprn 1 customer 1 create
            description "RSVP-based MVPN"
            ecmp 2
            autonomous-system 64500
            route-distinguisher 64500:103
            auto-bind-tunnel
                resolution-filter
                    ldp
                    rsvp
                exit
                resolution filter
            exit
            vrf-target target:64500:1
            interface "int-PE-3-H-3" create
                description "to receiver Host-3"
                address 172.16.3.2/30
                ipv6
                    address 2001:db8:3::2/126
                exit
                sap 1/1/4 create
                exit
            exit
            igmp
```

```
                interface "int-PE-3-H-3"
                    static
                        group 232.0.0.1
                            source 172.16.1.1
                        exit
                    exit
                    no shutdown
                exit
                no shutdown
            exit
            mld
                interface "int-PE-3-H-3"
                    static
                        group ff3e::8000:1
                            source 2001:db8:1::1
                        exit
                    exit
                    no shutdown
                exit
                no shutdown
            exit
            pim
                no ipv6-multicast-disable
                apply-to all
            exit
            mvpn
                auto-discovery default
                c-mcast-signaling bgp
                provider-tunnel
                    inclusive
                        rsvp
                            lsp-template "MVPN-P2MP-LSP"
                            no shutdown
                        exit
                    exit
                exit
                vrf-target unicast
                exit
            exit
            service-name "RSVP-based MVPN"
            no shutdown
        exit
    exit
exit
```

The PIM instance must be **shutdown** before the mdt-type is modified; this leads to a multicast service disruption. Trying to change the mdt-type with PIM instance active will result in the following message being displayed.

```
*A:PE-1# configure service vprn 1 mvpn mdt-type receiver-only
MINOR: PIM #1100 PIM instance must be shutdown before changing this configuration
```

RSVP-Based MVPN Verification and Debugging

MDT-Type Verification

The status of the MVPN can be checked using the **show>router <***service-number***> mvpn** command:

The output for PE-1, PE-2 and PE-3 is as follows:

```
*A:PE-1# show router 1 mvpn

===============================================================================
MVPN 1 configuration data
===============================================================================
signaling          : Bgp                  auto-discovery     : Default
UMH Selection      : Highest-Ip           SA withdrawn       : Disabled
intersite-shared   : Enabled              Persist SA         : Disabled
vrf-import         : N/A
vrf-export         : N/A
vrf-target         : unicast
C-Mcast Import RT  : target:192.0.2.1:2

ipmsi              : rsvp MVPN-P2MP-LSP
i-pmsi P2MP AdmSt  : Up
i-pmsi Tunnel Name : MVPN-P2MP-LSP-1-73728
enable-bfd-root    : false                enable-bfd-leaf    : false
Mdt-type           : sender-only

BSR signalling     : none
Wildcard s-pmsi    : Disabled
Multistream-SPMSI  : Disabled
s-pmsi             : none
data-delay-interval: 3 seconds
enable-asm-mdt     : N/A

===============================================================================
*A:PE-1#


*A:PE-2# show router 1 mvpn

===============================================================================
MVPN 1 configuration data
===============================================================================
signaling          : Bgp                  auto-discovery     : Default
UMH Selection      : Highest-Ip           SA withdrawn       : Disabled
intersite-shared   : Enabled              Persist SA         : Disabled
vrf-import         : N/A
vrf-export         : N/A
vrf-target         : unicast
C-Mcast Import RT  : target:192.0.2.2:2

ipmsi              : rsvp MVPN-P2MP-LSP
i-pmsi P2MP AdmSt  : Up
i-pmsi Tunnel Name : mpls-virt-if-1005857
```

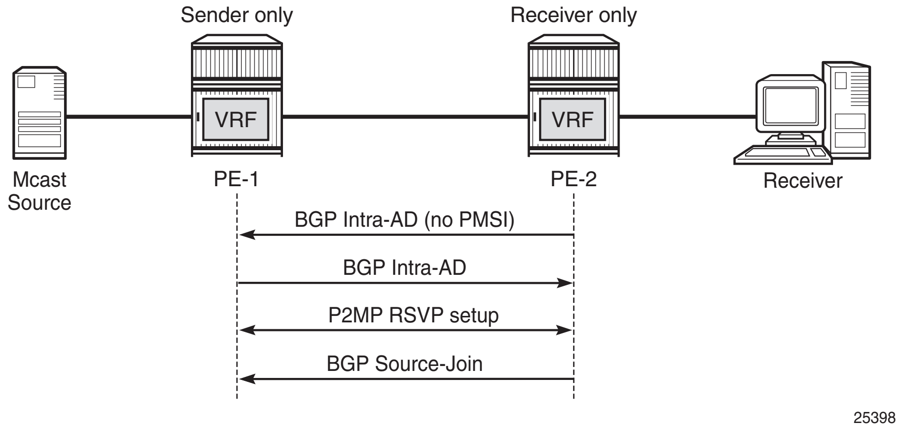
```
enable-bfd-root    : false               enable-bfd-leaf    : false
Mdt-type           : receiver-only

BSR signalling     : none
Wildcard s-pmsi    : Disabled
Multistream-SPMSI  : Disabled
s-pmsi             : none
data-delay-interval: 3 seconds
enable-asm-mdt     : N/A

===============================================================================
*A:PE-2#


*A:PE-3# show router 1 mvpn

===============================================================================
MVPN 1 configuration data
===============================================================================
signaling          : Bgp                 auto-discovery     : Default
UMH Selection      : Highest-Ip          SA withdrawn       : Disabled
intersite-shared   : Enabled             Persist SA         : Disabled
vrf-import         : N/A
vrf-export         : N/A
vrf-target         : unicast
C-Mcast Import RT  : target:192.0.2.3:2

ipmsi              : rsvp MVPN-P2MP-LSP
i-pmsi P2MP AdmSt  : Up
i-pmsi Tunnel Name : MVPN-P2MP-LSP-1-73728
enable-bfd-root    : false               enable-bfd-leaf    : false
Mdt-type           : sender-receiver

BSR signalling     : none
Wildcard s-pmsi    : Disabled
Multistream-SPMSI  : Disabled
s-pmsi             : none
data-delay-interval: 3 seconds
enable-asm-mdt     : N/A

===============================================================================
*A:PE-3#
```

BGP Verification and Debugging

When the MDT type is changed, the BGP signaling is slightly modified in order to achieve the signaling optimization.

The PE router does not include the PMSI part in the Intra-AD BGP messages when the MVPN is configured with mdt-type as **receiver-only**. The message flow is presented in Figure 175.

Figure 175 **RSVP-Based BGP Message Flow Between PE-1 and PE-2**



The following BGP debug output is taken from PE-2 and demonstrates the message flow between PE-1 and PE-2 for the MVPN-IPv4 address family.

There is no PMSI part in the BGP Intra-AD message sent by PE-2 (message 7), but the PMSI part is present in the BGP Intra-AD message received from **sender-only** PE-1 (message 1).

```
1 2017/10/02 13:35:34.946 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 86
    Flag: 0x90 Type: 14 Len: 23 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.1
        Type: Intra-AD Len: 12 RD: 64500:101 Orig: 192.0.2.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64500:1
    Flag: 0xc0 Type: 22 Len: 17 PMSI:
        Tunnel-type RSVP-TE P2MP LSP (1)
        Flags: (0x0)[Type: None BM: 0 U: 0 Leaf: not required]
        MPLS Label 0
        P2MP-ID 0x1, Tunnel-ID: 61441, Extended-Tunnel-ID 192.0.2.1
"


7 2017/10/02 13:35:45.192 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 66
    Flag: 0x90 Type: 14 Len: 23 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.2
        Type: Intra-AD Len: 12 RD: 64500:102 Orig: 192.0.2.2
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64500:1
"

19 2017/10/02 13:35:48.580 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 76
    Flag: 0x90 Type: 14 Len: 33 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.2
        Type: Source-Join Len:22 RD: 64500:101 SrcAS: 64500
                          Src: 172.16.1.1 Grp: 232.0.0.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:192.0.2.1:2
"
```

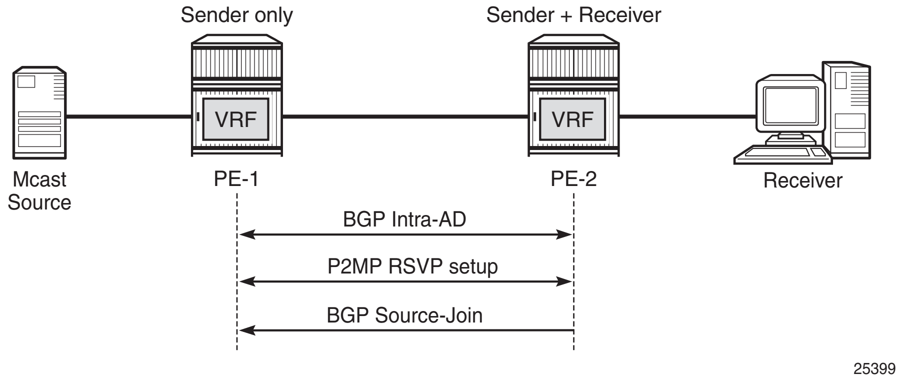
Similar behavior is observed for IPv6 multicast.The following BGP debug output is also taken from PE-2 and demonstrates the message flow between PE-1 and PE-2 for the MVPN-IPv6 address family.

There is no PMSI part in the Intra-AD message sent by PE-2 (message 8).

```
2 2017/10/02 13:35:34.946 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 86
    Flag: 0x90 Type: 14 Len: 23 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV6
        NextHop len 4 NextHop 192.0.2.1
        Type: Intra-AD Len: 12 RD: 64500:101 Orig: 192.0.2.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64500:1
    Flag: 0xc0 Type: 22 Len: 17 PMSI:
        Tunnel-type RSVP-TE P2MP LSP (1)
        Flags: (0x0)[Type: None BM: 0 U: 0 Leaf: not required]
        MPLS Label 0
        P2MP-ID 0x1, Tunnel-ID: 61441, Extended-Tunnel-ID 192.0.2.1
"

8 2017/10/02 13:35:45.192 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 66
    Flag: 0x90 Type: 14 Len: 23 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV6
        NextHop len 4 NextHop 192.0.2.2
        Type: Intra-AD Len: 12 RD: 64500:102 Orig: 192.0.2.2
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64500:1
"

20 2017/10/02 13:35:48.580 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 100
    Flag: 0x90 Type: 14 Len: 57 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV6
        NextHop len 4 NextHop 192.0.2.2
        Type: Source-Join Len:46 RD: 64500:101 SrcAS: 64500
                          Src: 2001:db8:1::1 Grp: ff3e::8000:1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:192.0.2.1:2
"
```

The PE router does not change its BGP behavior when the MVPN is configured with mdt-type as **sender-only**. The message flow is presented in Figure 176.

Figure 176 **RSVP-Based BGP Message Flow Between PE-1 and PE-3**



The BGP following debug output is taken from PE-3 and demonstrates the message flow between PE-1 and PE-3 for the MVPN-IPv4 address family.

The PMSI part is present in debug message 1, which is sent by PE-1 (**sender-only**).

```
1 2017/10/02 13:35:34.945 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 86
    Flag: 0x90 Type: 14 Len: 23 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.1
        Type: Intra-AD Len: 12 RD: 64500:101 Orig: 192.0.2.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64500:1
    Flag: 0xc0 Type: 22 Len: 17 PMSI:
        Tunnel-type RSVP-TE P2MP LSP (1)
        Flags: (0x0)[Type: None BM: 0 U: 0 Leaf: not required]
        MPLS Label 0
        P2MP-ID 0x1, Tunnel-ID: 61441, Extended-Tunnel-ID 192.0.2.1
"


13 2017/10/02 13:35:59.756 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 86
    Flag: 0x90 Type: 14 Len: 23 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.3
        Type: Intra-AD Len: 12 RD: 64500:103 Orig: 192.0.2.3
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64500:1
    Flag: 0xc0 Type: 22 Len: 17 PMSI:
        Tunnel-type RSVP-TE P2MP LSP (1)
        Flags: (0x0)[Type: None BM: 0 U: 0 Leaf: not required]
        MPLS Label 0
        P2MP-ID 0x1, Tunnel-ID: 61441, Extended-Tunnel-ID 192.0.2.3
"


31 2017/10/02 13:36:03.129 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 76
    Flag: 0x90 Type: 14 Len: 33 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.3
        Type: Source-Join Len:22 RD: 64500:101 SrcAS: 64500
                          Src: 172.16.1.1 Grp: 232.0.0.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:192.0.2.1:2
"
```

Similar behavior is observed for IPv6 multicast.

The following BGP debug output is taken from PE-3 and demonstrates the message flow between PE-1 and PE-3 for the MVPN-IPv6 address family.

The PMSI part is present in debug message 4, which is sent by PE-1 (**sender-only**).

```
2 2017/10/02 13:35:34.945 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 86
    Flag: 0x90 Type: 14 Len: 23 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV6
        NextHop len 4 NextHop 192.0.2.1
        Type: Intra-AD Len: 12 RD: 64500:101 Orig: 192.0.2.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64500:1
    Flag: 0xc0 Type: 22 Len: 17 PMSI:
        Tunnel-type RSVP-TE P2MP LSP (1)
        Flags: (0x0)[Type: None BM: 0 U: 0 Leaf: not required]
        MPLS Label 0
        P2MP-ID 0x1, Tunnel-ID: 61441, Extended-Tunnel-ID 192.0.2.1
"

14 2017/10/02 13:35:59.756 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 86
    Flag: 0x90 Type: 14 Len: 23 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV6
        NextHop len 4 NextHop 192.0.2.3
        Type: Intra-AD Len: 12 RD: 64500:103 Orig: 192.0.2.3
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64500:1
    Flag: 0xc0 Type: 22 Len: 17 PMSI:
        Tunnel-type RSVP-TE P2MP LSP (1)
        Flags: (0x0)[Type: None BM: 0 U: 0 Leaf: not required]
        MPLS Label 0
        P2MP-ID 0x1, Tunnel-ID: 61441, Extended-Tunnel-ID 192.0.2.3
"

32 2017/10/02 13:36:03.129 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 100
    Flag: 0x90 Type: 14 Len: 57 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV6
        NextHop len 4 NextHop 192.0.2.3
        Type: Source-Join Len:46 RD: 64500:101 SrcAS: 64500
                          Src: 2001:db8:1::1 Grp: ff3e::8000:1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:192.0.2.1:2
"
```

The BGP routing table of each router is populated accordingly.

PE-1 (**sender-only**) has two Intra-Ad and two Source-Join messages from PE-2 and PE-3.

```
*A:PE-1# show router bgp routes mvpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
Flag  RouteType                   OriginatorIP           LocalPref   MED
      RD                          SourceAS               Path-Id     Label
      Nexthop                     SourceIP
      As-Path                     GroupIP
-------------------------------------------------------------------------------
u*>i  Source-Join                 -                      100         0
      64500:101                   64500                  None        -
      192.0.2.2                   172.16.1.1
      No As-Path                  232.0.0.1
*>i   Source-Join                 -                      100         0
      64500:101                   64500                  None        -
      192.0.2.3                   172.16.1.1
      No As-Path                  232.0.0.1
u*>i  Intra-Ad                    192.0.2.2              100         0
      64500:102                   -                      None        -
      192.0.2.2                   -
      No As-Path                  -
u*>i  Intra-Ad                    192.0.2.3              100         0
      64500:103                   -                      None        -
      192.0.2.3                   -
      No As-Path                  -
-------------------------------------------------------------------------------
Routes : 4
===============================================================================
*A:PE-1#
```

PE-2 (receiver-only) has two Intra-Ad messages from PE-1 and PE-3.

```
*A:PE-2# show router bgp routes mvpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
Flag  RouteType                   OriginatorIP           LocalPref   MED
      RD                          SourceAS               Path-Id     Label
      Nexthop                     SourceIP
      As-Path                     GroupIP
-------------------------------------------------------------------------------
u*>i  Intra-Ad                    192.0.2.1              100         0
```

```
      64500:101                   -                       None        -
      192.0.2.1                   -
      No As-Path                  -
u*>i  Intra-Ad                    192.0.2.3               100         0
      64500:103                   -                       None        -
      192.0.2.3                   -
      No As-Path                  -
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-2#
```

PE-3 (**sender-receiver**) has two Intra-Ad messages: one from PE-1 and one from PE-2.

```
*A:PE-3# show router bgp routes mvpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.3        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
Flag  RouteType                   OriginatorIP            LocalPref   MED
      RD                          SourceAS                Path-Id     Label
      Nexthop                     SourceIP
      As-Path                     GroupIP
-------------------------------------------------------------------------------
u*>i  Intra-Ad                    192.0.2.1               100         0
      64500:101                   -                       None        -
      192.0.2.1                   -
      No As-Path                  -
u*>i  Intra-Ad                    192.0.2.2               100         0
      64500:102                   -                       None        -
      192.0.2.2                   -
      No As-Path                  -
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-3#
```

RSVP Verification and Debugging

When BGP intra-AD messages are exchanged, every PE starts to build multicast tunnels based on the following criteria:

- PE nodes which are configured as **sender-only** for an MVPN do not join P2MP LSPs from other PEs in this MVPN.

- PE nodes which are configured as receiver-only for an MVPN do not originate P2MP LSPs to other PEs in this MVPN.

The RSVP session can be checked with the **show>router>rsvp>session** command:

PE-1 (192.0.2.1) has two originating LSPs: one toward PE-2 (192.0.2.2) and one toward PE-3 (192.0.2.3). PE-1 also has one incoming LSP from PE-3 (**mdt-type sender-receiver**).

```
*A:PE-1# show router rsvp session

===============================================================================
RSVP Sessions
===============================================================================
From            To              Tunnel LSP   Name                          State
                                ID     ID
-------------------------------------------------------------------------------
192.0.2.1       192.0.2.2       61441  16896 MVPN-P2MP-LSP-1-73728::EMPTY Up
192.0.2.1       192.0.2.3       61441  16896 MVPN-P2MP-LSP-1-73728::EMPTY Up
192.0.2.3       192.0.2.1       61441  36864 MVPN-P2MP-LSP-1-73728::EMPTY Up
-------------------------------------------------------------------------------
Sessions : 3
===============================================================================
*A:PE-1#
```

PE-2 (192.0.2.2) has two incoming LSPs from PE-1 (192.0.2.1) and PE-3 (192.0.2.3) and no originating LSPs due to the fact that PE-2 has **mdt-type receiver-only**.

```
*A:PE-2# show router rsvp session

===============================================================================
RSVP Sessions
===============================================================================
From            To              Tunnel LSP   Name                          State
                                ID     ID
-------------------------------------------------------------------------------
192.0.2.1       192.0.2.2       61441  16896 MVPN-P2MP-LSP-1-73728::EMPTY Up
192.0.2.3       192.0.2.2       61441  36864 MVPN-P2MP-LSP-1-73728::EMPTY Up
-------------------------------------------------------------------------------
Sessions : 2
===============================================================================
*A:PE-2#
```

PE-3 (192.0.2.3) has two originating LSPs: one toward PE-2 (192.0.2.2) and one toward PE-1 (192.0.2.1). PE-3 also has one incoming LSP from PE-1 (**mdt-type sender-only**).

Theoretically there is no need for the LSP from PE-3 toward PE-1, because PE-1 is a sender-only; this minor limitation should be taken into account during planning phase.

```
*A:PE-3# show router rsvp session
```

```
===============================================================================
RSVP Sessions
===============================================================================
From            To              Tunnel LSP   Name                         State
                                ID     ID
-------------------------------------------------------------------------------
192.0.2.1       192.0.2.3       61441  16896 MVPN-P2MP-LSP-1-73728::EMPTY Up
192.0.2.3       192.0.2.2       61441  36864 MVPN-P2MP-LSP-1-73728::EMPTY Up
192.0.2.3       192.0.2.1       61441  36864 MVPN-P2MP-LSP-1-73728::EMPTY Up
-------------------------------------------------------------------------------
Sessions : 3
===============================================================================
*A:PE-3#
```

Additional details about originating P2MP paths can be found using the following command:

show>router>mpls p2mp-lsp <*lsp name*> p2mp-instance <*service number*> s2l

The output for PE-1, PE-2 and PE-3 is as follows:

```
*A:PE-1# show router mpls p2mp-lsp "MVPN-P2MP-LSP-1-73728" p2mp-instance "1" s2l

===============================================================================
MPLS LSP MVPN-P2MP-LSP-1-73728 S2L
===============================================================================
-------------------------------------------------------------------------------
LSP Name         : MVPN-P2MP-LSP-1-737* P2MP ID              : 1
Adm State        : Up                   Oper State           : Up
P2MPInstance     : 1                    Inst-type            : Primary
Adm State        : Up                   Oper State           : Up
-------------------------------------------------------------------------------
S2l Name                       To              Next Hop        Adm  Opr
-------------------------------------------------------------------------------
EMPTY                          192.0.2.2       192.168.12.2    Up   Up
EMPTY                          192.0.2.3       192.168.13.2    Up   Up
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-1#
```

```
*A:PE-2# show router mpls p2mp-lsp

===============================================================================
MPLS P2MP LSPs (Originating)
===============================================================================
LSP Name                                             Tun     Fastfail  Adm  Opr
                                                     Id      Config
-------------------------------------------------------------------------------
No Matching Entries Found
===============================================================================
*A:PE-2#
```

```
*A:PE-3# show router mpls p2mp-lsp "MVPN-P2MP-LSP-1-73728" p2mp-instance "1" s2l

===============================================================================
MPLS LSP MVPN-P2MP-LSP-1-73728 S2L
===============================================================================
-------------------------------------------------------------------------------
LSP Name         : MVPN-P2MP-LSP-1-737* P2MP ID            : 1
Adm State        : Up                   Oper State         : Up
P2MPInstance     : 1                    Inst-type          : Primary
Adm State        : Up                   Oper State         : Up
-------------------------------------------------------------------------------
S2l Name                        To              Next Hop         Adm  Opr
-------------------------------------------------------------------------------
EMPTY                           192.0.2.1       192.168.13.1     Up   Up
EMPTY                           192.0.2.2       192.168.23.1     Up   Up
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-3#
```

Multicast Stream Verification

The status of the multicast groups/streams can be verified using the **show>router <*sid*>>pim group detail ipv6** command:

There is an IPv4 sender connected to PE-1. The physical interface where the sender is connected is used as the incoming interface. An I-PMSI is used as the outgoing interface.

```
*A:PE-1# show router 1 pim group detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 232.0.0.1
Source Address     : 172.16.1.1
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 172.16.1.1
MRIB Src Flags     : direct
Keepalive Timer    : Not Running
Up Time            : 0d 00:03:55        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:00
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 172.16.1.1
Incoming Intf      : int-PE-1-S-1
Outgoing Intf List : mpls-if-73728

Curr Fwding Rate   : 1072.6 kbps
Forwarded Packets  : 1619               Discarded Packets  : 0
Forwarded Octets   : 2425262            RPF Mismatches     : 0
```

```
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-1#
```

There is an IPv4 receiver connected to PE-2. An I-PMSI is used as the incoming interface and the physical interface where the receiver is connected is used as the outgoing interface.

```
*A:PE-2# show router 1 pim group detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 232.0.0.1
Source Address     : 172.16.1.1
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 192.0.2.1
MRIB Src Flags     : remote
Keepalive Timer    : Not Running
Up Time            : 0d 00:04:00        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:01
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 192.0.2.1
Incoming Intf      : mpls-if-73728
Outgoing Intf List : int-PE-2-H-2

Curr Fwding Rate   : 1072.6 kbps
Forwarded Packets  : 1906               Discarded Packets  : 0
Forwarded Octets   : 2855188            RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-2#
```

There is an IPv4 receiver connected to PE-3. An I-PMSI is used as the incoming interface and the physical interface where receiver is connected is used as the outgoing interface.

```
*A:PE-3# show router 1 pim group detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 232.0.0.1
Source Address     : 172.16.1.1
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 192.0.2.1
MRIB Src Flags     : remote
Keepalive Timer    : Not Running
Up Time            : 0d 00:03:49        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:14
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 192.0.2.1
Incoming Intf      : mpls-if-73729
Outgoing Intf List : int-PE-3-H-3

Curr Fwding Rate   : 1072.6 kbps
Forwarded Packets  : 2135               Discarded Packets  : 0
Forwarded Octets   : 3198230            RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-3#
```

Similar behavior is observed for IPv6 multicast.

An IPv6 sender is connected to PE-1. The physical interface where the sender is connected is used as the incoming interface. An I-PMSI is used as the outgoing interface.

```
*A:PE-1# show router 1 pim group detail ipv6

===============================================================================
PIM Source Group ipv6
===============================================================================
Group Address      : ff3e::8000:1
Source Address     : 2001:db8:1::1
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 2001:db8:1::1
MRIB Src Flags     : direct
Keepalive Timer    : Not Running
Up Time            : 0d 00:04:08        Resolved By        : rtable6-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:00
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
```

```
Reg From Anycast RP: No

Rpf Neighbor       : 2001:db8:1::1
Incoming Intf      : int-PE-1-S-1
Outgoing Intf List : mpls-if-73728

--- snipped ---

-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-1#
```

An IPv6 receiver is connected to PE-2. An I-PMSI is used as the incoming interface and the physical interface where the receiver is connected is used as the outgoing interface.

```
*A:PE-2# show router 1 pim group detail ipv6

===============================================================================
PIM Source Group ipv6
===============================================================================
Group Address      : ff3e::8000:1
Source Address     : 2001:db8:1::1
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 192.0.2.1
MRIB Src Flags     : remote
Keepalive Timer    : Not Running
Up Time            : 0d 00:04:12        Resolved By        : rtable6-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:49
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 192.0.2.1
Incoming Intf      : mpls-if-73728
Outgoing Intf List : int-PE-2-H-2

--- snipped ---

-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-2#
```

An IPv6 receiver is connected to PE-3. An I-PMSI is used as the incoming interface and the physical interface where the receiver is connected is used as the outgoing interface.

```
*A:PE-3# show router 1 pim group detail ipv6
```

```
===============================================================================
PIM Source Group ipv6
===============================================================================
Group Address      : ff3e::8000:1
Source Address     : 2001:db8:1::1
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 192.0.2.1
MRIB Src Flags     : remote
Keepalive Timer    : Not Running
Up Time            : 0d 00:04:00        Resolved By        : rtable6-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:00
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 192.0.2.1
Incoming Intf      : mpls-if-73729
Outgoing Intf List : int-PE-3-H-3

--- snipped ---

-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-3#
```

mLDP-Based MVPN Configuration

Step 1. Reconfigure VPRN 1 to make it mLDP-based. The resolution-filter should only be LDP (no RSVP anymore) for auto-bind-tunnel. The MVPN context also changes: the inclusive provider-tunnel is mLDP-based. The MDT-type remains the same: PE-1 is sender-only, PE-2 is receiver-only and PE-3 is sender-receiver (default).

PE-2 and PE-3 have static joins for the IPv4/IPv6 multicast groups:

- group 232.0.0.1,source 172.16.1.1
- group FF3E::8000:1, source 2001:DB8:1::1

Step 2.The VPRN 1 configuration on PE-1 is as follows:

```
# on PE-1
configure
    service
        vprn 1 customer 1 create
            description "mLDP-based MVPN"
            ecmp 2
            autonomous-system 64500
            route-distinguisher 64500:101
            ignore-nh-metric
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64500:1
            interface "int-PE-1-S-1" create
                description "to multicast source"
                address 172.16.1.2/30
                ipv6
                    address 2001:db8:1::2/126
                exit
                sap 1/1/3 create
                exit
            exit
            pim
                no ipv6-multicast-disable
                apply-to all
            exit
            mvpn
                auto-discovery default
                c-mcast-signaling bgp
                mdt-type sender-only
                provider-tunnel
                    inclusive
                        mldp
                            no shutdown
                        exit
                    exit
                exit
                vrf-target unicast
                exit
            exit
            service-name "mLDP-based MVPN"
            no shutdown
        exit
```

Based on the example topology, PE-2 is configured as receiver-only for the MVPN. PE-2 has also static joins for the IPv4 and IPv6 multicast groups:

- group 232.0.0.1,source 172.16.3.1
- group FF3E::8000:1, source 2001:DB8:3::1

```
# on PE-2
configure
    service
        vprn 1 customer 1 create
            description "mLDP-based MVPN"
            ecmp 2
            autonomous-system 64500
            route-distinguisher 64500:102
            ignore-nh-metric
            auto-bind-tunnel
                resolution-filter
```

```
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64500:1
            interface "int-PE-2-H-2" create
                description "to receiver Host-2"
                address 172.16.2.2/30
                ipv6
                    address 2001:db8:2::2/126
                exit
                sap 1/1/4 create
                exit
            exit
            igmp
                interface "int-PE-2-H-2"
                    static
                        group 232.0.0.1
                            source 172.16.1.1
                        exit
                    exit
                    no shutdown
                exit
                no shutdown
            exit
            mld
                interface "int-PE-2-H-2"
                    static
                        group ff3e::8000:1 source 2001:db8:1::1
                    exit
                    no shutdown
                exit
                no shutdown
            exit
            pim
                no ipv6-multicast-disable
                apply-to all
            exit
            mvpn
                auto-discovery default
                c-mcast-signaling bgp
                mdt-type receiver-only
                provider-tunnel
                    inclusive
                        mldp
                            no shutdown
                        exit
                    exit
                exit
                vrf-target unicast
                exit
            exit
            service-name "mLDP-based MVPN"
            no shutdown
        exit
```

Based on the example topology, PE-3 is configured as **sender-receiver** (default) for the MVPN. PE-3 has also static joins for the IPv4 and IPv6 multicast groups:

- group 232.0.0.1,source 172.16.3.1
- group FF3E::8000:1, source 2001:DB8:3::1

```
# on PE-3
configure service
        vprn 1 customer 1 create
            description "mLDP-based MVPN"
            ecmp 2
            autonomous-system 64500
            route-distinguisher 64500:103
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64500:1
            interface "int-PE-3-H-3" create
                description "to receiver Host-3"
                address 172.16.3.2/30
                ipv6
                    address 2001:db8:3::2/126
                exit
                sap 1/1/4 create
                exit
            exit
            igmp
                interface "int-PE-3-H-3"
                    static
                        group 232.0.0.1
                            source 172.16.1.1
                        exit
                    exit
                    no shutdown
                exit
                no shutdown
            exit
            mld
                interface "int-PE-3-H-3"
                    static
                        group ff3e::8000:1
                            source 2001:db8:1::1
                        exit
                    exit
                    no shutdown
                exit
                no shutdown
            exit
            pim
                no ipv6-multicast-disable
                apply-to all
            exit
            mvpn
                auto-discovery default
                c-mcast-signaling bgp
                provider-tunnel
                    inclusive
                        mldp
                            no shutdown
                        exit
                    exit
                exit
                vrf-target unicast
                exit
            exit
            service-name "mLDP-based MVPN"
            no shutdown
        exit
```

mLDP-Based MVPN Verification and Debugging

MDT-Type Verification

The status of the MVPN can be checked using the following command:

```
show router <service-number> mvpn
```

The output for PE-1, PE-2 and PE-3 is as follows:

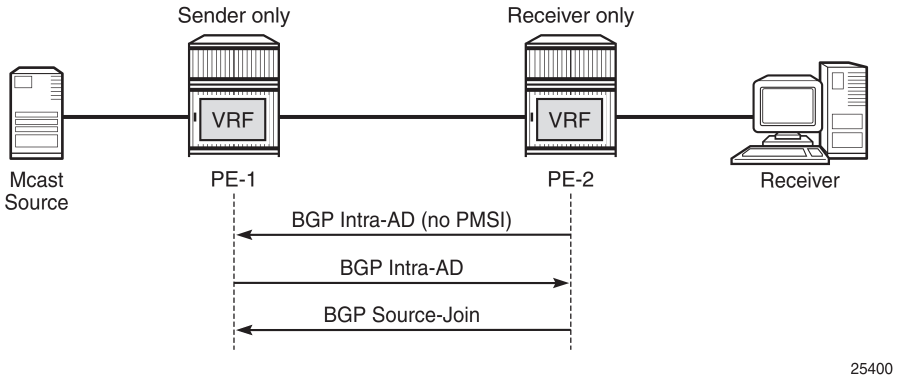
```
*A:PE-1# show router 1 mvpn

===============================================================================
MVPN 1 configuration data
===============================================================================
signaling          : Bgp                  auto-discovery     : Default
UMH Selection      : Highest-Ip           SA withdrawn       : Disabled
intersite-shared   : Enabled              Persist SA         : Disabled
vrf-import         : N/A
vrf-export         : N/A
vrf-target         : unicast
C-Mcast Import RT  : target:192.0.2.1:2

ipmsi              : ldp
i-pmsi P2MP AdmSt  : Up
i-pmsi Tunnel Name : mpls-if-73729
Mdt-type           : sender-only

BSR signalling     : none
Wildcard s-pmsi    : Disabled
Multistream-SPMSI  : Disabled
s-pmsi             : none
data-delay-interval: 3 seconds
enable-asm-mdt     : N/A

===============================================================================
*A:PE-1#


*A:PE-2# show router 1 mvpn

===============================================================================
MVPN 1 configuration data
===============================================================================
signaling          : Bgp                    auto-discovery     : Default
UMH Selection      : Highest-Ip             SA withdrawn       : Disabled
intersite-shared   : Enabled                Persist SA         : Disabled
vrf-import         : N/A
vrf-export         : N/A
vrf-target         : unicast
C-Mcast Import RT  : target:192.0.2.2:2

ipmsi              : ldp
i-pmsi P2MP AdmSt  : Up
i-pmsi Tunnel Name : mpls-virt-if-1005858
Mdt-type           : receiver-only

BSR signalling     : none
Wildcard s-pmsi    : Disabled
Multistream-SPMSI  : Disabled
s-pmsi             : none
data-delay-interval: 3 seconds
enable-asm-mdt     : N/A

===============================================================================
*A:PE-2#


*A:PE-3# show router 1 mvpn

===============================================================================
MVPN 1 configuration data
===============================================================================
signaling          : Bgp                    auto-discovery     : Default
UMH Selection      : Highest-Ip             SA withdrawn       : Disabled
intersite-shared   : Enabled                Persist SA         : Disabled
vrf-import         : N/A
vrf-export         : N/A
vrf-target         : unicast
C-Mcast Import RT  : target:192.0.2.3:2

ipmsi              : ldp
i-pmsi P2MP AdmSt  : Up
i-pmsi Tunnel Name : mpls-if-73730
Mdt-type           : sender-receiver

BSR signalling     : none
Wildcard s-pmsi    : Disabled
Multistream-SPMSI  : Disabled
s-pmsi             : none
data-delay-interval: 3 seconds
enable-asm-mdt     : N/A

===============================================================================
*A:PE-3#
```

BGP Verification and Debugging

When the MDT type is changed, the BGP signaling is slightly modified in order to achieve the signaling optimization.The PE router does not include the PMSI part in Intra-AD BGP messages when the MVPN is configured with mdt-type as **receiver-only**.

The message flow is presented in Figure 177.

Figure 177 **mLDP-Based BGP Message Flow Between PE-1 and PE-2**



In order to demonstrate the BGP message flow sequence the following initialization steps are taken on PE-2:

1. Bring down the VPRN service, PIM protocol in a VPRN and IGMP/MLD protocol. As a result, the state of all signaling protocols is cleared.
2. Bring up the VPRN service. BGP exchanges unicast routing information.
3. Bring up the IPv4 PIM protocol. BGP exchanges IPv4 multicast routing information in order to build the PMSI infrastructure.
4. Bring up IGMP and add a static IGMP join where it is applicable. BGP exchanges IPv4 multicast routing information in order to propagate the multicast traffic to the receiver.
5. Bring up the IPv6 PIM protocol. BGP exchanges IPv6 multicast routing information in order to build the PMSI infrastructure.
6. Bring up MLD and add a static MLD join where it is applicable. BGP exchanges IPv6 multicast routing information in order to propagate the multicast traffic to the receiver.

The following BGP debug is taken from PE-2 and demonstrates the message flow between PE-2 and PE-1. VPN-IPv4 and VPN-IPv6 updates are not present in this output.

Step 1. Bring down the VPRN service and protocols to clear the state of all signaling protocols.

```
# on PE-2
configure
    service
        vprn 1
            shutdown
            pim shutdown
            pim ipv6-multicast-disable
            igmp shutdown
            mld shutdown
        exit
    exit
```

Step 2. Enable the VPRN service on PE-2.

PE-2 immediately receives Intra-AD messages from PE-1 because the remote VPRN service is already enabled for IPv4 and IPv6 multicast propagation.

```
*A:PE-2# configure service vprn 1 no shutdown


13 2017/10/02 13:43:58.579 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 91
    Flag: 0x90 Type: 14 Len: 23 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.1
        Type: Intra-AD Len: 12 RD: 64500:101 Orig: 192.0.2.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64500:1
    Flag: 0xc0 Type: 22 Len: 22 PMSI:
        Tunnel-type LDP P2MP LSP (2)
        Flags: (0x0)[Type: None BM: 0 U: 0 Leaf: not required]
        MPLS Label 0
        Root-Node 192.0.2.1, LSP-ID 0x2001
"


14 2017/10/02 13:43:58.579 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 91
    Flag: 0x90 Type: 14 Len: 23 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV6
        NextHop len 4 NextHop 192.0.2.1
        Type: Intra-AD Len: 12 RD: 64500:101 Orig: 192.0.2.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64500:1
    Flag: 0xc0 Type: 22 Len: 22 PMSI:
        Tunnel-type LDP P2MP LSP (2)
        Flags: (0x0)[Type: None BM: 0 U: 0 Leaf: not required]
        MPLS Label 0
        Root-Node 192.0.2.1, LSP-ID 0x2001
"
```

Step 3. Enable only PIM IPv4 for the service on PE-2.

PE-2 immediately sends Intra-AD messages to PE-1. Note that no PMSI part is present in the debug message sent by receiver-only PE-2.

```
*A:PE-2# configure service vprn 1 pim no shutdown
```

```
16 2017/10/02 13:44:03.949 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 66
    Flag: 0x90 Type: 14 Len: 23 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.2
        Type: Intra-AD Len: 12 RD: 64500:102 Orig: 192.0.2.2
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64500:1
"
```

Step 4. Bring up IGMP and add a static IGMP join for the service on a PE-2.

PE-2 immediately sends a source-join message to PE-3 and receives a source-AD message from PE-1.

```
# on PE-2
configure
    service
        vprn 1
            igmp
                interface "int-PE-2-H-2"
                    static
                        group 232.0.0.1 source 172.16.1.1
                    exit
                    no shutdown
                exit
                no shutdown
            exit
        exit

18 2017/10/02 13:44:37.060 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 76
    Flag: 0x90 Type: 14 Len: 33 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.2
        Type: Source-Join Len:22 RD: 64500:101 SrcAS: 64500
                          Src: 172.16.1.1 Grp: 232.0.0.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:192.0.2.1:2
"
```

Step 5. Enable PIM IPv6 for the service on PE-2.

PE-2 immediately sends Intra-AD messages to PE-3.

```
*A:PE-2# configure service vprn 1 pim no ipv6-multicast-disable

20 2017/10/02 13:44:49.326 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 66
    Flag: 0x90 Type: 14 Len: 23 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV6
        NextHop len 4 NextHop 192.0.2.2
        Type: Intra-AD Len: 12 RD: 64500:102 Orig: 192.0.2.2
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64500:1
"
```

Step 6. Bring up MLD and add a static MLD join for the service on a PE-2.

PE-2 immediately sends a source-join message to PE-3 and receives a source-AD message from PE-3.

```
# on PE-2
configure
    service
        vprn 1
            mld
                interface "int-PE-2-H-2"
                    static
                        group ff3e::8000:1 source 2001:db8:1::1
                    exit
                    no shutdown
                exit
                no shutdown
            exit
        exit
```

```
22 2017/10/02 13:45:12.064 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 100
    Flag: 0x90 Type: 14 Len: 57 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV6
        NextHop len 4 NextHop 192.0.2.2
        Type: Source-Join Len:46 RD: 64500:101 SrcAS: 64500
                          Src: 2001:db8:1::1 Grp: ff3e::8000:1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:192.0.2.1:2
"
```

The same information can be gathered using the following show commands.

show>router>bgp>neighbor <*peer*> advertised-routes [mvpn-ipv4|mvpn-ipv6]

show>router>bgp>neighbor <*peer*> received-routes [mvpn-ipv4|mvpn-ipv6]

PE-2 output for the advertised routes for the mvpn-ipv4 address family is as follows:

```
*A:PE-2# show router bgp neighbor 192.0.2.1 advertised-routes mvpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
Flag  RouteType                   OriginatorIP           LocalPref   MED
      RD                          SourceAS               Path-Id     Label
      Nexthop                     SourceIP
      As-Path                     GroupIP
-------------------------------------------------------------------------------
i     Source-Join                 -                      100         0
      64500:101                   64500                  None        -
      192.0.2.2                   172.16.1.1
      No As-Path                  232.0.0.1
i     Intra-Ad                    192.0.2.2              100         0
      64500:102                   -                      None        -
      192.0.2.2                   -
      No As-Path                  -
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-2#
```

PE-2 output for the advertised routers for the mvpn-ipv6 address family is as follows:

```
*A:PE-2# show router bgp neighbor 192.0.2.1 advertised-routes mvpn-ipv6
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv6 Routes
===============================================================================
Flag  RouteType                   OriginatorIP           LocalPref   MED
      RD                          SourceAS               Path-Id     Label
      Nexthop                     SourceIP
      As-Path                     GroupIP
-------------------------------------------------------------------------------
i     Source-Join                 -                      100         0
      64500:101                   64500                  None        -
      192.0.2.2                   2001:db8:1::1
      No As-Path                  ff3e::8000:1
i     Intra-Ad                    192.0.2.2              100         0
      64500:102                   -                      None        -
      192.0.2.2                   -
      No As-Path                  -
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-2#
```

PE-2 output for the received routes for the mvpn-ipv4 address family is as follows:

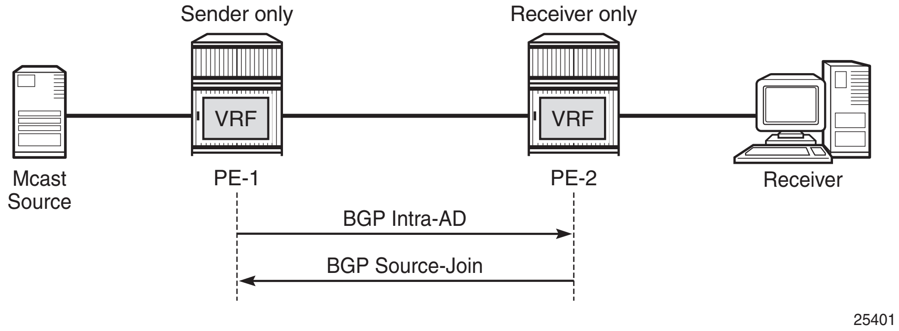
```
*A:PE-2# show router bgp neighbor 192.0.2.1 received-routes mvpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
Flag  RouteType                   OriginatorIP           LocalPref   MED
      RD                          SourceAS               Path-Id     Label
      Nexthop                     SourceIP
      As-Path                     GroupIP
-------------------------------------------------------------------------------
u*>i  Intra-Ad                    192.0.2.1              100         0
      64500:101                   -                      None        -
      192.0.2.1                   -
      No As-Path                  -
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-2#
```

PE-2 output for the received routes for the mvpn-ipv6 address family is as follows:

```
*A:PE-2# show router bgp neighbor 192.0.2.1 received-routes mvpn-ipv6
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv6 Routes
===============================================================================
Flag  RouteType                   OriginatorIP           LocalPref   MED
      RD                          SourceAS               Path-Id     Label
      Nexthop                     SourceIP
      As-Path                     GroupIP
-------------------------------------------------------------------------------
u*>i  Intra-Ad                    192.0.2.1              100         0
      64500:101                   -                      None        -
      192.0.2.1                   -
      No As-Path                  -
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-2#
```

The PE router does not change the BGP behavior when the MVPN is configured with mdt-type as **sender-only**. A schematic of the message flow is presented in Figure 178.

Figure 178 **mLDP-Based BGP Message Flow Between PE-1 and PE-3**



In order to demonstrate the BGP message flow sequence, the following initialization steps are taken:

1. Bring down the VPRN service, PIM protocol in the VPRN and IGMP/MLD protocol. As a result, the state of all signaling protocols is cleared.
2. Bring up the VPRN service. BGP exchanges unicast routing information.
3. Bring up the IPv4 PIM protocol. BGP exchanges IPv4 multicast routing information in order to build the PMSI infrastructure.
4. Bring up IGMP and add a static IGMP join where it is applicable. BGP exchanges IPv4 multicast routing information in order to propagate the multicast traffic to the receiver.
5. Bring up the IPv6 PIM protocol. BGP exchanges IPv6 multicast routing information in order to build the PMSI infrastructure.
6. Bring up MLD and add a static MLD join where it is applicable. BGP exchanges IPv6 multicast routing information in order to propagate the multicast traffic to the receiver.

The following BGP debug output is taken from PE-3 and demonstrates the message flow between PE-1 and PE-3.

The PMSI part is present in debug messages sent by PE-1 (**sender-only**).

Step 1. Bring down the VPRN service and protocols to clear the state of all signaling protocols.

```
# on PE-3
configure
    service
        vprn 1
            shutdown
            pim shutdown
            pim ipv6-multicast-disable
            igmp shutdown
            mld shutdown
        exit
    exit
```

Step 2. Enable the VPRN service on PE-3. PE-3 immediately receives Intra-AD messages from PE-1 because the remote VPRN service is already enabled for IPv4 and IPv6 multicast propagation. The PMSI attribute is present in both messages.

```
*A:PE-3# configure service vprn 1 no shutdown

9 2017/10/02 13:46:23.126 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 91
    Flag: 0x90 Type: 14 Len: 23 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.1
        Type: Intra-AD Len: 12 RD: 64500:101 Orig: 192.0.2.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64500:1
    Flag: 0xc0 Type: 22 Len: 22 PMSI:
        Tunnel-type LDP P2MP LSP (2)
        Flags: (0x0)[Type: None BM: 0 U: 0 Leaf: not required]
        MPLS Label 0
        Root-Node 192.0.2.1, LSP-ID 0x2001
"

10 2017/10/02 13:46:23.126 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 91
    Flag: 0x90 Type: 14 Len: 23 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV6
        NextHop len 4 NextHop 192.0.2.1
        Type: Intra-AD Len: 12 RD: 64500:101 Orig: 192.0.2.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64500:1
    Flag: 0xc0 Type: 22 Len: 22 PMSI:
        Tunnel-type LDP P2MP LSP (2)
        Flags: (0x0)[Type: None BM: 0 U: 0 Leaf: not required]
        MPLS Label 0
        Root-Node 192.0.2.1, LSP-ID 0x2001
"
```

Step 3. Enable PIM IPv4 only for the service on PE-3. PE-3 immediately sends Intra-AD messages to PE-1.

```
*A:PE-3# configure service vprn 1 pim no shutdown
```

```
6 2017/10/02 13:46:22.257 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 91
    Flag: 0x90 Type: 14 Len: 23 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.3
        Type: Intra-AD Len: 12 RD: 64500:103 Orig: 192.0.2.3
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64500:1
    Flag: 0xc0 Type: 22 Len: 22 PMSI:
        Tunnel-type LDP P2MP LSP (2)
        Flags: (0x0)[Type: None BM: 0 U: 0 Leaf: not required]
        MPLS Label 0
        Root-Node 192.0.2.3, LSP-ID 0x2001
"
```

Step 4. Bring up IGMP and add a static IGMP join for the service on a PE-3. PE-3 immediately sends a source-join message to PE-1 and receives a source-AD message from PE-1.

```
*A:PE-3# configure service vprn 1 igmp no shutdown
```

```
*A:PE-3# configure service vprn 1 igmp interface "int-PE-3-H-3"
static group 232.0.0.1 source 172.16.1.1
```

```
18 2017/10/02 13:46:33.123 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 76
    Flag: 0x90 Type: 14 Len: 33 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.3
        Type: Source-Join Len:22 RD: 64500:101 SrcAS: 64500
                          Src: 172.16.1.1 Grp: 232.0.0.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:192.0.2.1:2
"
```

Step 5. Enable PIM IPv6 for the service on PE-3. PE-3 immediately sends Intra-AD messages to PE-1.

```
*A:PE-3# configure service vprn 1 pim no ipv6-multicast-disable
```

```
20 2017/10/02 13:46:41.334 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 91
    Flag: 0x90 Type: 14 Len: 23 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV6
        NextHop len 4 NextHop 192.0.2.3
        Type: Intra-AD Len: 12 RD: 64500:103 Orig: 192.0.2.3
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64500:1
    Flag: 0xc0 Type: 22 Len: 22 PMSI:
        Tunnel-type LDP P2MP LSP (2)
        Flags: (0x0)[Type: None BM: 0 U: 0 Leaf: not required]
        MPLS Label 0
        Root-Node 192.0.2.3, LSP-ID 0x2001
"
```

Step 6. Bring up MLD and add a static MLD join for the service on a PE-3. PE-3 immediately sends a source-join message to PE-1.

```
*A:PE-3# configure service vprn 1 mld no shutdown
```

```
*A:PE-3# configure service vprn 1 mld interface "int-PE-3-H-3" static
group ff3e::8000:1 source 2001:db8:1::1
```

```
22 2017/10/02 13:47:00.145 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 100
    Flag: 0x90 Type: 14 Len: 57 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV6
        NextHop len 4 NextHop 192.0.2.3
        Type: Source-Join Len:46 RD: 64500:101 SrcAS: 64500
                          Src: 2001:db8:1::1 Grp: ff3e::8000:1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:192.0.2.1:2
"
```

The same information can be gathered using the following show commands.

show router bgp neighbor <*peer*> advertised-routes [**mvpn-ipv4**|**mvpn-ipv6**]

show router bgp neighbor <*peer*> received-routes [**mvpn-ipv4**|**mvpn-ipv6**]

PE-3 output for the advertised routes for the mvpn-ipv4 address family is as follows:

```
*A:PE-3# show router bgp neighbor 192.0.2.1 advertised-routes mvpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.3        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
Flag  RouteType                   OriginatorIP           LocalPref   MED
      RD                          SourceAS               Path-Id     Label
      Nexthop                     SourceIP
      As-Path                     GroupIP
-------------------------------------------------------------------------------
i     Source-Join                 -                      100         0
      64500:101                   64500                  None        -
      192.0.2.3                   172.16.1.1
      No As-Path                  232.0.0.1
i     Intra-Ad                    192.0.2.3              100         0
      64500:103                   -                      None        -
      192.0.2.3                   -
      No As-Path                  -
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-3#
```

PE-3 output for the advertised routes for the mvpn-ipv6 address family is as follows:

```
*A:PE-3# show router bgp neighbor 192.0.2.1 advertised-routes mvpn-ipv6
===============================================================================
 BGP Router ID:192.0.2.3        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv6 Routes
```

```
===============================================================================
Flag  RouteType                   OriginatorIP           LocalPref   MED
      RD                          SourceAS               Path-Id     Label
      Nexthop                     SourceIP
      As-Path                     GroupIP
-------------------------------------------------------------------------------
i     Source-Join                 -                      100         0
      64500:101                   64500                  None        -
      192.0.2.3                   2001:db8:1::1
      No As-Path                  ff3e::8000:1
i     Intra-Ad                    192.0.2.3              100         0
      64500:103                   -                      None        -
      192.0.2.3                   -
      No As-Path                  -
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-3#
```

PE-3 output for the received routes for the mvpn-ipv4 address family is as follows:

```
*A:PE-3# show router bgp neighbor 192.0.2.1 received-routes mvpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.3        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
Flag  RouteType                   OriginatorIP           LocalPref   MED
      RD                          SourceAS               Path-Id     Label
      Nexthop                     SourceIP
      As-Path                     GroupIP
-------------------------------------------------------------------------------
u*>i  Intra-Ad                    192.0.2.1              100         0
      64500:101                   -                      None        -
      192.0.2.1                   -
      No As-Path                  -
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-3#
```

PE-3 output for the received routes for the mvpn-ipv6 address family is as follows:

```
*A:PE-3# show router bgp neighbor 192.0.2.1 received-routes mvpn-ipv6
===============================================================================
 BGP Router ID:192.0.2.3        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
```

```
===============================================================================
BGP MVPN-IPv6 Routes
===============================================================================
Flag  RouteType                   OriginatorIP           LocalPref   MED
      RD                          SourceAS               Path-Id     Label
      Nexthop                     SourceIP
      As-Path                     GroupIP
-------------------------------------------------------------------------------
u*>i  Intra-Ad                    192.0.2.1              100         0
      64500:101                   -                      None        -
      192.0.2.1                   -
      No As-Path                  -
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-3#
```

LDP Verification and Debugging

When BGP intra-AD messages are exchanged, every PE starts to build a multicast tunnel based on the following criteria:

PE nodes which are configured as **sender-only** do not distribute mLDP forward equivalence classes (FECs) to remote PEs for this MVPN.

PE nodes which are configured as receiver-only do not include the PMSI part for intra-AD messages and remote PEs do not send mLDP FECs for this MVPN.

LDP bindings can be verified using the following command:

show router ldp bindings p2mp

PE-1 (192.0.2.1) has two egress FECs due to the fact that PE-1 has the mdt-type sender-only.

```
*A:PE-1# show router ldp bindings p2mp ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID 2001:db8::1)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
===============================================================================
LDP Generic IPv4 P2MP Bindings
===============================================================================
P2MP-Id
```

```
RootAddr                                    Interface       IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
Peer
--------------------------------------------------------------------------------
8193
192.0.2.1                                   73729             --      262136
192.168.12.2                                1/1/1
192.0.2.2:0

8193
192.0.2.1                                   73729             --      262136
192.168.13.2                                1/1/2
192.0.2.3:0

--------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Bindings: 2
================================================================================
--- snipped ---
================================================================================
*A:PE-1#
```

PE-2 (192.0.2.2) has two ingress FECs due to the fact that PE-2 has mdt-type receiver-only.

```
*A:PE-2# show router ldp bindings p2mp ipv4

================================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID 2001:db8::2)
================================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
================================================================================
LDP Generic IPv4 P2MP Bindings
================================================================================
P2MP-Id
RootAddr                                    Interface       IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
Peer
--------------------------------------------------------------------------------
8193
192.0.2.1                                   73732           262136U     --
  --                                          --
192.0.2.1:0

8193
192.0.2.3                                   73734           262135U     --
  --                                          --
192.0.2.3:0

--------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Bindings: 2
================================================================================
 --- snipped ---
```

```
===============================================================================
*A:PE-2#
```

PE-3 (192.0.2.3) has one ingress FEC and one egress FECs due to the fact that PE-3 has the default mdt-type sender-receiver. There is only an egress FEC to PE-2 (receiver-only), but not to PE-1. PE-1 can never be a receiver, since it is configured as sender-only.

```
*A:PE-3# show router ldp bindings p2mp ipv4 opaque-type generic

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.3)
             (IPv6 LSR ID 2001:db8::3)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
===============================================================================
LDP Generic IPv4 P2MP Bindings
===============================================================================
P2MP-Id
RootAddr                                    Interface       IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
Peer
-------------------------------------------------------------------------------
8193
192.0.2.1                                   73733           262136U     --
  --                                          --
192.0.2.1:0

8193
192.0.2.3                                   73732             --      262135
192.168.23.1                                1/1/2
192.0.2.2:0

-------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Bindings: 2
===============================================================================
*A:PE-3#
```

Multicast Stream Verification

The status of a multicast group/stream can be verified using the following command:

show router <*sid*> pim group detail [**ipv6**]

An IPv4 sender is connected to PE-1. The physical interface where the source is connected is used as incoming interface and the I-PMSI is used as outgoing interface.

```
*A:PE-1# show router 1 pim group detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 232.0.0.1
Source Address     : 172.16.1.1
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 172.16.1.1
MRIB Src Flags     : direct
Keepalive Timer    : Not Running
Up Time            : 0d 00:05:51        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:00
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 172.16.1.1
Incoming Intf      : int-PE-1-S-1
Outgoing Intf List : mpls-if-73729

Curr Fwding Rate   : 1066.6 kbps
Forwarded Packets  : 31272              Discarded Packets  : 0
Forwarded Octets   : 46845456           RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-1#
```

There is an IPv4 receiver connected to PE-2. The I-PMSI is used as incoming interface and the physical interface where the receiver is connected is used as outgoing.

```
*A:PE-2# show router 1 pim group detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 232.0.0.1
Source Address     : 172.16.1.1
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 192.0.2.1
MRIB Src Flags     : remote
Keepalive Timer    : Not Running
Up Time            : 0d 00:04:12        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:47
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00
```

```
Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 192.0.2.1
Incoming Intf      : mpls-if-73732
Outgoing Intf List : int-PE-2-H-2

Curr Fwding Rate   : 1066.6 kbps
Forwarded Packets  : 22474             Discarded Packets  : 0
Forwarded Octets   : 33666052          RPF Mismatches     : 0
Spt threshold      : 0 kbps            ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-2#
```

There is IPv4 receiver connected to PE-3. The I-PMSI is used as incoming interface and the physical interface where the receiver is connected is used as outgoing.

```
*A:PE-3# show router 1 pim group detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 232.0.0.1
Source Address     : 172.16.1.1
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              :                   Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 192.0.2.1
MRIB Src Flags     : remote
Keepalive Timer    : Not Running
Up Time            : 0d 00:02:18       Resolved By        : rtable-u

Up JP State        : Joined            Up JP Expiry       : 0d 00:00:41
Up JP Rpt          : Not Joined StarG  Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 192.0.2.1
Incoming Intf      : mpls-if-73733
Outgoing Intf List : int-PE-3-H-3

Curr Fwding Rate   : 1072.6 kbps
Forwarded Packets  : 12354             Discarded Packets  : 0
Forwarded Octets   : 18506292          RPF Mismatches     : 0
Spt threshold      : 0 kbps            ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-3#
```

Similar behavior is observed for IPv6 multicast.

There is an IPv6 sender connected to PE-1. The physical interface where the sender is connected is used as the incoming interface and the I-PMSI is used as the outgoing interface.

```
*A:PE-1# show router 1 pim group detail ipv6

===============================================================================
PIM Source Group ipv6
===============================================================================
Group Address      : ff3e::8000:1
Source Address     : 2001:db8:1::1
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 2001:db8:1::1
MRIB Src Flags     : direct
Keepalive Timer    : Not Running
Up Time            : 0d 00:05:51        Resolved By        : rtable6-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:00
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 2001:db8:1::1
Incoming Intf      : int-PE-1-S-1
Outgoing Intf List : mpls-if-73729

--- snipped ---

-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-1#
```

There is an IPv6 receiver connected to PE-2. An I-PMSI is used as the incoming interface and the physical interface where the receiver is connected is used as the outgoing interface.

```
*A:PE-2# show router 1 pim group detail ipv6
===============================================================================
PIM Source Group ipv6
===============================================================================
Group Address      : ff3e::8000:1
Source Address     : 2001:db8:1::1
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 192.0.2.1
MRIB Src Flags     : remote
Keepalive Timer    : Not Running
```

```
Up Time            : 0d 00:03:37        Resolved By        : rtable6-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:22
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 192.0.2.1
Incoming Intf      : mpls-if-73732
Outgoing Intf List : int-PE-2-H-2

--- snipped ---

-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-2#
```

There is an IPv6 receiver connected to PE-3. An I-PMSI is used as the incoming interface and the physical interface where the receiver is connected is used as the outgoing interface.

```
*A:PE-3# show router 1 pim group detail ipv6

===============================================================================
PIM Source Group ipv6
===============================================================================
Group Address      : ff3e::8000:1
Source Address     : 2001:db8:1::1
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 192.0.2.1
MRIB Src Flags     : remote
Keepalive Timer    : Not Running
Up Time            : 0d 00:01:51        Resolved By        : rtable6-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:08
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 192.0.2.1
Incoming Intf      : mpls-if-73733
Outgoing Intf List : int-PE-3-H-3

--- snipped ---

-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-3#
```

Conclusion

The sender-only/receiver-only feature provides significant signaling optimization in the core network for RSVP and LDP protocols and is recommended to be used when such functionality is required. The following are required before implementing this feature in the network:

- MDT-types **sender-only**, **receiver-only** and **sender-receiver** are enabled per MVPN.
- The default mdt-type is **sender-receiver** mode for backward compatibility.
- This is purely a control plane feature and there are no hardware dependencies.
- Rosen MPVN or MDT-SAFI based MVPNs are not supported.
- IPv4 and IPv6 C-signaling are supported.
- mLDP and RSVP demonstrate slightly different behavior due to the nature of each protocol.
- mLDP provides a better optimization than RSVP in all cases, as mLDP does not initiate LSPs to sender-only routers.

NG-MVPN Source Redundancy

This chapter provides information about MVPN source redundancy.

Topics in this chapter include:

- Applicability
- Summary
- Overview
- Configuration
- Conclusion

Applicability

The chapter was initially written for release 12.0.R1, using multicast LDP as the provider tunnel signaling mechanism for IPv4 multi-casting. The customer multicast signaling protocol within the VPN must be BGP. The CLI in the current edition corresponds to 15.0.R5.

Summary

Multicast source redundancy allows operators to provide multiple geo-redundant sources for the same multicast group in a multicast virtual private network (MVPN). For instance, in an IPTV environment where a TV channel maps to a multicast group, the same TV channel can be provided from sources in a geographically diverse manner where a national broadcaster can have multiple sources from two or more regional distribution centers.

Knowledge of Multi-Protocol BGP (MP-BGP) and RFC 4364, *BGP/MPLS IP Virtual Private Networks (VPNs)*, is assumed throughout this chapter, as well as Protocol Independent Multicast (PIM), RFC 6513, *Multicast in MPLS/BGP IP VPNs*, and RFC 6514, *BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs*.

Overview

Hosts connected to receiver PEs can receive TV channels from a specific source, with a regional backup source available in case of a failure.

Figure 179 **Source Redundancy Example.**

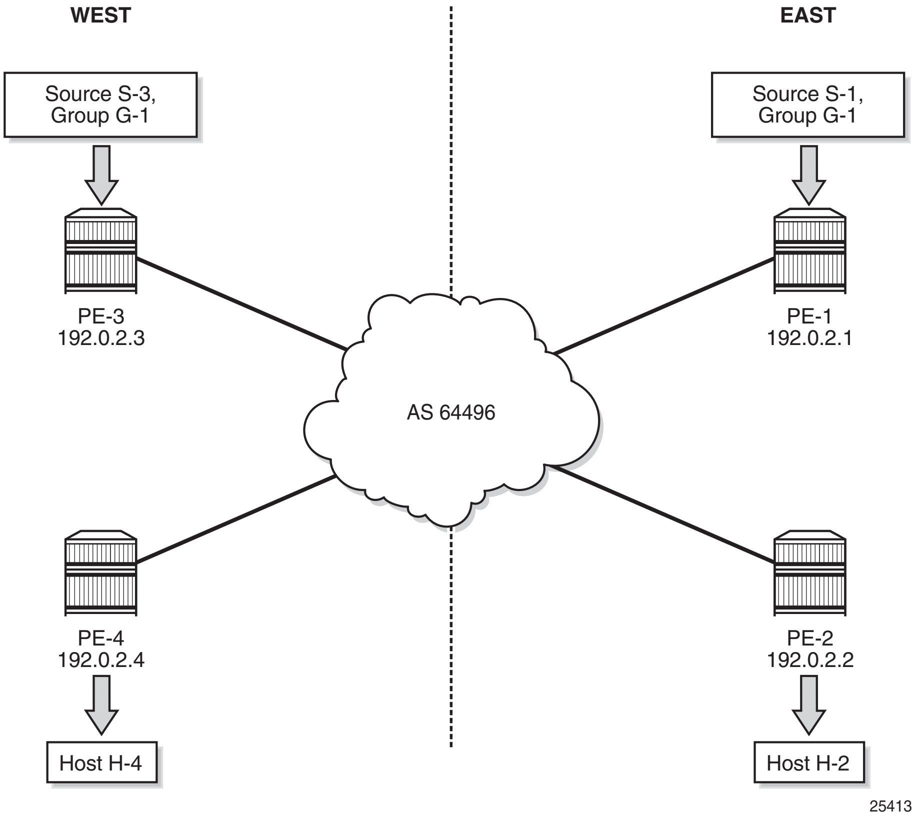


Figure 179 shows the concept of source redundancy. PE-1 and PE-3 have directly connected multicast sources. For clarity, consider a single multicast group with two separate sources connected at different sites. The content of each group is identical at a given time (allowing for transmission delay), as is expected for an IPTV channel. PE-1 and PE-3 are referred to as sender PEs because they are closer to the source; PE-2 and PE-4 are referred to as receiver PEs because they are closer to hosts H-2 and H-4.

A multicast group, group 1 (G-1) has two sources: Source 1 (S-1) in the east region and Source 3 (S-3) in the west region which are connected to PE-1 and PE-3 respectively. Receivers connected to PE-2 in the east region will join group (S-1,G-1) and receivers connected to PE-4 in the west region will join group (S-3,G-1). The presence of each source is declared within the multicast VPN by the sender PE. When a multicast group becomes active, a BGP Source Active auto-discovery (SA) route is advertised to all PEs within the multicast VPN. This must occur even if no receiver indicates that it wishes to become a member of this group. In other words, the SA must be persistently present in the receiver PEs when the source is available.

Should either source fail or become unavailable, then the sender PE will notify the receiver PEs by sending an NLRI unreachable BGP SA Route that declares the absence of the source. All hosts that are members of this group will then switch to receive traffic from the remaining active source. Only customer multicast joins received as IGMP (*,G) queries or PIM (*,G) joins at the receiver PE are valid, because the source address is not specified.

Source redundancy is achieved by:

- Configuring a list of redundant sources within each receiver PE.
- Configuring the sender PEs to originate a BGP Source Active Auto Discovery for each detected active multicast source, regardless of whether a receiver is joined to the multicast group or not. As a result, a Source Active route is originated on a per (S,G) basis.

  For multiple SAs to be persistently present in the receiver PEs, one of the following two conditions must be configured within the sender PEs:

  – Either disable inter-site shared trees on the sender PEs, such that there is no c-tree with root at the RP. Any active source will announce its presence using a BGP SA to all receiver PEs so no shared joins are sent by receiver PEs to RP, or
  – Leave inter-site shared trees as enabled, but configured so that the SA AD route for each multicast group is persistently present in the receiver PEs, even in the absence of requesting hosts for each group. Shared and Source Joins are sent by the receiver PEs.

  Both of the preceding options are supported. The default behavior has inter-site shared trees enabled without persistency. In this example, inter-site shared trees at the sender PEs are enabled with Source Active routes set to be persistent.
- Ensuring that the preferred source is IP reachable within the VPRN from the receiver PE. This must be a remote source advertised from a remote PE within the VPRN.
- Receiver PEs will accept the Source Active route(s) into the appropriate Multicast VRF.

- Ensuring the preferred active source should have a higher BGP Local Preference. This is achieved using a route policy. Any other sources from the redundant list should exist as suppressed standby sources, but the (S,G) state should exist if the source is active – when a valid BGP MVPN Source Active route for that source has been received.

All of these conditions are achieved by configuration.

In order to allow each receiver PE to choose a preferred source, each SA route advertised by the sender PE will be tagged with a community value. Each receiver PE can then use the community value contained within each SA route update received to set the Local Preference BGP attribute to a value such that the receiver PE can choose the most preferred active source.

The objectives are:

- To configure multicast in a VPRN on PE-1 to PE-4 with inter-site-shared trees enabled on the receiver PEs and Source Active routes persistently present, for reasons previously described.
- To connect redundant sources to the sender PE-1 and PE-3, with each multicast source having the same group address. For ease of configuration, a single redundant source is used.
- To advertise each source to the receiver PEs (PE-2 and PE-4), using appropriate route policies for adding community strings to the BGP Source Active Auto-Discovery routes.
- To configure appropriate route policies that allow each BGP SA route to have the correct Local Preference set, based on the community strings present.
- To allow receivers to connect to the appropriate source, using (*,G) joins.

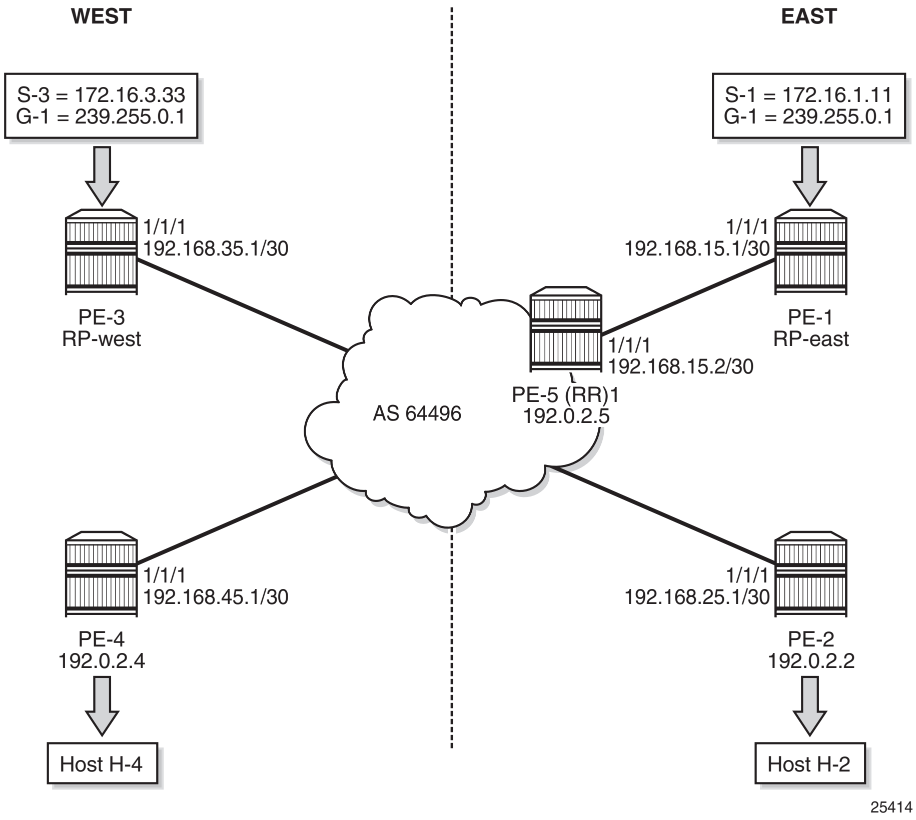
The following configuration tasks should be completed as a pre-requisite:

- Full mesh IS-IS or OSPF between each of the PE routers and the route reflector.
- Link-layer LDP between all PEs. RSVP could also be used.
- Multicast LDP is used as the provider tunnel signaling protocol. This is enabled by default when link layer LDP is enabled. RSVP and PIM SSM are also supported as provider tunnel signaling mechanisms and could be used.

Configuration

The example topology is shown in Figure 180, containing the four PEs plus the route reflector at P-5.

Figure 180 **Schematic Topology**



Global BGP Configuration

The first step is to configure an iBGP session between each of the PEs and the route reflector (RR) seen in Figure 180. The address families negotiated between the iBGP peers are vpn-ipv4 (unicast routing) and mvpn-ipv4 (multicast routing).The BGP configuration for all PE nodes is identical:

```
# on PE-1
configure
    router
        bgp
            group "INTERNAL"
                family vpn-ipv4 mvpn-ipv4
                type internal
                neighbor 192.0.2.5
                exit
            exit
            no shutdown
        exit
```

The configuration for the Route Reflector at P-5 is:

```
# on P-5
configure
    router
        bgp
            group "RRclients"
                family vpn-ipv4 mvpn-ipv4
                type internal
                cluster 1.1.1.1
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
            exit
            no shutdown
        exit
```

On PE-1, verify that the BGP session with RR at P-5 is established with address families “vpn-ipv4” and “mvpn-ipv4” capabilities negotiated:

```
*A:PE-1# show router bgp summary
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
BGP Admin State         : Up          BGP Oper State              : Up
Total Peer Groups       : 1           Total Peers                 : 1
Total VPN Peer Groups   : 0           Total VPN Peers             : 0
Total BGP Paths         : 12          Total Path Memory           : 3168

--- snipped ---

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
                   AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.5
               64496        3    0 00h00m18s 0/0/0 (VpnIPv4)
                            3    0           0/0/0 (MvpnIPv4)
-------------------------------------------------------------------------------
*A:PE-1#
```

The same command can be used on the other PEs to verify their BGP sessions to the RR.

Configuring VPRN on PEs

The following outputs show the VPRN configurations for each PE. The specific MVPN configuration is shown later.

PE-1

The VPRN configuration for PE-1 is as follows:

```
# on PE-1
configure
    service
        vprn 1 customer 1 create
            route-distinguisher 64496:1
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64496:1
            interface "int-PE-1-S-1" create
                address 172.16.1.1/24
                sap 1/1/3 create
                exit
            exit
            interface "RP" create
                address 10.10.10.1/32
                loopback
            exit
            pim
                apply-to all
                rp
                    static
                        address 10.10.10.1
                            group-prefix 239.0.0.0/8
                        exit
                    exit
                exit
                no shutdown
            exit
        exit
```

There is a single interface toward S-1 from which the multicast group is generated.

If the customer signaling uses PIM ASM, then the customer Rendezvous Point (RP) must be positioned on the sender PE because registration of the source with the RP causes the SA to be sent to the remote source PEs.

A loopback interface called “**RP**” acts as the RP for all group prefixes in the 239.0.0.0/8 range. This will be the RP for the East groups.

PE-2

PE-2 has a receiver attached, and a single interface is configured to accommodate this. The RP configured is that of the East region and has a configuration as follows:

```
# on PE-2
configure service
        vprn 1 customer 1 create
            route-distinguisher 64496:1
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64496:1
            interface "int-PE-2-H-2" create
                address 172.16.2.1/24
                sap 1/1/3 create
                exit
            exit
            igmp
                interface "int-PE-2-H-2"
                exit
            exit
            pim
                apply-to all
                rp
                    static
                        address 10.10.10.1
                            group-prefix 239.0.0.0/8
                        exit
                    exit
                exit
                no shutdown
            exit
        exit
```

PE-3

PE-3 serves as the RP for the West region and uses a different IP address for the Rendezvous Point interface.

```
# on PE-3
configure service
        vprn 1 customer 1 create
            route-distinguisher 64496:1
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64496:1
            interface "int-PE-3-S-3" create
                address 172.16.3.1/24
                sap 1/1/3 create
                exit
            exit
            interface "RP" create
                address 10.10.10.3/32
                loopback
            exit
            pim
                apply-to all
                rp
                    static
                        address 10.10.10.3
                            group-prefix 239.0.0.0/8
                        exit
                    exit
                exit
                no shutdown
            exit
        exit
```

PE-4

PE-4 also has a receiver, and uses the West sender PE (PE-3) as the Rendezvous Point.

```
# on PE-4
configure service
        vprn 1 customer 1 create
            route-distinguisher 64496:1
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64496:1
            interface "int-PE-4-H-4" create
                address 172.16.2.1/24
                sap 1/1/3 create
                exit
            exit
            igmp
                interface "int-PE-4-H-4"
                exit
            exit
            pim
                apply-to all
                rp
                    static
                        address 10.10.10.3
                            group-prefix 239.0.0.0/8
                        exit
                    exit
                exit
                no shutdown
            exit
        exit
```

MVPN Configuration for Source PEs

At the PEs closest to the sources (PE-1 and PE-3), Source Active auto-discovery BGP routes are generated when the source is active.

This applies for PIM-ASM (*,G) joins only, or IGMP (*,G) membership queries received by the provider domain. These are received by all PEs.

Inter-site trees must be disabled for this to occur. Alternatively, inter-site trees can be enabled such that when a source is discovered, a Source Active is advertised to each other PE in the MVPN. This occurs regardless of whether any receivers wish to become members of the multicast groups.

As previously stated, the presence of the SA in the receiver PEs means that no shared joins routes are generated toward the C-RPs.

The MVPN configuration enables BGP as both auto-discovery mechanism and the customer multicast signaling protocol across the VPRN. The provider tunnel between PEs within the MVPN is signaled using Multicast LDP.

The MVPN configuration for each PE should be as follows:

```
# on PE-1
configure
    service
        vprn 1
            mvpn
                auto-discovery default
                c-mcast-signaling bgp
                provider-tunnel
                    inclusive
                        mldp
                            no shutdown
                        exit
                    exit
                exit
            exit
```

The VPRN MVPN configuration for PE-2, PE-3, and PE-4 is identical.

Sender PE Route Policies

The choice of active and standby sources by the receiver PEs is determined by the “best route” policy. PE-1 and PE-3 advertise BGP Source Active Auto Discovery routes when a source is active. This is received by all PEs within the MVPN. As two different sources advertise the same group, it is necessary to differentiate between them.

Assuming that receiver PE-2 prefers the source from PE-1, and PE-4 prefers the source active on PE-3, then the export policy for MVPN routes on PE-1 requires the following steps:

1. Set a community value at PE-1 for the (S,G) multicast group – call this “blue” with value 64496:11.
2. Set the route target community for the VPRN – 64496:1.
3. Create a policy statement that becomes the export policy for MVPN routes within PE-1.
4. Create a policy statement entry (entry 10) that adds the community value “blue” along with the route target for Source Active AD BGP routes. Source Active AD routes are MVPN type 5 routes.
5. Create a policy statement entry default-action that adds the route target for all other MVPN AD BGP routes (such as Intra-AD (type 1)) that are exported to the MVPN PEs.

```
# on PE-1
configure
    router
        policy-options
            begin
            community "blue" members "64496:11"
            community "MVPN1_RT" members "target:64496:1"
            policy-statement "MVPN1_export"
                entry 10
                    description "match MVPN routes - type 5 Source AD -
                                 add RT and 'blue' community"
                    from
                        mvpn-type 5
                        family mvpn-ipv4
                    exit
                    action accept
                        community add "blue" "MVPN1_RT"
                    exit
                exit
                default-action accept
                    community add "MVPN1_RT"
                exit
            exit
            commit
        exit
```

6. Apply as an export policy within the MVPN context.

The import policy requires that all imported MVPN BGP routes have the correct route target extended community value, specifically “target:64496:1”.

1. Create a policy statement that becomes the import policy for PE-1.
2. Create a policy statement entry (entry 10) that matches the community of the route target extended community for all MVPN BGP routes. These include the Intra-AD and Source-Join routes.

```
# on PE-1
configure
    router
        policy-options
            begin
                policy-statement "MVPN1_import"
                    entry 10
                        from
                            community "MVPN1_RT"
                        exit
                        action accept
                        exit
                    exit
                exit
            commit
        exit
```

Enable the inter-site-shared type 5 advertisement persistency so that source ADs are advertised when multicast sources are active. Alternatively, inter-site shared trees can be disabled using the **no intersite shared** command. In this example, only inter-site shared MVPN type 5 persistency is shown.

The additional configuration in the MVPN context is as follows, where the PIM instance must be shut down when the intersite-shared configuration is modified.

```
# on PE-1
configure
    service
        vprn 1
            mvpn
                intersite-shared persistent-type5-adv
                vrf-import "MVPN1_import"
                vrf-export "MVPN1_export"
            exit
```

For PE-3 (the other sender PE), similar import and export policies are required. In this case, the community will be called “red” and is added to the Source Active AD route generated when the source is active.

The requirements for the export policy for PE-3 are as follows:

```
# on PE-3
configure
    router
        policy-options
            begin
            community "red" members "64496:33"
            community "MVPN1_RT" members "target:64496:1"
            policy-statement "MVPN1_export"
                entry 10
                    description "match MVPN routes - type 5 Source AD -
                                 add RT and 'red' community"
                    from
                        mvpn-type 5
                        family mvpn-ipv4
                    exit
                    action accept
                        community add "red" "MVPN1_RT"
                    exit
                exit
                default-action accept
                    community add "MVPN1_RT"
                exit
            exit
            commit
        exit
```

The import policy is exactly the same as for PE-1.

Apply the import and export policies to the MVPN context of the sender PE (PE-3) and enable inter-site-shared type 5 advertisement persistency with the same command as on PE-1.

Receiver PE Configuration

PE-2 and PE-4 are the receiver PEs. These will receive the Source Active AD routes and initiate Joins toward the preferred source.

When a Source-Active AD route is received, the community value is examined and the Local Preference value of the route is set using a Route Policy. The preferred source is determined by the SA AD route with the highest Local Preference value.

In the case of PE-2, the preferred source is that advertised by PE-1, the “blue” source as previously referenced. PE-2 sets the Local Preference to 200. The SA AD tagged with the “red” community has the Local Preference set to 50.

For PE-4, the reverse applies: SA AD routes tagged with the “red” community have the Local Preference set to 200, and “blue” SA AD routes have the Local Preference set to 50.

Once again, assuming that the PE-2 receiver prefers the source from PE-1 and PE-4 prefers the source active on PE-3, the import policy for MVPN routes on PE-2 requires the following steps:

1. Set a community value at PE-2 for the (S,G), call this “blue” with value 64496:11.
2. Set the route target community for the VPRN to 64496:1.
3. Create a prefix list that matches the multicast group address, in this case 239.255.0.0/24.
4. Create a policy statement that becomes the import policy for MVPN routes within PE-1.

5. Create a policy statement entry (entry 10) that matches the following attributes:
   - Source Active AD BGP routes type. Source Active AD routes are classed as MVPN type 5 routes, and
   - Community value “blue” AND Route Target extended community, and
   - Group address prefix 239.255.0.0/24

   If the BGP route matches all three conditions, then set the Local Preference to 200.
6. Create a policy statement default-action that accepts all other MVPN BGP routes, including SA routes tagged with the “red” community value.

   The import policy statement looks like:

```
# on PE-2
configure
    router
        policy-options
            begin
            prefix-list "group_239.255.x.y"
                prefix 239.255.0.0/16 longer
            exit
            community "red" members "64496:33"
            community "blue" members "64496:11"
            community "MVPN1_RT" members "target:64496:1"
            policy-statement "MVPN1_import"
                entry 10
                    description "allow MVPN source-ad - set LP to 200 for 'blue'"
                    from
                        community expression "[blue] AND [MVPN1_RT]"
                        mvpn-type 5
                        group-address "group_239.255.x.y"
                    exit
                    action accept
                        local-preference 200
                    exit
                exit
                entry 20
                    description "allow MVPN source-ad - set LP to 50 for 'red'"
                    from
                        community expression "[red] AND [MVPN1_RT]"
                        mvpn-type 5
                        group-address "group_239.255.x.y"
                    exit
                    action accept
                        local-preference 50
                    exit
                exit
                default-action accept
                exit
            exit
            commit
```

The export policy for PE-2 MVPN routes requires each MVPN route to be tagged with the route target extended community for VPRN 1. The following policy statement is created:

```
# on PE-2
configure
    router
        policy-options
            begin
            policy-statement "MVPN1_export"
                entry 10
                    from
                        family mvpn-ipv4
                    exit
                    action accept
                        community add "MVPN1_RT"
                    exit
                exit
            exit
            commit
```

7. Create a list of redundant sources. This is a list of prefixes that match the source addresses of redundant multicast groups. This is an important parameter because the receiver PEs only create active and standby (S,G) states for groups with source address prefixes that are contained in this list.
8. Before any hosts attempt to join the multicast groups, the decision must be made to enable or disable inter-site shared trees at the receiver PEs. In this example, only the Inter-site shared trees disabled option will be considered. In order to make this configuration change, it is necessary to shut the PIM protocol down before and re-enable when completed.

The additional MVPN configuration for PE-2 is shown in the following output, where the redundant source prefix list is included, and inter-site shared trees are disabled.

```
# on PE-2
configure
    service
        vprn 1
            mvpn
                no intersite-shared
                red-source-list
                    src-prefix 172.16.1.0/24
                    src-prefix 172.16.3.0/24
                exit
                vrf-import "MVPN1_import"
                vrf-export "MVPN1_export"
            exit
```

PE-4 requires a similar set of import and export policies. In this case, the “red” sources have the highest Local Preference value, based on the community string added by the export policy of PE-3.

```
# on PE-4
configure
    router
        policy-options
            begin
            prefix-list "group_239.255.x.y"
                prefix 239.255.0.0/16 longer
            exit
            community "red" members "64496:33"
            community "blue" members "64496:11"
            community "MVPN1_RT" members "target:64496:1"
            policy-statement "MVPN1_import"
                entry 10
                    description "allow MVPN source-ad - set LP to 200 for 'red'"
                    from
                        community expression "[red] AND [MVPN1_RT]"
                        mvpn-type 5
                        group-address "group_239.255.x.y"
                    exit
                    action accept
                        local-preference 200
                    exit
                exit
                entry 20
                    description "allow MVPN source-ad - set LP to 50 for 'blue'"
                    from
                        community expression "[blue] AND [MVPN1_RT]"
                        mvpn-type 5
                        group-address "group_239.255.x.y"
                    exit
                    action accept
                        local-preference 50
                    exit
                exit
                default-action accept
                exit
            exit
            commit
```

The export policy for MVPN routes adds the route target extended community. It is exactly the same export policy as for PE-2.

The additional MVPN configuration for VPRN 1 on PE-4 is identically the same as for PE-2.

Each PE within the MVPN originates an Intra-AD BGP route. This notifies the other PEs within the VPRN. This is used to create a set of Inclusive Provider Multicast Service Interfaces (I-PMSI) between each PE. In this case, I-PMSIs are signaled using mLDP.

Using PE-1 as an example, the set of Intra-AD routes can be seen using the following command:

```
*A:PE-1# show router bgp routes mvpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
Flag  RouteType                   OriginatorIP           LocalPref   MED
      RD                          SourceAS               Path-Id     Label
      Nexthop                     SourceIP
      As-Path                     GroupIP
-------------------------------------------------------------------------------
i     Intra-Ad                    192.0.2.1              100         0
      64496:1                     -                      None        -
      192.0.2.1                   -
      No As-Path                  -
u*>i  Intra-Ad                    192.0.2.2              100         0
      64496:1                     -                      None        -
      192.0.2.2                   -
      No As-Path                  -
u*>i  Intra-Ad                    192.0.2.3              100         0
      64496:1                     -                      None        -
      192.0.2.3                   -
      No As-Path                  -
u*>i  Intra-Ad                    192.0.2.4              100         0
      64496:1                     -                      None        -
      192.0.2.4                   -
      No As-Path                  -
-------------------------------------------------------------------------------
Routes : 4
===============================================================================
*A:PE-1#
```

At this moment, there are no connected sources detected and no receivers wishing to join any multicast sources.

Each I-PMSI is seen as a PIM tunnel interface. As there are four routers in the MVPN, there are four I-PMSIs.

```
*A:PE-1# show router 1 pim tunnel-interface

===============================================================================
PIM Interfaces ipv4
===============================================================================
Interface                        Originator Address   Adm  Opr  Transport Type
-------------------------------------------------------------------------------
mpls-if-73729                    192.0.2.1            Up   Up   Tx-IPMSI
mpls-if-73730                    192.0.2.3            Up   Up   Rx-IPMSI
mpls-if-73731                    192.0.2.2            Up   Up   Rx-IPMSI
mpls-if-73732                    192.0.2.4            Up   Up   Rx-IPMSI
-------------------------------------------------------------------------------
Interfaces : 4
===============================================================================
*A:PE-1#
```

In order to be able to reach the source, a route for each source is included in the VRF for VPRN 1.

For PE-2, this looks as follows:

```
*A:PE-2# show router 1 route-table

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                  Metric
-------------------------------------------------------------------------------
10.10.10.1/32                                 Remote  BGP VPN   00h04m38s  170
       192.0.2.1 (tunneled)                                         0
10.10.10.3/32                                 Remote  BGP VPN   00h04m38s  170
       192.0.2.3 (tunneled)                                         0
172.16.1.0/24                                 Remote  BGP VPN   00h04m38s  170
       192.0.2.1 (tunneled)                                         0
172.16.2.0/24                                 Local   Local     00h07m00s  0
       int-PE-2-H-2                                                 0
172.16.3.0/24                                 Remote  BGP VPN   00h04m38s  170
       192.0.2.3 (tunneled)                                         0
172.16.4.0/24                                 Remote  BGP VPN   00h03m08s  170
       192.0.2.4 (tunneled)                                         0
-------------------------------------------------------------------------------
No. of Routes: 6
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-2#
```

The sources at 172.16.1.0/24 and 172.16.3.0/24 are learned as BGP VPN routes.

The following output shows the BGP routes for these prefixes, for example for prefix 172.16.1.0/24 on PE-2:

```
*A:PE-2# show router bgp routes 172.16.1.0/24 vpn-ipv4 hunt
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : 172.16.1.0/24
Nexthop        : 192.0.2.1
Route Dist.    : 64496:1                VPN Label      : 262138
Path Id        : None
From           : 192.0.2.5
Res. Nexthop   : n/a
```

```
Local Pref.    : 100                    Interface Name : int-PE-2-P-5
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : target:64496:1 l2-vpn/vrf-imp:192.0.2.1:2
                 source-as:64496:0
Cluster        : 1.1.1.1
Originator Id  : 192.0.2.1              Peer Router Id : 192.0.2.5
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h08m43s
VPRN Imported  :  1

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-2#
```

This prefix is advertised with three communities:

- A route target extended community
- An l2-vpn/vrf-import extended community.
- A source-AS extended community (not used in Intra-AS context).

The l2-vpn/vrf-import extended community is significant as it is a unique value. It represents a specific MVPN on a specific PE and is comprised of a 32 bit value that identifies the PE plus an index identifying the VRF. The 32 bit value is the system address. The index (3) can be derived from the command:

```
*A:PE-2# admin display-config index | match vprn1
        virtual-router "vprn1" 2 0
*A:PE-2#
```

Therefore, the l2-vpn/vrf-import community for VPRN 1 on PE-1 is 192.0.2.1:2

This community attribute is included within the source-join BGP route that is sent in a BGP update by a receiver PE as it tries to join a multicast group with a source address that matches the 172.16.1.0/24 prefix. This ensures that the source-join route is only accepted as a valid route and imported by the PE that originated the source address prefix. This is explained in the following section.

Enable Redundant Sources

The redundant sources are now enabled so that multicast traffic flows into both PE-1 and PE-3, using groups (S-1,G-1) and (S-3,G-1), respectively.

On each of these PEs, a source active AD route is generated. By examining each receiver PE, these can be clearly seen.

For PE-2, the source active AD routes can be seen using the following command.

```
*A:PE-2# show router bgp routes mvpn-ipv4 type source-ad
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
Flag  RouteType                   OriginatorIP           LocalPref   MED
      RD                          SourceAS               Path-Id     Label
      Nexthop                     SourceIP
      As-Path                     GroupIP
-------------------------------------------------------------------------------
u*>i  Source-Ad                   -                      200         0
      64496:1                     -                      None        -
      192.0.2.1                   172.16.1.11
      No As-Path                  239.255.0.1
u*>i  Source-Ad                   -                      50          0
      64496:1                     -                      None        -
      192.0.2.3                   172.16.3.33
      No As-Path                  239.255.0.1
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-2#
```

There are two routes present, one from each source for the same group from PE-1 and PE-3.

The PIM groups can now be seen on PE-1 as follows:

```
*A:PE-1# show router 1 pim group

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type               Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                 State    Inc Intf(S)
-------------------------------------------------------------------------------
```

```
239.255.0.1                 (S,G)                        int-PE-1-S-1   0
   172.16.1.11                 10.10.10.1
239.255.0.1                 (S,G)                        mpls-if-73730  0
   172.16.3.33                 10.10.10.1
-------------------------------------------------------------------------------
Groups : 2
===============================================================================
*A:PE-1#
```

There are two groups at PE-1. In addition to its locally connected source, PE-1 has also received a source active from PE-3 which has an incoming interface of the I-PMSI toward PE-3. The outgoing interface list is empty as there is no host wishing to become a group member.

Similarly, on the other sender, PE-3.

```
*A:PE-3# show router 1 pim group

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
239.255.0.1                 (S,G)                        mpls-if-73730  0
   172.16.1.11                 10.10.10.3
239.255.0.1                 (S,G)                        int-PE-3-S-3   0
   172.16.3.33                 10.10.10.3
-------------------------------------------------------------------------------
Groups : 2
===============================================================================
*A:PE-3#
```

By examining the receiver PE-2, it can be seen that the Source AD route for (S,G) (172.16.1.2, 239.255.0.1) from PE-1 has a higher local preference so it is chosen as the preferred (active) source. Examining these routes in more detail shows that each route is tagged with two communities: the route target extended community and the “red” or “blue” community, as seen in the following output.

```
*A:PE-2# show router bgp routes mvpn-ipv4 type source-ad hunt
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
```

```
-------------------------------------------------------------------------------
Route Type     : Source-Ad
Route Dist.    : 64496:1
Source IP      : 172.16.1.11
Group IP       : 239.255.0.1
Nexthop        : 192.0.2.1
Path Id        : None
From           : 192.0.2.5
Res. Nexthop   : 0.0.0.0
Local Pref.    : 200                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : 64496:11 no-export target:64496:1
Cluster        : 1.1.1.1
Originator Id  : 192.0.2.1              Peer Router Id : 192.0.2.5
Flags          : Used  Valid  Best  IGP
--- snipped ---
VPRN Imported  :  1

Route Type     : Source-Ad
Route Dist.    : 64496:1
Source IP      : 172.16.3.33
Group IP       : 239.255.0.1
Nexthop        : 192.0.2.3
Path Id        : None
From           : 192.0.2.5
Res. Nexthop   : 0.0.0.0
Local Pref.    : 50                     Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : 64496:33 no-export target:64496:1
Cluster        : 1.1.1.1
Originator Id  : 192.0.2.3              Peer Router Id : 192.0.2.5
Flags          : Used  Valid  Best  IGP
--- snipped ---
VPRN Imported  :  1

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-2#
```

The local preference is set based on these community values.

A debug of the received BGP Source AD routes is as follows for PE-2:

```
1 2017/10/12 09:54:47.907 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.5
"Peer 1: 192.0.2.5: UPDATE
Peer 1: 192.0.2.5 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 90
    Flag: 0x90 Type: 14 Len: 29 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.1
        Type: Source-AD Len: 18 RD: 64496:1 Src: 172.16.1.11 Grp: 239.255.0.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 8 Community:
        64496:11
        no-export
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.1
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        1.1.1.1
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64496:1
"
2 2017/10/12 09:55:51.907 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.5
"Peer 1: 192.0.2.5: UPDATE
Peer 1: 192.0.2.5 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 90
    Flag: 0x90 Type: 14 Len: 29 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.3
        Type: Source-AD Len: 18 RD: 64496:1 Src: 172.16.3.33 Grp: 239.255.0.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 8 Community:
        64496:33
        no-export
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.3
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        1.1.1.1
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:64496:1
"
```

Similarly, the Source Active routes on receiver PE-4 show that the highest local preference value of 200 is set for the SA route received from PE-3 with an originator ID of 192.0.2.3, as follows:

```
*A:PE-4# show router bgp routes mvpn-ipv4 type source-ad hunt
===============================================================================
 BGP Router ID:192.0.2.4        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
```

```
-------------------------------------------------------------------------------
Route Type     : Source-Ad
Route Dist.    : 64496:1
Source IP      : 172.16.1.11
Group IP       : 239.255.0.1
Nexthop        : 192.0.2.1
Path Id        : None
From           : 192.0.2.5
Res. Nexthop   : 0.0.0.0
Local Pref.    : 50                     Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : 64496:11 no-export target:64496:1
Cluster        : 1.1.1.1
Originator Id  : 192.0.2.1              Peer Router Id : 192.0.2.5
Flags          : Used  Valid  Best  IGP
--- snipped ---
Last Modified  : 00h06m54s
VPRN Imported  :  1

Route Type     : Source-Ad
Route Dist.    : 64496:1
Source IP      : 172.16.3.33
Group IP       : 239.255.0.1
Nexthop        : 192.0.2.3
Path Id        : None
From           : 192.0.2.5
Res. Nexthop   : 0.0.0.0
Local Pref.    : 200                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : 64496:33 no-export target:64496:1
Cluster        : 1.1.1.1
Originator Id  : 192.0.2.3              Peer Router Id : 192.0.2.5
Flags          : Used  Valid  Best  IGP
--- snipped ---
Last Modified  : 00h06m54s
VPRN Imported  :  1

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-4#
```

Host Group Membership

If the hosts then send a (*,G) request to join the group, a source-join route is originated by each receiver PE toward the preferred source from the redundant list.

The following output shows a join originated by PE-2:

```
*A:PE-2# show debug
debug
    router "Base"
        bgp
            update neighbor 192.0.2.5
        exit
    exit
exit
```

```
3 2017/10/12 10:08:34.595 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.5
"Peer 1: 192.0.2.5: UPDATE
Peer 1: 192.0.2.5 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 84
    Flag: 0x90 Type: 14 Len: 33 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.2
        Type: Source-Join Len:22 RD: 64496:1 SrcAS: 64496 Src: 172.16.1.11
                                             Grp: 239.255.0.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 16 Extended Community:
        target:64496:1
        target:192.0.2.1:2
"
```

When an active source AD route is present, there is no shared join sent to the RP. Because the source address is known, only a source-join needs to be sent. The source-join is trying to become a member of group 239.255.0.1 with a source address of 172.16.1.11. As this is sent as a BGP routing update, this must be accepted by the MVPN VRF at the PE that originated the unicast route that represents the c-multicast source. As previously mentioned, there are two extended community values. The second of these is the l2-vpn/vrf-import route target for 192.0.2.1 (PE-1), so only PE-1 will accept this route.

Examining the PIM state table for PE-2 shows the presence of a group with multiple sources.

```
*A:PE-2# show router 1 pim group

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
239.255.0.1                (*,G)                         mpls-if-73730  1
```

```
   *                           10.10.10.1
239.255.0.1                 (S,G)                spt      mpls-if-73730  1
   172.16.1.11                 10.10.10.1          A
239.255.0.1                 (S,G)                         mpls-if-73731  1
   172.16.3.33                 10.10.10.1          S
-------------------------------------------------------------------------------
Groups : 3
===============================================================================
*A:PE-2#
```

Each (S,G) has a state of either Active (A) or Standby (S), and the active group is chosen based on the Source Active AD with the highest local preference.

As a direct comparison, PE-4 also has the same two (S,G) states, but has a reversed active and standby source.

```
*A:PE-4# show router 1 pim group

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
239.255.0.1                 (*,G)                         mpls-if-73731  1
   *                           10.10.10.3
239.255.0.1                 (S,G)                         mpls-if-73730  1
   172.16.1.11                 10.10.10.3          S
239.255.0.1                 (S,G)                spt      mpls-if-73731  1
   172.16.3.33                 10.10.10.3          A
-------------------------------------------------------------------------------
Groups : 3
===============================================================================
*A:PE-4#
```

The Source Active ADs received on PE-4 have their local preference values based on the community string value.

```
*A:PE-4# show router bgp routes mvpn-ipv4 type source-ad hunt
===============================================================================
 BGP Router ID:192.0.2.4       AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Route Type     : Source-Ad
Route Dist.    : 64496:1
Source IP      : 172.16.1.11
Group IP       : 239.255.0.1
Nexthop        : 192.0.2.1
Path Id        : None
From           : 192.0.2.5
Res. Nexthop   : 0.0.0.0
Local Pref.    : 50                     Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : 64496:11 no-export target:64496:1
Cluster        : 1.1.1.1
Originator Id  : 192.0.2.1              Peer Router Id : 192.0.2.5
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
--- snipped ---
Last Modified  : 00h13m07s
VPRN Imported  :  1

Route Type     : Source-Ad
Route Dist.    : 64496:1
Source IP      : 172.16.3.33
Group IP       : 239.255.0.1
Nexthop        : 192.0.2.3
Path Id        : None
From           : 192.0.2.5
Res. Nexthop   : 0.0.0.0
Local Pref.    : 200                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : 64496:33 no-export target:64496:1
Cluster        : 1.1.1.1
Originator Id  : 192.0.2.3              Peer Router Id : 192.0.2.5
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
--- snipped ---
Last Modified  : 00h13m07s
VPRN Imported  :  1

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-4#
```

Sender PE MVPN Status

The MVPN status of the PE-1 sender PE is as follows:

```
*A:PE-1# show router 1 mvpn

===============================================================================
MVPN 1 configuration data
===============================================================================
signaling          : Bgp                  auto-discovery     : Default
UMH Selection      : Highest-Ip           SA withdrawn       : Disabled
intersite-shared   : Enabled              Persist SA         : Enabled
vrf-import         : MVPN1_import
vrf-export         : MVPN1_export
vrf-target         : N/A
C-Mcast Import RT  : target:192.0.2.1:2

ipmsi              : ldp
i-pmsi P2MP AdmSt  : Up
i-pmsi Tunnel Name : mpls-if-73729
Mdt-type           : sender-receiver

BSR signalling     : none
Wildcard s-pmsi    : Disabled
Multistream-SPMSI  : Disabled
s-pmsi             : none
data-delay-interval: 3 seconds
enable-asm-mdt     : N/A

===============================================================================
*A:PE-1#
```

The C-Mcast Import RT is set to <system-address>:<VPRN index>.

The VPRN index is derived from the following command:

```
*A:PE-1# admin display-config index | match vprn1
        virtual-router "vprn1" 2 0
*A:PE-1#
```

Any Source Join received must include this attribute along with the route target extended community. As previously stated, this is advertised within the VPN-IPv4 routes as a BGP attribute.

A source join received from PE-2 to join (S,G) (172.16.1.2, 239.255.0.1) is as follows:

```
*A:PE-1# show router bgp routes mvpn-ipv4 type source-join hunt
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
```

```
-------------------------------------------------------------------------------
Route Type     : Source-Join
Route Dist.    : 64496:1
Source AS      : 64496
Source IP      : 172.16.1.11
Group IP       : 239.255.0.1
Nexthop        : 192.0.2.2
Path Id        : None
From           : 192.0.2.5
Res. Nexthop   : 0.0.0.0
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : no-export target:64496:1 target:192.0.2.1:2
Cluster        : 1.1.1.1
Originator Id  : 192.0.2.2              Peer Router Id : 192.0.2.5
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h06m28s
VPRN Imported  :  1

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-1#
```

The PIM status for this group on sender PE-1 is as follows:

```
*A:PE-1# show router 1 pim group 239.255.0.1 source 172.16.1.11 detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 239.255.0.1
Source Address     : 172.16.1.11
RP Address         : 10.10.10.1
Advt Router        : 192.0.2.1
Flags              : spt                Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 172.16.1.11
MRIB Src Flags     : direct
Keepalive Timer Exp: 0d 00:00:37
Up Time            : 0d 00:19:42        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:00
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00
```

```
Register State     : Pruned            Register Stop Exp  : 0d 00:00:45
Reg From Anycast RP: No

Rpf Neighbor       : 172.16.1.11
Incoming Intf      : int-PE-1-S-1
Outgoing Intf List : mpls-if-73729

Curr Fwding Rate   : 1048.6 kbps
Forwarded Packets  : 102896            Discarded Packets  : 0
Forwarded Octets   : 154138208         RPF Mismatches     : 0
Spt threshold      : 0 kbps            ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-1#
```

The outgoing interface list is the I-PMSI, and traffic is seen to be flowing because the current forwarding rate is non-zero.

Similarly for sender PE-3, the MVPN status is:

```
*A:PE-3# show router 1 mvpn

===============================================================================
MVPN 1 configuration data
===============================================================================
signaling          : Bgp                 auto-discovery     : Default
UMH Selection      : Highest-Ip          SA withdrawn       : Disabled
intersite-shared   : Enabled             Persist SA         : Enabled
vrf-import         : MVPN1_import
vrf-export         : MVPN1_export
vrf-target         : N/A
C-Mcast Import RT  : target:192.0.2.3:2

ipmsi              : ldp
i-pmsi P2MP AdmSt  : Up
i-pmsi Tunnel Name : mpls-if-73729
Mdt-type           : sender-receiver

BSR signalling     : none
Wildcard s-pmsi    : Disabled
Multistream-SPMSI  : Disabled
s-pmsi             : none
data-delay-interval: 3 seconds
enable-asm-mdt     : N/A

===============================================================================
*A:PE-3#
```

The Source-Join route on PE-3 for this multicast group is:

```
*A:PE-3# show router bgp routes mvpn-ipv4 type source-join hunt
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Route Type     : Source-Join
Route Dist.    : 64496:1
Source AS      : 64496
Source IP      : 172.16.3.33
Group IP       : 239.255.0.1
Nexthop        : 192.0.2.4
Path Id        : None
From           : 192.0.2.5
Res. Nexthop   : 0.0.0.0
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : no-export target:64496:1 target:192.0.2.3:2
Cluster        : 1.1.1.1
Originator Id  : 192.0.2.4              Peer Router Id : 192.0.2.5
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h14m31s
VPRN Imported  :  1

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-3#
```

The PIM state for this group is as follows:

```
*A:PE-3# show router 1 pim group 239.255.0.1 source 172.16.3.33 detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address       : 239.255.0.1
Source Address      : 172.16.3.33
RP Address          : 10.10.10.3
Advt Router         : 192.0.2.3
Flags               : spt                Type               : (S,G)
```

```
Mode               : sparse
MRIB Next Hop      : 172.16.3.33
MRIB Src Flags     : direct
Keepalive Timer Exp: 0d 00:01:10
Up Time            : 0d 00:26:23        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:00
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : Pruned             Register Stop Exp  : 0d 00:01:21
Reg From Anycast RP: No

Rpf Neighbor       : 172.16.3.33
Incoming Intf      : int-PE-3-S-3
Outgoing Intf List : mpls-if-73729

Curr Fwding Rate   : 1042.6 kbps
Forwarded Packets  : 137847             Discarded Packets  : 0
Forwarded Octets   : 206494806          RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-3#
```

The preferred source remains active unless:

- The multicast source ceases to exist, the source PE withdraws the Source Active AD route
- Or a Source Active AD is received with a higher local preference.

Conclusion

MVPN Source Redundancy provides an optimal solution for multicast routing in a VPRN. This protocol provides simple configuration, operation and guaranteed fast protection time. It could be utilized in a regionalized IPTV solution where multiple sources for the same TV channel are used.

NG-MVPN Wildcard S-PMSI

This chapter provides information about next generation multicast virtual private networks (NG-MVPNs): use of wildcard selective PMSI.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

The chapter was initially written based on SR OS release 13.0.R4, but the CLI in the current edition is based on release 15.0.R5.

MPLS provider tunnels can be set up using multicast label distribution protocol (mLDP) or point-to-multipoint (P2MP) resource reservation protocol with traffic engineering (RSVP-TE) label switched paths (LSPs). SR OS release 12.0.R4 or later is required for route reflectors (RRs) peering with multicast virtual private network (MVPN) PEs.

Provider multicast service interfaces (PMSIs) are signaled using mLDP. PE MVPN auto-discovery uses BGP MVPN IPv4 network layer routing.

Knowledge of multi-protocol BGP (MP-BGP), RFC 4364, *BGP/MPLS IP Virtual Private Networks (VPNs)*, RFC 6513, *Multicast in MPLS/BGP IP VPNs*/RFC 6514, *BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs*, and RFC 6625, *Wildcards in Multicast VPN Auto-Discovery Routes*, is assumed throughout this chapter.

Overview

Consider a service provider core network used to deliver multicast services to a number of receiver PEs using Next Generation MVPN techniques, as defined in RFC 6513/6514, where multicast traffic is forwarded between PEs across a mesh of provider tunnels.

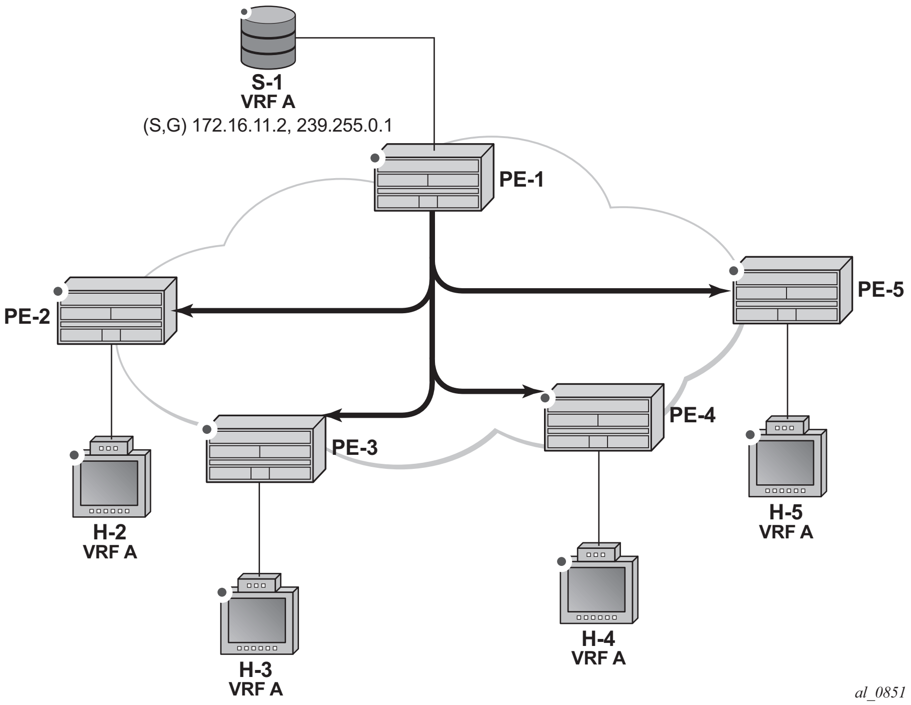
The provider tunnel used is the default Inclusive PMSI (I-PMSI) that is instantiated between all source and receiver PEs. This results in a full mesh of provider tunnels between all PEs in the MVPN. In a large network, this can result in an inefficient use of bandwidth because multicast traffic is forwarded to all PEs regardless of whether the PE has an interested receiver. Some of the mesh scaling issues can be mitigated by using source-only/destination-only configuration on PEs. However, this technique requires additional configuration and is not fully optimal when mLDP is used in the core.

To address the preceding limitation, wildcard Selective PMSI (S-PMSI) has been developed, as per RFC 6625. In the standard customer signaling notation of (C-S,C-G), this becomes (C-*,C-*). Using methods defined in RFC 6625, it is possible to use a (C-*,C-*) S-PMSI as the default tunnel, where the receiver PE can join the S-PMSI by mapping the channel join to a wildcard channel group. Multiple channels can be transported by the wildcard (C-*,C-*) S-PMSI, where an S-PMSI auto-discovery route is advertised with an empty channel group and source address:

1. Bandwidth savings can be achieved by the delivery of multicast channels on S-PMSIs, because traffic is not forwarded to PEs that have no interested receivers.
2. Control plane savings can be achieved by eliminating the need for the full tunnel mesh between all PES. The wildcard S-PMSI is only signaled on PEs containing attached upstream multicast sources, for which the PE is resolved as an upstream multicast hop (UMH) within the MVPN.

Figure 181 shows the concept of an MVPN.

Figure 181 **Multicast VPN**



In Figure 181, PE-1 has a directly connected multicast source (S-1). For clarity, consider this MVPN as a single multicast group. PE-1 is configured as a sender PE because it is the PE closest to the source. PE-2, PE-3, PE-4, and PE-5 are configured as receiver-only PEs because they have connected receiver hosts H-2, H-3, H-4, and H-5, respectively. Hosts H-2 to H-5 connected to receiver PEs can receive multicast channels from the source, S-1, connected to the source PE, PE-1, within the same virtual private routed network (VPRN).

Within the provider network, multicast traffic is delivered from the source PE to the receiver PE across a PMSI tunnel. This tunnel is, in this case, a P2MP LSP, with its root at PE-1 and with a leaf at each of the receiver PEs. Any traffic that is forwarded into the tunnel interface is replicated so that a single copy of a multicast stream is received at all PEs.

The PMSI tunnel is created after each PE has declared themselves as a member of the MVPN using BGP MVPN auto-discovery techniques.

There are two choices of PMSI:

- An I-PMSI, which is created on each PE within the MVPN, with a root at each PE and a leaf at all other PEs that are members of the MVPN. The I-PMSI is the default tunnel for all multicast traffic carried between sender PE and receiver PEs. When at least one receiver PE has a host interested in becoming a member of a multicast group, traffic for that group is delivered to all PEs via the I-PMSI, regardless of whether they have an interested host. Receiver PEs with no such interested host then drop the traffic.
- An S-PMSI, which is created to carry multicast traffic to the subset of receiver PEs that have connected hosts interested in receiving multicast traffic. This can be for a specific group, so that one S-PMSI carries traffic for one multicast group, denoted as (C-S,C-G) or (C-*,C-G). The S-PMSI can also be signaled to carry traffic for multiple multicast groups, denoted using a wildcard: (C-*,C-*) or (C-S,C-*). The wildcard S-PMSI can be signaled in place of the I-PMSI, so that all traffic can be carried on the S-PMSI by default. In this case, no I-PMSI is signaled.

In the case of an I-PMSI, the tunnel type is included in the BGP auto-discovery intra-AD route originated and advertised to all other PEs within the VPRN.

If a wildcard S-PMSI is to be used and no I-PMSI tunnel is to be signaled, then an intra-AD route update for I-PMSI is advertised with no tunnel type attribute included. In addition, the source PE will originate an additional S-PMSI auto-discovery route containing no source-encoding wildcard information.

Table 17 shows the S-PMSI MVPN BGP network layer reachability information (NLRI) advertisement.

Table 17 **S-PMSI Auto-Discovery BGP NLRI**

Route Distinguisher (8 octets)
Multicast Source Length (1 octet)
Multicast Source (variable)
Multicast Group Length (1 octet)
Multicast Group (variable)
Originating Router IP Address

To signal the S-PMSI as wildcard (C-*,C-*) S-PMSI, the multicast source length and multicast group length fields are encoded with the value of zero (0), representing (C-*,C-*) wildcard.

The objectives are to:

- Configure multicast in a VPRN on PE-1 to PE-5 using mLDP as the tunnel signaling method.
- Connect multicast sources to the sender PE-1.
- Create receiver hosts that can receive multicast traffic from the source, and to observe the effect on the provider network.

The following configuration tasks should be completed as a prerequisite:

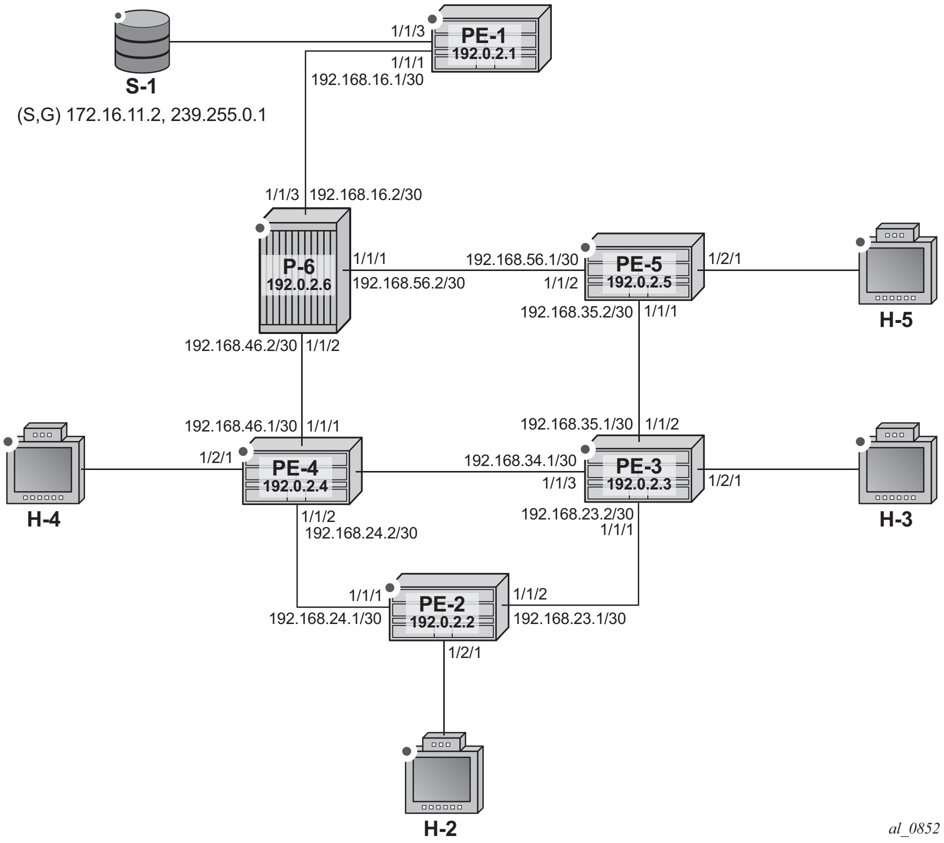
- Full mesh IS-IS or OSPF between each of the PE routers and the RR.
- Link-layer LDP between all PEs.
- mLDP used as the provider tunnel signaling protocol. This is enabled by default when link-layer LDP is enabled.

RSVP-TE is also supported as a provider tunnel signaling mechanism and could be used instead of mLDP.

Configuration

The example topology is shown in Figure 182, containing five PE routers. P-6 acts as an RR.

Figure 182 **Schematic Topology**



Global BGP Configuration

The first step is to configure an IBGP session between each of the PEs and the RR (PE-6) shown in Figure 182. The address families negotiated between the IBGP peers are vpn-ipv4 (unicast routing) and mvpn-ipv4 (multicast routing).

The configuration for PE1 is:

```
configure
    router
        bgp
            group INTERNAL
                family vpn-ipv4 mvpn-ipv4
                type internal
                neighbor 192.0.2.6
            exit
        exit
```

The configuration for the other PE nodes is exactly the same.

The configuration for the RR at P-6 is:

```
configure
    router
        bgp
            cluster 0.0.0.1
            group "RR_CLIENTS"
                family vpn-ipv4 mvpn-ipv4
                type internal
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
                neighbor 192.0.2.5
                exit
            exit
```

On PE-1, verify that the BGP session with RR at P-6 is established with address families vpn-ipv4 and mvpn-ipv4 capabilities negotiated:

```
*A:PE-1# show router bgp summary
===============================================================================
 BGP Router ID:192.0.2.1        AS:65545       Local AS:65545
===============================================================================
BGP Admin State         : Up          BGP Oper State              : Up
Total Peer Groups       : 1           Total Peers                 : 1
Total VPN Peer Groups   : 0           Total VPN Peers             : 0
Total BGP Paths         : 17          Total Path Memory           : 4488

--- snipped ---

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
                   AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.6
               65545        3    0 00h00m19s 0/0/0 (VpnIPv4)
                            3    0           0/0/0 (MvpnIPv4)
-------------------------------------------------------------------------------
*A:PE-1#
```

The same command can be used on the other PEs to verify their BGP sessions to the RR.

Configuring VPRN on PEs

The following outputs show the VPRN configurations for each PE. The specific MVPN configuration is shown later.

The VPRN configuration for PE-1 is:

```
# on PE-1
configure
    service
        vprn 1 customer 1 create
            route-distinguisher 65545:1
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:65545:1
            interface "int-PE-1-S-1" create
                address 172.16.11.1/24
                sap 1/1/3 create
                exit
            exit
            interface "rp" create
                address 10.0.0.1/32
                loopback
            exit
            pim
                apply-to all
                rp
                    static
                        address 10.0.0.1
                            group-prefix 239.255.0.0/16
                        exit
                    exit
                exit
                no shutdown
            exit
            no shutdown
```

There is a single interface toward S-1 from which the multicast group is transmitted.

If the customer signaling uses PIM ASM, a customer Rendezvous Point (RP) is required.

A loopback interface called “rp” acts as the RP for all group prefixes in the 239.255.0.0/16 range. This will be the RP for all groups.

MVPN configuration enables BGP as both the auto-discovery mechanism and the customer multicast signaling protocol across the VPRN. The provider tunnel between PEs within the MVPN is signaled using mLDP.

PE-2 contains an attached receiver, therefore a single interface is configured to accommodate this, as follows. The RP is configured as a static RP:

```
# on PE-2
configure
    service
        vprn 1 customer 1 create
            route-distinguisher 65545:1
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:65545:1
            interface "int-PE-2-H-2" create
                address 172.16.22.1/24
                sap 1/2/1 create
                exit
            exit
            igmp
                interface "int-PE-2-H-2”
                    no shutdown
                exit
                no shutdown
            exit
            pim
                apply-to all
                rp
                    static
                        address 10.0.0.1
                            group-prefix 239.255.0.0/16
                        exit
                    exit
                exit
                no shutdown
            exit
            no shutdown
```

PE-3 also has an attached receiver, as follows:

```
# on PE-3
configure
    service
        vprn 1 customer 1 create
            route-distinguisher 65545:1
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:65545:1
            interface "int-PE-3-H-3" create
                address 172.16.33.1/24
                sap 1/2/1 create
                exit
            exit
            igmp
                interface "int-PE-3-H-3"
                    no shutdown
                exit
                no shutdown
            exit
            pim
                apply-to all
                rp
                    static
                        address 10.0.0.1
                            group-prefix 239.255.0.0/16
                        exit
                    exit
                exit
                no shutdown
            exit
            no shutdown
```

The configuration for PE-4 and PE-5 is similar.

MVPN Configuration for PEs

The provider-tunnel inclusive configuration specifies that a wildcard S-PMSI will be used instead of an I-PMSI as the default PMSI. The MVPN configuration for all PEs is:

```
# on all PE’s
configure
    service
        vprn 1
            mvpn
                auto-discovery default
                c-mcast-signaling bgp
                provider-tunnel
                    inclusive
                        mldp
                            no shutdown
                        exit
                        wildcard-spmsi
                    exit
                exit
                vrf-target unicast
                exit
            exit
```

The keyword **wildcard-spmsi** reduces the number of PMSIs signaled. If there are no group sources on the receiver PEs, there will be no S-PMSI signaled. This has an effect similar to configuring the receiver PEs as MDT-type receiver-only.

Provider Tunnel Signaling

Each PE originates BGP MVPN intra-AD routes to determine membership of an MVPN.

The provider tunnels constructed between the PEs within the VPRN are signaled on receipt of an intra-AD route update from other PEs. The intra-AD update message contains details of the originator, along with the VRF route-target extended community. If an I-PMSI is to be signaled, a PMSI tunnel attribute is included that determines the tunnel type, such as LDP P2MP LSP. PEs that receive this intra-AD update will import the route into the MVPN, then signal a P2MP LDP label map toward the originator, which is the root of the LDP P2MP LSP.

However, if a wildcard S-PMSI is to be used as the default PMSI, no PMSI tunnel attribute is included within the intra-AD update.

The following output shows a BGP update originated by PE-1, and received by PE-2:

```
*A:PE-2# show router bgp routes mvpn-ipv4 type intra-ad rd 65545:1 detail
                                                       originator-ip 192.0.2.1
===============================================================================
 BGP Router ID:192.0.2.2        AS:65545       Local AS:65545
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
Original Attributes

Route Type     : Intra-Ad
Route Dist.    : 65545:1
Originator IP  : 192.0.2.1
Nexthop        : 192.0.2.1
Path Id        : None
From           : 192.0.2.6
Res. Nexthop   : 0.0.0.0
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : no-export target:65545:1
Cluster        : 0.0.0.1
Originator Id  : 192.0.2.1              Peer Router Id : 192.0.2.6
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
```

```
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h00m25s
VPRN Imported  :  1

Modified Attributes

Route Type     : Intra-Ad
Route Dist.    : 65545:1
--- snipped ---
VPRN Imported  :  1

-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-2#
```

There is no PMSI tunnel attribute included, and the route is imported into the correct VPRN (VPRN 1).

The intra-AD originated by PE-2 is:

```
*A:PE-1# show router bgp routes mvpn-ipv4 type intra-ad rd 65545:1
                                                  originator-ip 192.0.2.2 hunt
===============================================================================
 BGP Router ID:192.0.2.1        AS:65545       Local AS:65545
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Route Type     : Intra-Ad
Route Dist.    : 65545:1
Originator IP  : 192.0.2.2
Nexthop        : 192.0.2.2
Path Id        : None
From           : 192.0.2.6
Res. Nexthop   : 0.0.0.0
Local Pref.    : 100                    Interface Name : NotAvailable
Community      : no-export target:65545:1
Cluster        : 0.0.0.1
Originator Id  : 192.0.2.2              Peer Router Id : 192.0.2.6
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
```

```
Last Modified  : 00h00m50s
VPRN Imported  :  1

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-1#
```

This output also contains no PMSI tunnel attribute: PE-2 has no group source and there is no S-PMSI signaled. All other receiver PEs will originate a similar intra-AD route.

The following output shows all intra-AD routes originated by the PEs within the VPRN, as received by PE-1:

```
*A:PE-1# show router bgp routes mvpn-ipv4 type intra-ad rd 65545:1
===============================================================================
 BGP Router ID:192.0.2.1        AS:65545       Local AS:65545
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
Flag  RouteType                   OriginatorIP           LocalPref   MED
      RD                          SourceAS               Path-Id     Label
      Nexthop                     SourceIP
      As-Path                     GroupIP
-------------------------------------------------------------------------------
i     Intra-Ad                    192.0.2.1              100         0
      65545:1                     -                      None        -
      192.0.2.1                   -
      No As-Path                  -
u*>i  Intra-Ad                    192.0.2.2              100         0
      65545:1                     -                      None        -
      192.0.2.2                   -
      No As-Path                  -
u*>i  Intra-Ad                    192.0.2.3              100         0
      65545:1                     -                      None        -
      192.0.2.3                   -
      No As-Path                  -
u*>i  Intra-Ad                    192.0.2.4              100         0
      65545:1                     -                      None        -
      192.0.2.4                   -
      No As-Path                  -
u*>i  Intra-Ad                    192.0.2.5              100         0
      65545:1                     -                      None        -
      192.0.2.5                   -
      No As-Path                  -
-------------------------------------------------------------------------------
Routes : 5
===============================================================================
*A:PE-1#
```

Instead of an I-PMSI being signaled, an S-PMSI AD route update is advertised by PE-1 to all receiver PEs within the MVPN. The NLRI encoding has a zero length field for both source and group addresses, so is seen to represent multicast group (C-*,C-*). This is wildcard nomenclature for both source and group addresses.

The BGP route as advertised by PE-1:

```
*A:PE-1# show router bgp routes mvpn-ipv4 type spmsi-ad rd 65545:1 hunt
===============================================================================
 BGP Router ID:192.0.2.1       AS:65545       Local AS:65545
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
--- snipped ---
-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
Route Type      : Spmsi-Ad
Route Dist.     : 65545:1
Originator IP   : 192.0.2.1
Source IP       : 0.0.0.0
Group IP        : 0.0.0.0
Nexthop         : 192.0.2.1
Path Id         : None
To              : 192.0.2.6
Res. Nexthop    : n/a
Local Pref.     : 100                   Interface Name : NotAvailable
Aggregator AS   : None                  Aggregator     : None
--- snipped ---
Community       : no-export target:65545:1
Cluster         : No Cluster Members
Originator Id   : None                  Peer Router Id : 192.0.2.6
Origin          : IGP
AS-Path         : No As-Path
Route Tag       : 0
Neighbor-AS     : N/A
Orig Validation: N/A
Source Class    : 0                     Dest Class     : 0
-------------------------------------------------------------------------------
PMSI Tunnel Attributes :
Tunnel-type     : LDP P2MP LSP
Flags           : Type: RNVE(0) BM: 0 U: 0 Leaf: not required
MPLS Label      : 0
Root-Node       : 192.0.2.1             LSP-ID         : 8193
-------------------------------------------------------------------------------
Routes : 6
===============================================================================
*A:PE-1#
```

The source IP and group IP address fields are advertised as 0.0.0.0, and the tunnel type attribute is now included as an LDP P2MP LSP.

The following output shows the MVPN status at PE-1, with the I-PMSI tunnel name containing a wildcard S-PMSI denoted by (W):

```
*A:PE-1# show router 1 mvpn

===============================================================================
MVPN 1 configuration data
===============================================================================
signaling          : Bgp                  auto-discovery     : Default
UMH Selection      : Highest-Ip           SA withdrawn       : Disabled
intersite-shared   : Enabled              Persist SA         : Disabled
vrf-import         : N/A
vrf-export         : N/A
vrf-target         : unicast
C-Mcast Import RT  : target:192.0.2.1:2

ipmsi              : ldp
i-pmsi P2MP AdmSt  : Up
i-pmsi Tunnel Name : mpls-if-73728(W)
Mdt-type           : sender-receiver

BSR signalling     : none
Wildcard s-pmsi    : Enabled
Multistream-SPMSI  : Disabled
s-pmsi             : none
data-delay-interval: 3 seconds
enable-asm-mdt     : N/A

===============================================================================
*A:PE-1#
```

At this point, there is no interested host and no customer multicast flow (c-flow), so there is no S-PMSI LDP P2MP LSP signaled.

Data Transmission at Source PE

Multicast traffic for a particular group will be forwarded between the sender and receiver PE over a provider tunnel (PMSI). Because there is no default I-PMSI present, the receiver PE has to choose an S-PMSI to be used for forwarding, based on the NLRI contained within the S-PMSI AD routes.

The provider tunnel is signaled using a P2MP LDP label mapping message toward the root address signaled in the wildcard S-PMSI AD BGP update message. As previously shown, the update message is based on whether traffic is being forwarded on the shared or source-based shortest path tree.

When joining the shared tree, a c-multicast shared-join is sent toward the appropriate PE, which represents the UMH toward the RP. The UMH is chosen from the unicast route that represents the RP address. When joining the shortest path tree, a source-join c-multicast route is sent toward the UMH chosen from the unicast route that represents the actual source address. In both cases, the VPN-IPv4 unicast route advertises a VRF route import community that contains the system address as a next hop. Upon receipt of these updates, the UMH PE will forward traffic along the signaled wildcard S-PMSI.

Each S-PMSI is bound to one or more flows, as determined by the NLRI contained within the S-PMSI BGP update. The use of the wildcard within the BGP update to replace the c-source and c-group allows multiple flows to be bound to a single provider tunnel.

Traffic will only be forwarded upon reception of a BGP MVPN source-join or shared-join BGP route at the sender PE.

Data Reception at Receiver PE

When the sender PE has originated an S-PMSI AD route update, each receiver PE will install the route into its VRF. The S-PMSIs installed are used to select an appropriate upstream multicast router for a c-flow when an attached receiver is interested in receiving traffic from that c-flow.

The receiver PE will receive a flow based on the best match of the incoming (C-S,C-G) or (C-*,C-G) IGMP/MLD or PIM join.

If an IGMP/MLD group membership query or PIM join is received by the receiver PE over an attachment circuit for a group, the contained (C-S,C-G) or (C-*,C-G) must match the (C-S,C-G) contained within any installed S-PMSI AD route. In the case of the wildcard S-PMSI being the only installed NLRI, this will be a match; that is, incoming (C-*,C-G) or (C-S,C-G) will match the S-PMSI (C-*,C-*).

In this example, the c-group flow is 239.255.0.1.

Traffic Flow

A traffic stream representing a c-flow is enabled on PE-1: group address 239.255.0.1 with source address of 172.16.11.2. The RP for this group is found locally on PE-1. The outgoing interface list is empty:

```
*A:PE-1# show router 1 pim group detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 239.255.0.1
Source Address     : 172.16.11.2
RP Address         : 10.0.0.1
Advt Router        : 192.0.2.1
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 172.16.11.2
MRIB Src Flags     : direct
Keepalive Timer Exp: 0d 00:03:22
Up Time            : 0d 00:00:08        Resolved By        : rtable-u

Up JP State        : Not Joined         Up JP Expiry       : 0d 00:00:00
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : Pruned             Register Stop Exp  : 0d 00:00:46
Reg From Anycast RP: No

Rpf Neighbor       : 172.16.11.2
Incoming Intf      : int-PE-1-S-1
Outgoing Intf List :
Outgoing Sap List  :
Outgoing Host List :

Curr Fwding Rate   : 1018.6 kbps
Forwarded Packets  : 653                Discarded Packets  : 0
Forwarded Octets   : 978194             RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-1#
```

A host connected to PE-2 sends a (C-*,C-G) IGMP v2 group membership query for group 239.255.0.1.

PE-2 sends a BGP MVPN shared-join route update toward PE-1, where the RP address of the group 10.0.0.1 is found.

The following debug output from PE-2 shows the shared-join BGP route update transmitted by PE-2:

```
1 2017/10/11 12:05:18.893 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.6
"Peer 1: 192.0.2.6: UPDATE
Peer 1: 192.0.2.6 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 76
    Flag: 0x90 Type: 14 Len: 33 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.2
        Type: Shared-Join Len:22 RD: 65545:1 SrcAS: 65545 Src: 10.0.0.1
                                             Grp: 239.255.0.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:192.0.2.1:2
"
```

Upon receipt of the shared-join, traffic flows on the shared tree toward the receiver PE. This will flow on the default wildcard S-PMSI, as shown in the outgoing interface list:

```
*A:PE-1# show router 1 pim group 239.255.0.1 type starg detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 239.255.0.1
Source Address     : *
RP Address         : 10.0.0.1
Advt Router        : 192.0.2.1
Flags              :                    Type               : (*,G)
Mode               : sparse
MRIB Next Hop      :
MRIB Src Flags     : self
Keepalive Timer    : Not Running
Up Time            : 0d 00:04:16        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:43
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Rpf Neighbor       :
Incoming Intf      :
Outgoing Intf List : mpls-if-73728(W)

Curr Fwding Rate   : 0.0 kbps
Forwarded Packets  : 0                  Discarded Packets  : 0
Forwarded Octets   : 0                  RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-1#
```

When traffic is received on the shared tree by PE-2, the source address is learned, so a source-join BGP route update is sent toward the UMH PE, which contains the source address of 172.16.11.2. The UMH is chosen from the unicast route-table using a lookup for the best route matching the source address.

The following debug output from PE-2 shows the BGP source-join route update toward the source of group 239.255.0.1, as transmitted by PE-2:

```
3 2017/10/11 12:05:47.191 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.6
"Peer 1: 192.0.2.6: UPDATE
Peer 1: 192.0.2.6 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 76
    Flag: 0x90 Type: 14 Len: 33 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.2
        Type: Source-Join Len:22 RD: 65545:1 SrcAS: 65545 Src: 172.16.11.2
                                             Grp: 239.255.0.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:192.0.2.1:2
"
```

The c-flow toward host H-2 flows along the shortest path tree, and on PE-1 the outgoing interface list is populated with the wildcard S-PMSI:

```
*A:PE-1# show router 1 pim group detail 239.255.0.1 source 172.16.11.2

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 239.255.0.1
Source Address     : 172.16.11.2
RP Address         : 10.0.0.1
Advt Router        : 192.0.2.1
Flags              : spt, rpt-prn-des   Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 172.16.11.2
MRIB Src Flags     : direct
Keepalive Timer Exp: 0d 00:00:47
Up Time            : 0d 00:16:49        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:00
Up JP Rpt          : Pruned             Up JP Rpt Override : 0d 00:00:00

Register State     : Pruned             Register Stop Exp  : 0d 00:00:05
Reg From Anycast RP: No

Rpf Neighbor       : 172.16.11.2
Incoming Intf      : int-PE-1-S-1
Outgoing Intf List : mpls-if-73728(W)

Curr Fwding Rate   : 1018.6 kbps
Forwarded Packets  : 85773              Discarded Packets  : 0
Forwarded Octets   : 128487954          RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-1#
```

The outgoing interface is the MPLS interface mpls-if-73728. This maps to a P2MP LDP label binding from which the p2mp-id can be derived:

```
*A:PE-1# show router ldp bindings active p2mp ipv4 opaque-type generic

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
===============================================================================
LDP Generic IPv4 P2MP Bindings (Active)
===============================================================================
P2MP-Id                                     Interface
RootAddr                                    Op              IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
-------------------------------------------------------------------------------
8193                                        73728
192.0.2.1                                   Push              --      262137
192.168.16.2                                1/1/1

-------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 1
===============================================================================
*A:PE-1#
```

After the source-join is received, the sender PE will advertise a source-active AD BGP route to all PEs within the MVPN, to announce the presence of a (C-S,C-G) group. The following debug output shows the source-active AD route received on PE-2:

```
5 2017/10/11 12:06:17.172 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.6
"Peer 1: 192.0.2.6: UPDATE
Peer 1: 192.0.2.6 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 86
    Flag: 0x90 Type: 14 Len: 29 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.1
        Type: Source-AD Len: 18 RD: 65545:1 Src: 172.16.11.2 Grp: 239.255.0.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.1
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        0.0.0.1
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:65545:1
"
```

The PIM status of the group on receiver PE-2 shows that the incoming interface is the wildcard S-PMSI originated on PE-1, as denoted by the (W):

```
*A:PE-2# show router 1 pim group 239.255.0.1 source 172.16.11.2 detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 239.255.0.1
Source Address     : 172.16.11.2
RP Address         : 10.0.0.1
Advt Router        : 192.0.2.1
Flags              : spt                Type              : (S,G)
Mode               : sparse
MRIB Next Hop      : 192.0.2.1
MRIB Src Flags     : remote
Keepalive Timer Exp: 0d 00:02:02
Up Time            : 0d 00:08:58        Resolved By       : rtable-u

Up JP State        : Joined             Up JP Expiry      : 0d 00:00:02
Up JP Rpt          : Not Pruned         Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 192.0.2.1
Incoming Intf      : mpls-if-73729(W)
Outgoing Intf List : int-PE-2-H-2

Curr Fwding Rate   : 1018.6 kbps
Forwarded Packets  : 45751              Discarded Packets  : 0
Forwarded Octets   : 68534998           RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-2#
```

The S-PMSI is an LDP P2MP LSP. The LDP label binding for P2MP LSP-Id 8193 at PE-2 shows that the label operation is a label pop:

```
*A:PE-2# show router ldp bindings active p2mp p2mp-id 8193 root 192.0.2.1

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
===============================================================================
LDP Generic IPv4 P2MP Bindings (Active)
===============================================================================
P2MP-Id                                     Interface
RootAddr                                    Op              IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
-------------------------------------------------------------------------------
8193                                        73729
192.0.2.1                                   Pop             262136      --
  --                                          --

-------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 1
===============================================================================
*A:PE-2#
```

PE-3 has no host joined to c-flow group 239.255.0.1. However, it contains the PIM state for this group due to the presence of the source-active AD route within the VRF. This route was received when the host connected to PE-2 joined the c-flow group.

The following output shows the source-active AD route within PE-3 for group 239.255.0.1 from source 172.16.11.2:

```
*A:PE-3# show router bgp routes mvpn-ipv4 type source-ad rd 65545:1
===============================================================================
 BGP Router ID:192.0.2.3        AS:65545       Local AS:65545
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MVPN-IPv4 Routes
===============================================================================
Flag  RouteType                   OriginatorIP           LocalPref   MED
      RD                          SourceAS               Path-Id     Label
      Nexthop                     SourceIP
      As-Path                     GroupIP
-------------------------------------------------------------------------------
u*>i  Source-Ad                   -                      100         0
      65545:1                     -                      None        -
      192.0.2.1                   172.16.11.2
      No As-Path                  239.255.0.1
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-3#
```

However, traffic is not received from the S-PMSI because there is no label binding for the LDP P2MP LSP. The following output shows that there is no label binding for the LSP Id 8193, which has its root on PE-1:

```
*A:PE-3# show router ldp bindings p2mp p2mp-id 8193 root 192.0.2.1

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.3)
             (IPv6 LSR ID ::)
```

```
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
===============================================================================
LDP Generic IPv4 P2MP Bindings
===============================================================================
P2MP-Id
RootAddr                                    Interface       IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
Peer
-------------------------------------------------------------------------------
No Matching Entries Found
===============================================================================
*A:PE-3#
```

A static IGMPv2 (*,G) group is created on interface int-PE-3-H3 for group 239.255.0.1 toward PE-3. The following debug output shows the process.

```
4 2017/10/11 11:32:30.123 UTC MINOR: DEBUG #2001 vprn1 IGMP[vprn1 inst 2]
"IGMP[vprn1 inst 2]: igmpIfSGStaticAdd
Adding Static SG (0.0.0.0,239.255.0.1) to IGMP interface int-PE-3-H-3 [ifIndex 5]"

5 2017/10/11 11:32:30.123 UTC MINOR: DEBUG #2001 vprn1 IGMP[vprn1 inst 2]
"IGMP[vprn1 inst 2]: igmpIfGroupAdd
Adding 239.255.0.1 to IGMP interface int-PE-3-H-3 [ifIndex 5] database"

6 2017/10/11 11:32:30.123 UTC MINOR: DEBUG #2001 vprn1 IGMP[vprn1 inst 2]
"IGMP[vprn1 inst 2]: igmpProcessGroupRec
Process group rec CHG_TO_EXCL received on interface int-PE-3-H-3 [ifIndex 5]
                            for group 239.255.0.1 in mode INCLUDE. Num srcs 0"

7 2017/10/11 11:32:30.123 UTC MINOR: DEBUG #2001 vprn1 IGMP[vprn1 inst 2]
"IGMP[vprn1 inst 2]: igmpIfSrcAdd
Adding i/f source entry for interface int-PE-3-H-3 [ifIndex 5] (*,239.255.0.1)
                            to IGMP fwdList Database, redir if N/A"
```

A similar process takes place when receiver host H-3 sends an unsolicited IGMP v2 group membership query for this group. The first message would correspond to the IGMP query instead.

After the IGMP interface source entry has been added for interface int-PE-3-H-3, an mLDP P2MP label mapping message is sent from PE-3 toward the root node PE-1, as follows:

```
8 2017/10/11 11:32:30.123 UTC MINOR: DEBUG #2001 Base LDP
"LDP: Binding
Sending Label mapping label 262136 for P2MP: root = 192.0.2.1, T: 1, L: 4,
 TunnelId: 8193 to peer 192.0.2.4:0."

9 2017/10/11 11:32:30.123 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
```

```
Send Label Mapping packet (msgId 618) to 192.0.2.4:0
Protocol version = 1
Label 262136 advertised for the following FECs
P2MP: root = 192.0.2.1, T: 1, L: 4, TunnelId: 8193
"
```

BGP shared-join and source-join BGP route updates are sent via the RR toward the RP (source = 10.0.0.1) and the actual source (172.16.11.2), respectively:

```
10 2017/10/11 11:32:30.124 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.6
"Peer 1: 192.0.2.6: UPDATE
Peer 1: 192.0.2.6 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 100
    Flag: 0x90 Type: 14 Len: 57 Multiprotocol Reachable NLRI:
        Address Family MVPN_IPV4
        NextHop len 4 NextHop 192.0.2.3
        Type: Shared-Join Len:22 RD: 65545:1 SrcAS: 65545 Src: 10.0.0.1
                                                          Grp: 239.255.0.1
        Type: Source-Join Len:22 RD: 65545:1 SrcAS: 65545 Src: 172.16.11.2
                                                          Grp: 239.255.0.1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 8 Len: 4 Community:
        no-export
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:192.0.2.1:2
"
```

The PIM status for the c-group 239.255.0.1 on PE-3 is as follows:

```
*A:PE-3# show router 1 pim group

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
239.255.0.1                 (*,G)                        mpls-if-73729* 1
   *                           10.0.0.1
239.255.0.1                 (S,G)               spt      mpls-if-73729* 1
   172.16.11.2                 10.0.0.1
-------------------------------------------------------------------------------
Groups : 2
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-3#
```

Assume that a receiver on each of PE-4 and PE-5 needs to join group 239.255.0.1, as shown in Figure 183.

Figure 183 **S-PMSI P2MP LSP Schematic**

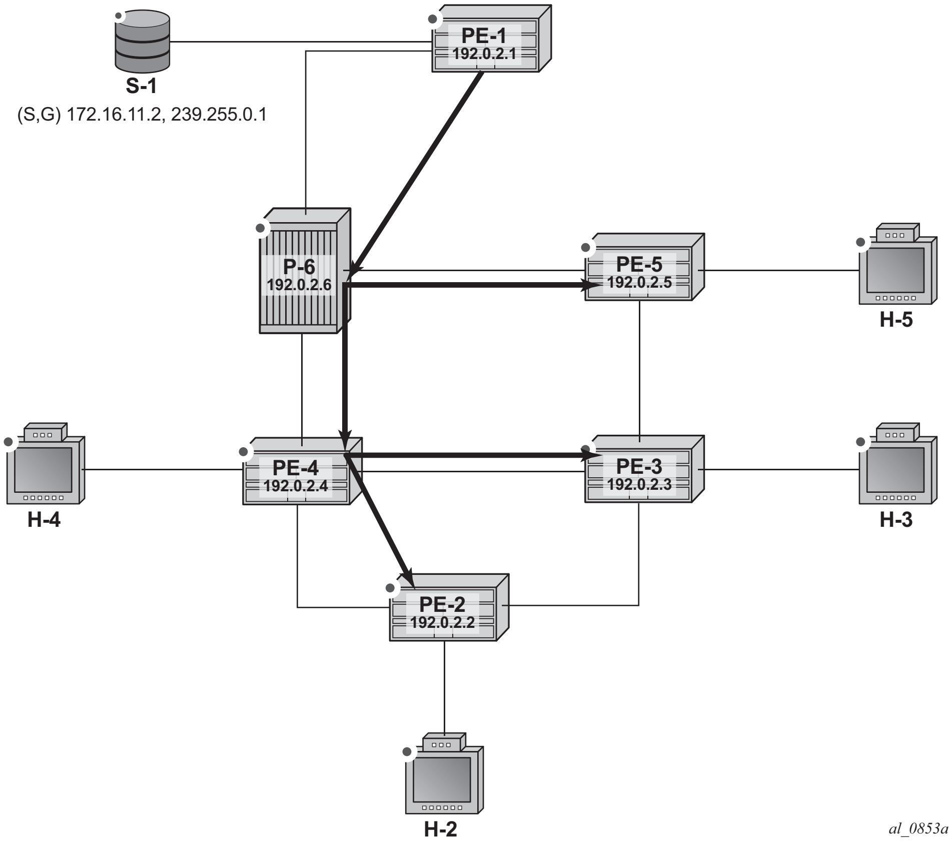


Figure 183 shows the S-PMSI P2MP LSP. The next set of outputs shows the P2MP label mapping of the LDP LSP between PE-1 and the receiver PEs.

The root of the S-PMSI is on PE-1, as follows:

```
*A:PE-1# show router ldp bindings active p2mp p2mp-id 8193 root 192.0.2.1
===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
===============================================================================
LDP Generic IPv4 P2MP Bindings (Active)
===============================================================================
P2MP-Id                                      Interface
RootAddr                                     Op              IngLbl    EgrLbl
EgrNH                                        EgrIf/LspId
```

```
--------------------------------------------------------------------------------
8193                                         73728
192.0.2.1                                    Push             --      262137
192.168.16.2                                 1/1/1
--------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 1
================================================================================
*A:PE-1#
```

The egress label on PE-1 becomes the ingress label on P-6. P-6 has two leaves: one toward PE-4 and one toward PE-5, as follows:

```
*A:P-6# show router ldp bindings active p2mp p2mp-id 8193 root 192.0.2.1
================================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.6)
             (IPv6 LSR ID ::)
================================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
================================================================================
LDP Generic IPv4 P2MP Bindings (Active)
================================================================================
P2MP-Id                                      Interface
RootAddr                                     Op               IngLbl    EgrLbl
EgrNH                                        EgrIf/LspId
--------------------------------------------------------------------------------
8193                                         Unknw
192.0.2.1                                    Swap             262137    262136
192.168.46.1                                 1/1/2

8193                                         Unknw
192.0.2.1                                    Swap             262137    262136
192.168.56.1                                 1/1/1
--------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 2
================================================================================
*A:P-6#
```

On PE-5, the following output shows that the LSP terminates as a leaf, as the operation (Op) is shown as “pop”:

```
*A:PE-5# show router ldp bindings active p2mp p2mp-id 8193 root 192.0.2.1
================================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.5)
             (IPv6 LSR ID ::)
================================================================================
--- snipped ---
================================================================================
LDP Generic IPv4 P2MP Bindings (Active)
================================================================================
P2MP-Id                                      Interface
RootAddr                                     Op               IngLbl    EgrLbl
EgrNH                                        EgrIf/LspId
```

```
--------------------------------------------------------------------------------
8193                                         73729
192.0.2.1                                    Pop             262136      --
  --                                           --

--------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 1
================================================================================
*A:PE-5#
```

On PE-4, the P2MP LSP has 3 entries: a pop operation to receiver H-4, and two label swaps toward PE-3 and PE-2:

```
*A:PE-4# show router ldp bindings active p2mp p2mp-id 8193 root 192.0.2.1
================================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.4)
             (IPv6 LSR ID ::)
================================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
================================================================================
LDP Generic IPv4 P2MP Bindings (Active)
================================================================================
P2MP-Id                                      Interface
RootAddr                                     Op              IngLbl    EgrLbl
EgrNH                                        EgrIf/LspId
--------------------------------------------------------------------------------
8193                                         73729
192.0.2.1                                    Pop             262136      --
  --                                           --

8193                                         73729
192.0.2.1                                    Swap            262136    262136
192.168.24.1                                 1/1/2

8193                                         73729
192.0.2.1                                    Swap            262136    262136
192.168.34.1                                 1/1/3
--------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 3
================================================================================
*A:PE-4#
```

PE-2 and PE-3 are termination PEs for P2MP leaf. On PE-2, the pop operation is shown:

```
*A:PE-2# show router ldp bindings active p2mp p2mp-id 8193 root 192.0.2.1
================================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
================================================================================
--- snipped ---
P2MP-Id                                      Interface
```

```
RootAddr                                     Op              IngLbl    EgrLbl
EgrNH                                        EgrIf/LspId
-------------------------------------------------------------------------------
8193                                         73729
192.0.2.1                                    Pop             262136      --
  --                                           --
-------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 1
===============================================================================
*A:PE-2#
```

On PE-3, the P2MP pop operation is shown:

```
*A:PE-3# show router ldp bindings active p2mp p2mp-id 8193 root 192.0.2.1
===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.3)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
===============================================================================
LDP Generic IPv4 P2MP Bindings (Active)
===============================================================================
P2MP-Id                                      Interface
RootAddr                                     Op              IngLbl    EgrLbl
EgrNH                                        EgrIf/LspId
-------------------------------------------------------------------------------
8193                                         73729
192.0.2.1                                    Pop             262136      --
  --                                           --

-------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 1
===============================================================================
*A:PE-3#
```

Conclusion

MVPN wildcard Selective PMSI (S-PMSI), developed as per RFC 6625, provides an optimal solution for multicast routing in a VPRN. This protocol provides simple configuration, operation, and fast protection time in conjunction with MPLS and LDP fast-failover schemes. Wildcard S-PMSI can be used in a multicast network to avoid a large full mesh of an I-PMSI.

Rosen MVPN Core Diversity

This chapter provides information about Rosen multicast virtual private network (MVPN) core diversity.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SR OS release 13.0.R4, using Rosen MVPN. The CLI in the current edition is based on SR OS release 15.0.R5.

Default multicast distribution trees (MDTs) for each virtual private routed network (VPRN) are signaled using protocol independent multicasting (PIM) and auto-discovery uses border gateway protocol multicast distribution tree sub-address family indicator (BGP MDT-SAFI) network layer routing.

Overview

This chapter describes a service provider core network used by multiple content providers to deliver multicast services to multiple customers using Rosen MVPN. If the same set of PEs is used to deliver the MVPN, the MDTs will all be routed across the same paths between the set of PEs. Because each MDT is signaled using PIM, and the source of all MDTs is the system address of the PE, the path to this source is the same.

Each remote PE then sends a PIM join toward this PE with its source address set to the system address. For multiple VPNs between the same set of PEs, the MDT will follow the same path.

If there is a requirement to deliver content from each content provider across different MVPNs that use diversely routed MDTs, multiple IGP instances can be used: up to three different instances of IGP, OSPF or ISIS, can exist. In this chapter, two instances of OSPF are used to create incongruent topologies providing isolation between the MDTs of two different MVPNs: a default OSPF instance and OSPF instance 1. A separate /32 loopback address can be used as the MDT source address that is advertised in the non-default IGP, which can also be used as the BGP next hop for labeled IPv4 routes representing the customer source addresses.

Knowledge of multi-protocol BGP (MP-BGP) and RFC 4364 (*BGP/MPLS IP VPNs*) is assumed throughout this chapter, as well as the original RFC 6037.

All PEs within an MVPN create a default MDT with their own system address as the source. Auto-discovery of PEs within a Rosen MVPN is achieved using the BGP route type of multicast data tree subsequent address family identifier (MDT-SAFI). Each PE originates an MDT-SAFI route update per MVPN. This route advertises the presence of the MVPN on a specific PE.

Each MDT-SAFI update contains attributes, including the following:

1. Route distinguisher
2. Route target extended community
3. MDT source address (usually the system address)
4. Group address of MDT

Upon receipt of an MDT-SAFI route update, each remote PE accepts or rejects the route based on the route target extended community value. If the route is accepted, a remote PE sends a PIM (S,G) join to this local PE. The (S,G) values are taken from the MDT-SAFI. The set of MDTs extend the c-multicast data tree across the MVPN and form PIM adjacencies between PEs within the MVPN. The neighbor address across the set of PIM-enabled tunnels is the default MDT source address, usually the system address.

When established, the default MDT is used to transport c-multicast PIM signaling between PEs.

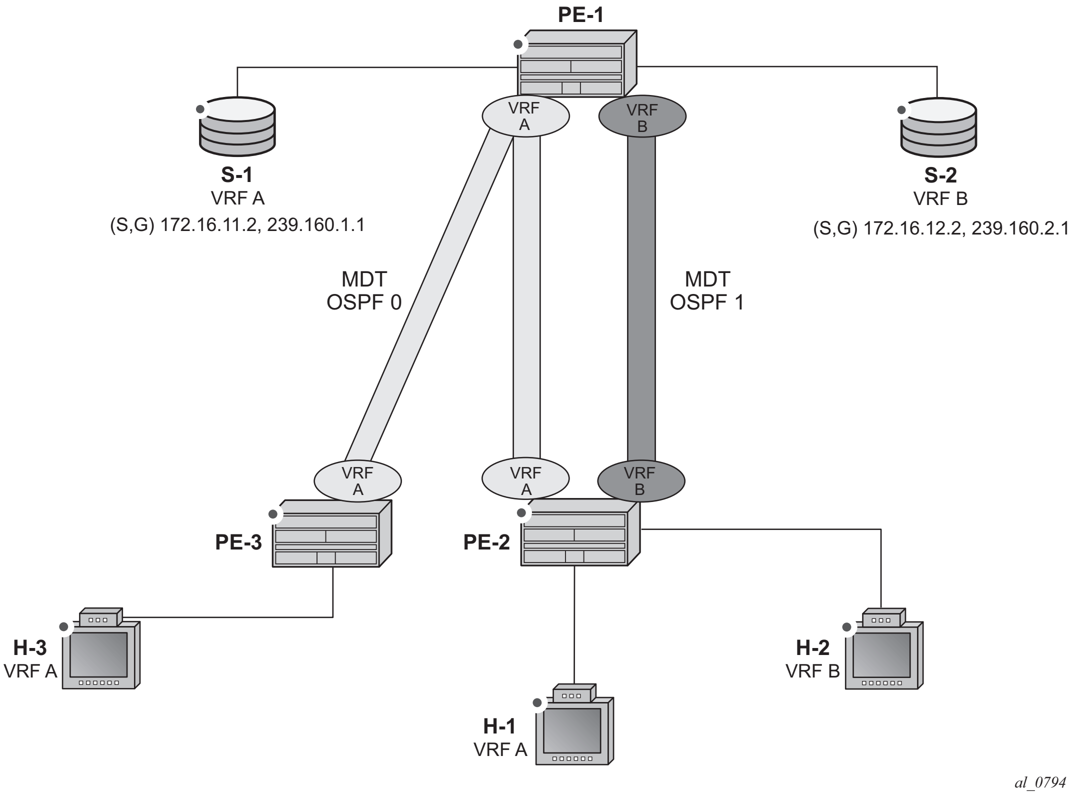
If a source S, of a c-multicast group G, is connected to a sender PE, the route to the source is advertised to remote PEs as a BGP-labeled VPN-IPv4 route.

Therefore, an (S,G) join toward this source at a remote PE will perform a reverse path forwarding (RPF) look-up of the unicast VRF table to find a suitable PIM-enabled interface. The next hop needs be resolved to the MDT source address of the sender PE. A PIM join must now be forwarded toward the sender PE that has a PIM neighbor that matches the next hop for this route, the system address of the sender PE. This is the default MDT.

The system address is a significant address in this process. Any other VPRN that uses the same set of PEs will also signal a set of default MDTs using a different group address, so they will follow the same path between PEs across the provider network.

Figure 184 shows an example of core diversity; multicast sources provided by two separate content providers are connected to a provider network. There is a requirement to provide topology diversity so that the default MDTs between the same PEs are routed across different paths within the core.

Figure 184 **Core Diversity Schematic**



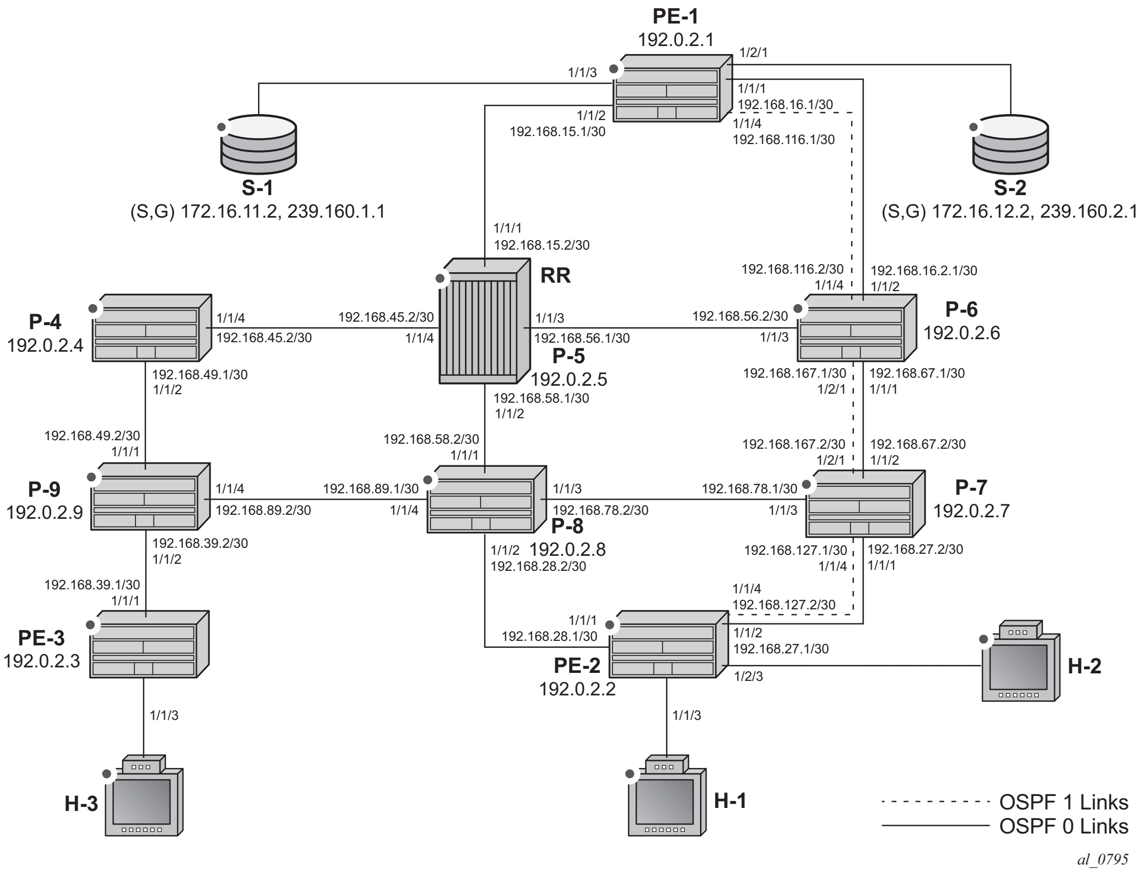
Content servers from two separate content providers are connected to PE-1 with directly connected multicast sources. For simplicity, this example uses only a single multicast group for provider S-1 and S-2.

Source S-1 is reachable via VRF A and source S-2 is reachable via VRF B.

Topology isolation for the multicast data trees of each VPRN can be provided by using two separate IGP instances; in this case, OSPF instances. Multi-instance IS-IS could also be used.

Figure 185 shows a schematic of the full network, including the c-multicast groups.

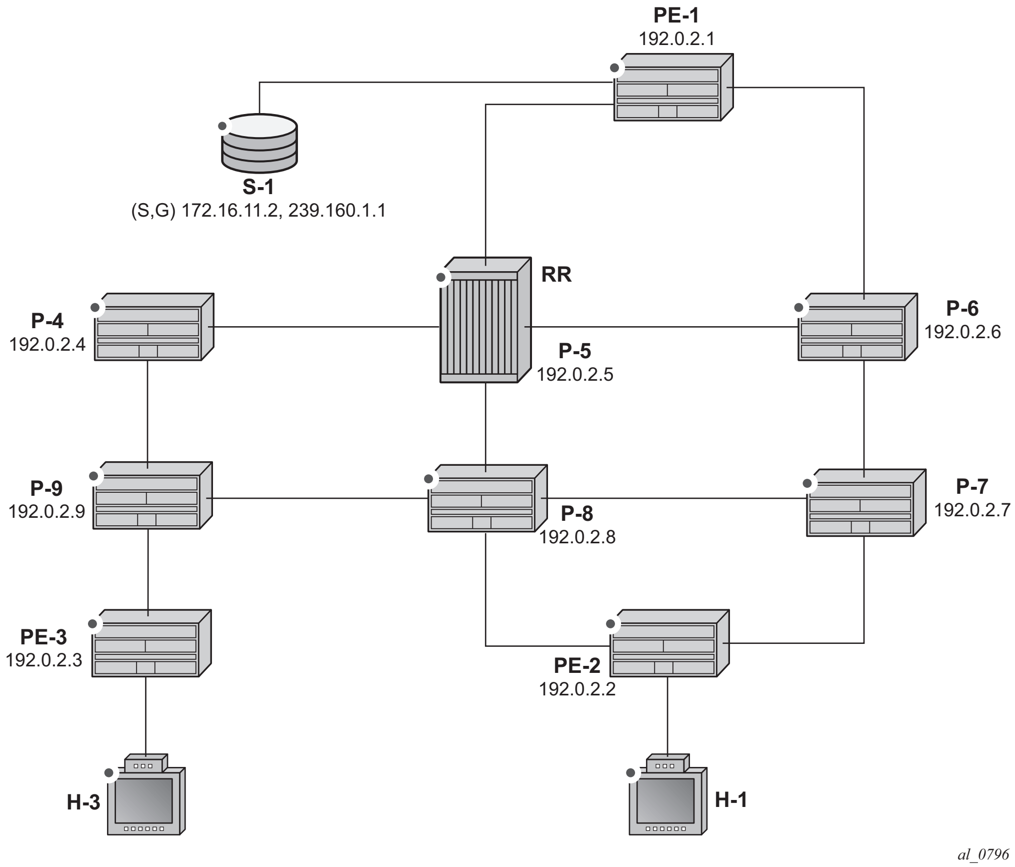
Figure 185 **Core Diversity Network**



All routers have interfaces in the OSPF base instance (instance 0) and routers interconnected by the dotted lines have interfaces in both the base instance and OSPF 1.

Figure 185 shows the extent of the OSFP base instance within the core network.

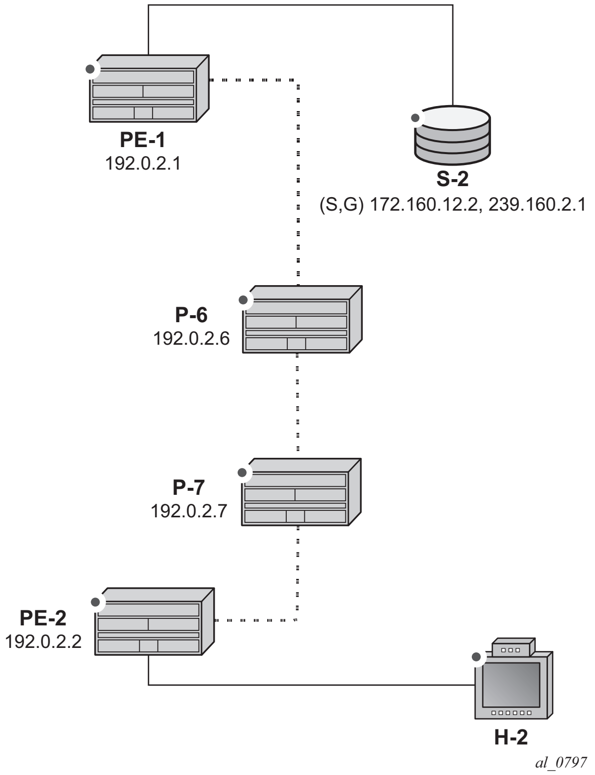
Figure 186 **Core Diversity Network — Base OSPF**



In this case, assume that the shortest path between PE-1 and PE-2 is the path via P-5 and P-8.

Similarly, Figure 187 shows the extent of OSPF instance 1.

Figure 187 **Core Diversity Network - OSPF Instance 1**



The only path available between PE-1 and PE-2 is now completely diverse from the shortest path advertised between the same pair of PEs in the base OSPF instance.

Therefore, for any MDT to be signaled across the OSFP 1 topology, only addresses advertised within OSPF 1 must be used. As previously stated, the system address is used as the default MDT source address. This system address is not advertised within the OSPF 1 topology, so a replacement /32 loopback address is used as the default MDT source address within OSPF 1.

VPN-IPv4 routes that may represent a customer multicast source address should be reachable via the default MDT. In the non-default topology, the c-multicast signaling across the MVPN must resolve the c-multicast route via the MDT, which has its root at the non-default /32 loopback. Therefore, the VPN-IPv4 prefix representing the possible source routes needs to be advertised containing the non-default /32 loopback.

This can be achieved in one of two ways:

1. Use a route policy at the advertising PE that changes the BGP next hop to match the MDT source address for non-default topology MDTs.

2. Use the BGP connector attribute for all VPN-IPv4 route prefixes within a multicast VPRN that has auto-discovery set to MDT-SAFI. The connector attribute contains the MDT source address within the originator field.

This chapter describes the use of the connector attribute.

If the default IGP instance is used, the BGP next hop of the VPN-IPv4 route matches the source address of the default MDT.

Therefore, if a second /32 loopback is used that replaces the system address as MDT source address and also as the next hop for source address RPF look-up, the loopback could be advertised within the non-default IGP instance, and the paths between the PEs would follow this topology.

Core diversity is achieved by configuring the following steps:

1. Configuring multiple OSPF instances, as shown in Figure 186 and Figure 187, and including the appropriate interfaces. This includes a separate loopback address per instance.
2. Configuring separate VPRNs with their own MDTs using BGP MDT-SAFI auto-discovery and PIM signaling across the appropriate PEs.
3. Configuring the VPRN that uses the base OSPF instance to use the system address as the source addresses for the MDTs (this is default behavior).
4. Configuring a separate loopback (/32) address that is advertised within OSFP instance 1 only.
5. Configuring the VPRN that uses the OSPF instance 1 to use the separate loopback system address as the source addresses for the MDTs.
6. Ensuring the unicast route that represents the c-source address is advertised as a VPN-IPv4 route and has a BGP connector attribute that contains an address that matches the MDT source address of the originating PE.

Configuration

The following configuration tasks must be completed as a prerequisite:

- Full mesh OSPF base instance between each of the nodes. However, IS-IS could also be used for any or all of the IGP instances.
- Link-layer LDP between each P and PE router.
- PIM enabled on each router network interface.

Global BGP Configuration

The first step is to configure an iBGP session between each of the PEs and the route reflector (P-5) shown in Figure 185. The address families negotiated between the iBGP peers are **vpn-ipv4**, for unicast routing, and **mdt-safi** for multicast routing.

The iBGP configuration for PE-1 is the following:

```
# on PE-1
configure router
        bgp
            group "INTERNAL"
                family vpn-ipv4 mdt-safi
                type internal
                neighbor 192.0.2.5
                exit
            exit
```

The configuration for the other PE nodes is the same.

P-5 is the route reflector for PE-1, PE-2, and PE-3, as follows:

```
# on P-5
configure router
        bgp
            cluster 0.0.0.1
            group "RR_CLIENTS"
                family vpn-ipv4 mdt-safi
                type internal
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
            exit
```

On PE-1, verify that the BGP session with the route reflector at P-5 is established with address families **mdt-safi** and **vpn-ipv4** capabilities negotiated:

```
*A:PE-1# show router bgp summary
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
BGP Admin State         : Up          BGP Oper State              : Up
Total Peer Groups       : 1           Total Peers                 : 1
Total VPN Peer Groups   : 0           Total VPN Peers             : 0
Total BGP Paths         : 20          Total Path Memory           : 5280
--- snipped ---
===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
                   AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.5
               64496        3    0 00h00m29s 0/0/0 (VpnIPv4)
                            3    0           0/0/0 (MdtSafi)
-------------------------------------------------------------------------------
*A:PE-1#
```

Configuring VPRN on PEs

There are two VPRNs:

- VPRN 1 using the base instance OSPF topology. This is present on PE-1, PE-2, and PE-3.
- VPRN 2 using OSPF instance 1. This is present on PE-1 and PE-2.

The following output displays the configuration for VPRN 1 for the sender PE-1.

```
# on PE-1
configure service
        vprn 1 customer 1 create
            route-distinguisher 64496:1
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64496:1
            interface "int-PE-1-S-1" create
                address 172.16.11.1/24
                sap 1/1/3 create
                exit
            exit
            pim
                apply-to all
                no shutdown
            exit
            mvpn
                auto-discovery mdt-safi
                provider-tunnel
                    inclusive
                        pim ssm 239.160.1.1
                        exit
                    exit
                exit
                vrf-target unicast
                exit
            exit
            no shutdown
```

There is a single interface toward S-1, from which the multicast group is received.

PIM is enabled and applied to all interfaces.

The MVPN configuration enables BGP MDT-SAFI as the auto-discovery mechanism. The provider tunnels between the PEs within the MVPN are PIM SSM multicast data trees with a group address of 239.160.1.1.

The configuration for VPRN 1 for the receiver PE-2 is the following.

```
# on PE-2
configure
    service
        vprn 1 customer 1 create
            route-distinguisher 64496:1
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64496:1
            interface "int-PE-2-H-1" create
                address 172.16.21.1/24
                sap 1/1/3 create
                exit
            exit
            igmp
                interface "int-PE-2-H-1
                    no shutdown
                exit
                no shutdown
            exit
            pim
                apply-to all
                no shutdown
            exit
            mvpn
                auto-discovery mdt-safi
                provider-tunnel
                    inclusive
                        pim ssm 239.160.1.1
                        exit
                    exit
                exit
                vrf-target unicast
                exit
            exit
            no shutdown
```

The configuration for VPRN 1 for receiver PE-3 is as follows.

```
# on PE-3
configure
    service
         vprn 1 customer 1 create
```

```
route-distinguisher 64496:1
auto-bind-tunnel
    resolution-filter
        ldp
    exit
    resolution filter
exit
vrf-target target:64496:1
interface "int-PE-3-H-3" create
    address 172.16.33.1/24
    sap 1/1/3 create
    exit
exit
igmp
    interface "int-PE-3-H-3"
        no shutdown
    exit
    no shutdown
exit
pim
    apply-to all
    no shutdown
exit
mvpn
    auto-discovery mdt-safi
    provider-tunnel
        inclusive
            pim ssm 239.160.1.1
            exit
        exit
    exit
    vrf-target unicast
    exit
exit
no shutdown
```

At PE-2, the MDT SAFI NLRI advertised by PE-1 is as follows:

```
*A:PE-2# show router bgp routes mdt-safi grp-address 239.160.1.1 source-ip 192.0.2.1
detail
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MDT-SAFI Routes
===============================================================================
Original Attributes

Route Dist.    : 64496:1
Source Addr    : 192.0.2.1
Group Addr     : 239.160.1.1
Nexthop        : 192.0.2.1
From           : 192.0.2.5
```

```
Res. Nexthop   : 0.0.0.0
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : target:64496:1
Cluster        : 0.0.0.1
Originator Id  : 192.0.2.1              Peer Router Id : 192.0.2.5
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h05m05s

Modified Attributes
 --- snipped ---
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-2#
```

The source and group address is used by PE-2 (and PE-3) to join the MDT that has its root at PE-1. The source address used is the system address of PE-1.

Examining the MDTs for this VPRN at PE-1 shows the state, as follows:

```
*A:PE-1# show router pim group 239.160.1.1

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
239.160.1.1                 (S,G)               spt      system         2
   192.0.2.1
239.160.1.1                 (S,G)               spt      int-PE-1-P-5   1
   192.0.2.2
239.160.1.1                 (S,G)               spt      int-PE-1-P-5   1
   192.0.2.3
-------------------------------------------------------------------------------
Groups : 3
===============================================================================
*A:PE-1#
```

The MDT with the root of its tree at PE-1 is as follows:

```
*A:PE-1# show router pim group 239.160.1.1 detail source 192.0.2.1

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 239.160.1.1
Source Address     : 192.0.2.1
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              : spt                Type              : (S,G)
Mode               : sparse
MRIB Next Hop      :
MRIB Src Flags     : self
Keepalive Timer Exp: 0d 00:03:28
Up Time            : 0d 00:06:36        Resolved By       : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:24
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       :
Incoming Intf      : system
Outgoing Intf List : system, int-PE-1-P-5

Curr Fwding Rate   : 0.0 kbps
Forwarded Packets  : 18                 Discarded Packets  : 0
Forwarded Octets   : 1404               RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-1#
```

The source address of the tree is the system address of the router, which is determined from the MDT SAFI NLRI that is advertised to all other PEs via the route reflector. Also, the outgoing interface list contains an interface (int-PE-1-P-5) that is OSPF enabled, and advertised within the base OSPF instance.

From the MDT on PE-2, which has its root on PE-1, the incoming interface is an OSPF interface advertised in the base OSPF instance, as shown.

```
*A:PE-2# show router pim group 239.160.1.1 detail source 192.0.2.1

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 239.160.1.1
Source Address     : 192.0.2.1
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              : spt                Type              : (S,G)
Mode               : sparse
MRIB Next Hop      : 192.168.28.2
MRIB Src Flags     : remote
```

```
Keepalive Timer Exp: 0d 00:03:24
Up Time            : 0d 00:06:09        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:50
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 192.168.28.2
Incoming Intf      : int-PE-2-P-8
Outgoing Intf List : system

Curr Fwding Rate   : 0.0 kbps
Forwarded Packets  : 14                 Discarded Packets  : 0
Forwarded Octets   : 1092               RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-2#
```

The incoming interface shown is “int-PE-2-P-8”. Similarly for PE-3, the incoming interface is “int-PE-3-P-9”.

```
*A:PE-3# show router pim group 239.160.1.1 detail source 192.0.2.1
===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 239.160.1.1
Source Address     : 192.0.2.1
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              : spt                Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 192.168.39.2
MRIB Src Flags     : remote
Keepalive Timer Exp: 0d 00:03:21
Up Time            : 0d 00:05:34        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:26
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 192.168.39.2
Incoming Intf      : int-PE-3-P-9
Outgoing Intf List : system
Curr Fwding Rate   : 0.0 kbps
Forwarded Packets  : 11                 Discarded Packets  : 0
Forwarded Octets   : 858                RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
Groups : 1
===============================================================================
*A:PE-3#
```

VPRN Using Non-Default IGP Instance

A VPRN instance is configured on each of PE-1 and PE-2 that uses an MDT topology governed by the non-default OSPF instance.

Additional interfaces need to be configured.

```
# on PE-1
configure
    router
        interface "int-PE-1-P-6a"
            address 192.168.116.1/30
            port 1/1/4
        exit
        interface "loop-1"
            address 192.0.3.1/32
            loopback
        exit
```

There are parallel links between PE-1 and P-6. The interface name of the second link contains the suffix **a.**

In a Rosen MVPN, each PE constructs a default MDT to all other PEs in the multicast VPN domain, as defined by the MDT SAFI BGP update received. The MDT update contains the source address of the MDT to which each PE should join.

When each of the other PEs receives the MDT SAFI network layer reachability information (NLRI), a PIM join is sent to the source address within the global PIM routing instance to create the MDT.

The MDT source address is usually the system address. Because the system address is advertised in the base instance of OSPF, another /32 address must be used as the source address for the default MDT. Therefore, a second loopback address is configured and used as the source address for the default MDT. For PE-1, this interface is called loop-1, and it is advertised in OSPF 1.

The interface loop-1 will be used as the source address for the MDTs and the next hop for the unicast route representing the source address of the c-multicast group.

The non-default OSPF instance for PE-1 is configured as follows, where 192.0.3.1 is the OSPF 1 router-ID. The router ID need not be equal to the IP address for loop-1, but in this case it is.

```
# on PE-1
configure
    router
        ospf 1 192.0.3.1
            area 0.0.0.0
                interface "int-PE-1-P-6a"
                    interface-type point-to-point
                exit
                interface "loop-1"
                exit
            exit
```

LDP is also required for BGP next-hop resolution and is configured as follows for PE-1.

```
# on PE-1
configure
    router
        ldp
            interface-parameters
                interface int-PE-1-P-6a
                    ipv4
                        transport-address interface
                    exit
                exit
            exit
```

The transport address is set to interface, rather than the default of system address; this is because the system address is not reachable within OSPF instance 1.

For completeness, the configuration of the additional interfaces, OSPF instance 1 and LDP of PE-2 is shown in the following three outputs.

```
# on PE-2
configure
    router
        interface "int-PE-2-P-7a"
            address 192.168.127.1/30
            port 1/1/4
        exit
        interface "loop-1"
            address 192.0.3.2/32
            loopback
        exit
```

The OSPF 1 instance configuration is as follows:

```
# on PE-2
configure
    router
        ospf 1 192.0.3.2
            area 0.0.0.0
                interface "int-PE-2-P-7a"
                    interface-type point-to-point
                exit
                interface "loop-1"
                exit
            exit
            no shutdown
        exit
```

The LDP configuration is as follows:

```
# on PE-2
configure
    router
        ldp
            interface-parameters
                interface "int-PE-2-P-7a"
                    ipv4
                        transport-address interface
                    exit
                exit
            exit
```

PIM needs to be enabled on all interfaces.

The MDT source address for VPRN 2 is the loop-1 address. Each PE within this VPRN has to join the MDT sourced at PE-1, so the MDT SAFI NLRI must advertise the source address of the MDT group as loop-1. This is achieved by specifying the MDT SAFI source address within the MVPN context. The following output displays the VPRN configuration for PE-1.

```
# on PE-1
configure
    service
         vprn 2 customer 1 create
            route-distinguisher 64496:2
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64496:2
            interface "int-PE-1-S-2" create
                address 172.16.12.1/24
                sap 1/2/1 create
                exit
            exit
            pim
                apply-to all
                no shutdown
            exit
            mvpn
                auto-discovery mdt-safi source-address 192.0.3.1
                provider-tunnel
                    inclusive
                        pim ssm 239.160.2.1
                        exit
                    exit
                exit
                vrf-target target:64496:2
                exit
            exit
            no shutdown
```

The MDT SAFI source address modification is only required on PEs that use the non-default /32 addresses. The system address must not be explicitly configured as the MDT source address for MVPNs that use the default IGP instance. As previously stated, only three MVPNs can be used to create core diversity, one of which must be the default instance. Configuring the system address as a source address prevents the creation of a third MVPN because only two MVPNs are allowed to use explicitly configured MDT source addresses.

Verification of Core Diversity

The MDT SAFI NLRI advertised by the PE-1 sender router contains the following information.

```
*A:PE-1# show router bgp routes mdt-safi hunt rd 64496:2 | match "RIB Out" post-lines
25 pre-lines 1
-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
Route Dist.    : 64496:2
Source Addr    : 192.0.3.1
Group Addr     : 239.160.2.1
Nexthop        : 192.0.2.1
To             : 192.0.2.5
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 0
AIGP Metric    : None
Connector      : None
Community      : target:64496:2
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.5
Origin         : IGP
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0

-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:PE-1#
```

The source address is set to 192.0.3.1, which is the address of the loopback address used in the non-default OSPF instance 1 of PE-1.

The following output shows the MDT that has its root at PE-1, and that the source address is set to 192.0.3.1. The outgoing interface list includes the router interface contained within the OSPF 1 instance, proving that the non-default OSPF instance is used.

```
*A:PE-1# show router pim group 239.160.2.1 source 192.0.3.1 detail
===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 239.160.2.1
Source Address     : 192.0.3.1
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              : spt                Type               : (S,G)
Mode               : sparse
MRIB Next Hop      :
MRIB Src Flags     : self
Keepalive Timer Exp: 0d 00:03:15
Up Time            : 0d 00:04:49        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:11
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       :
Incoming Intf      : loop-1
Outgoing Intf List : system, int-PE-1-P-6a

Curr Fwding Rate   : 0.0 kbps
Forwarded Packets  : 13                 Discarded Packets  : 0
Forwarded Octets   : 1014               RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-1#
```

The PIM interfaces within VPRN 2 are now present on PE-1, as follows:

```
*A:PE-1# show router 2 pim interface
===============================================================================
PIM Interfaces ipv4
===============================================================================
Interface                  Adm  Opr  DR Prty         Hello Intvl  Mcast Send
   DR
-------------------------------------------------------------------------------
int-PE-1-S-2               Up   Up   1               30           auto
   172.16.12.1
2-mt-239.160.2.1           Up   Up   1               N/A          auto
   192.0.3.2
-------------------------------------------------------------------------------
Interfaces : 2 Tunnel-Interfaces : 0
===============================================================================
*A:PE-1#
```

Likewise, for PE-2, the PIM interfaces within VPRN 2 are displayed, as follows:

```
*A:PE-2# show router 2 pim interface
===============================================================================
PIM Interfaces ipv4
===============================================================================
Interface                   Adm  Opr  DR Prty         Hello Intvl  Mcast Send
   DR
-------------------------------------------------------------------------------
int-PE-2-H-2                Up   Up   1               30           auto
   172.16.22.1
2-mt-239.160.2.1            Up   Up   1               N/A          auto
   192.0.3.2
-------------------------------------------------------------------------------
Interfaces : 2 Tunnel-Interfaces : 0
===============================================================================
*A:PE-2#
```

Within the VPRN, there are PIM neighbors shown via the MDT. On PE-2, the PIM neighbor is 192.0.3.1, as follows:

```
*A:PE-2# show router 2 pim neighbor
===============================================================================
PIM Neighbor ipv4
===============================================================================
Interface               Nbr DR Prty     Up Time       Expiry Time     Hold Time
   Nbr Address
-------------------------------------------------------------------------------
2-mt-239.160.2.1        1               0d 00:04:10   0d 00:01:35     105
   192.0.3.1
-------------------------------------------------------------------------------
Neighbors : 1
===============================================================================
*A:PE-2#
```

The PIM interface on PE-2 designated as 2-mt-239.160.2.1 with a neighbor address of 192.0.3.1 is the MDT interface toward PE-1.

The prefix that represents the source address on PE-1 is advertised as a VPN-IPv4 route, which contains a BGP connector attribute.

This can be shown when the VPN-IPv4 route is examined on PE-2, as follows:

```
*A:PE-2# show router bgp routes 172.16.12.0/24 vpn-ipv4 hunt
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : 172.16.12.0/24
Nexthop        : 192.0.2.1
Route Dist.    : 64496:2                VPN Label      : 262133
Path Id        : None
From           : 192.0.2.5
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : int-PE-2-P-8
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : RD 64496:2, Originator 192.0.3.1
Community      : target:64496:2
Cluster        : 0.0.0.1
Originator Id  : 192.0.2.1              Peer Router Id : 192.0.2.5
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h04m34s
VPRN Imported  :  2

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-2#
```

The originator value within the connector attribute is shown to be 192.0.3.1, which is the same as the MDT source address of PE-1. The BGP next hop is still set to the system address of PE-1, so the unicast route can still be resolved via an LDP tunnel.

PIM will now resolve the c-source address RPF using the originator value within the connector attribute.

Similarly, for VPRN 1, the route on PE-1 representing the source address is also advertised as a VPN-IPv4 address that contains a BGP connector attribute.

```
*A:PE-2# show router bgp routes 172.16.11.0/24 vpn-ipv4 hunt
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
```

```
===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : 172.16.11.0/24
Nexthop        : 192.0.2.1
Route Dist.    : 64496:1                VPN Label      : 262134
Path Id        : None
From           : 192.0.2.5
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : int-PE-2-P-8
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : RD 64496:1, Originator 192.0.2.1
Community      : target:64496:1
Cluster        : 0.0.0.1
Originator Id  : 192.0.2.1              Peer Router Id : 192.0.2.5
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h04m44s
VPRN Imported  :  1

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-2#
```

Verification of Multicast Traffic

An IGMPv3 query is initiated from all 3 hosts: H-1, H-2, and H-3 in Figure 1, and the multicast streams from S-1 and S-2 into interfaces on the two VPRNs are enabled.

Consider VPRN 1, which uses the default topology. On PE-1, the group 239.160.1.123 can be shown. The outgoing and incoming interface lists are populated, with the outgoing interface being the MDT interface for the VPRN:

```
*A:PE-1# show router 1 pim group detail
===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 239.160.1.123
```

```
Source Address     : 172.16.11.2
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 172.16.11.2
MRIB Src Flags     : direct
Keepalive Timer    : Not Running
Up Time            : 0d 00:02:06        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:00
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 172.16.11.2
Incoming Intf      : int-PE-1-S-1
Outgoing Intf List : 1-mt-239.160.1.1

Curr Fwding Rate   : 1018.6 kbps
Forwarded Packets  : 3035               Discarded Packets  : 0
Forwarded Octets   : 4546430            RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-1#
```

The same groups can be shown within VPRN 1 on PE-2.

```
*A:PE-2# show router 1 pim group detail
===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 239.160.1.123
Source Address     : 172.16.11.2
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 192.0.2.1
MRIB Src Flags     : remote
Keepalive Timer    : Not Running
Up Time            : 0d 00:02:10        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:29
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 192.0.2.1
Incoming Intf      : 1-mt-239.160.1.1
Outgoing Intf List : int-PE-2-H-1

Curr Fwding Rate   : 1024.6 kbps
```

```
Forwarded Packets  : 3337              Discarded Packets  : 0
Forwarded Octets   : 4998826           RPF Mismatches     : 0
Spt threshold      : 0 kbps            ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-2#
```

The MDT is now the incoming interface with an upstream RPF neighbor of 192.0.2.1, the system address of PE-1. Similarly for PE-3:

```
*A:PE-3# show router 1 pim group detail
===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 239.160.1.123
Source Address     : 172.16.11.2
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              :                   Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 192.0.2.1
MRIB Src Flags     : remote
Keepalive Timer    : Not Running
Up Time            : 0d 00:01:50       Resolved By        : rtable-u

Up JP State        : Joined            Up JP Expiry       : 0d 00:00:10
Up JP Rpt          : Not Joined StarG  Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 192.0.2.1
Incoming Intf      : 1-mt-239.160.1.1
Outgoing Intf List : int-PE-3-H-3

Curr Fwding Rate   : 1018.6 kbps
Forwarded Packets  : 3707              Discarded Packets  : 0
Forwarded Octets   : 5553086           RPF Mismatches     : 0
Spt threshold      : 0 kbps            ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-3#
```

Consider VPRN 2, which uses the non-default topology. On PE-1, the group 239.160.2.123 can be shown. The outgoing and incoming interface lists are populated, with the outgoing interface being the MDT interface for the VPRN.

```
*A:PE-1# show router 2 pim group detail
===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 239.160.2.123
```

```
Source Address     : 172.16.12.2
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 172.16.12.2
MRIB Src Flags     : direct
Keepalive Timer    : Not Running
Up Time            : 0d 00:02:25        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:00
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 172.16.12.2
Incoming Intf      : int-PE-1-S-2
Outgoing Intf List : 2-mt-239.160.2.1

Curr Fwding Rate   : 1018.6 kbps
Forwarded Packets  : 12308              Discarded Packets  : 0
Forwarded Octets   : 18437384           RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-1#
```

The outgoing interface list is again populated with the MDT being the interface. This MDT is encapsulated in the multicast tree shown in the global PIM context as multicast group 239.160.2.1 with source address 192.0.3.1. This can be shown to have an outgoing interface list containing the interface int-PE-1-P-6a, which is an OSPF 1 interface and was shown in a preceding output.

```
*A:PE-1# show router pim group detail 239.160.2.1
===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 239.160.2.1
Source Address     : 192.0.3.1
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              : spt                Type               : (S,G)
Mode               : sparse
MRIB Next Hop      :
MRIB Src Flags     : self
Keepalive Timer Exp: 0d 00:03:25
Up Time            : 0d 00:11:39        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:21
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No
```

```
Rpf Neighbor       :
Incoming Intf      : loop-1
Outgoing Intf List : system, int-PE-1-P-6a

Curr Fwding Rate   : 1018.6 kbps
Forwarded Packets  : 13184             Discarded Packets  : 0
Forwarded Octets   : 19712712          RPF Mismatches     : 0
Spt threshold      : 0 kbps            ECMP opt threshold : 7
Admin bandwidth    : 1 kbps

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 239.160.2.1
Source Address     : 192.0.3.2
RP Address         : 0
Advt Router        : 192.0.3.2
Flags              : spt               Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 192.168.116.2
MRIB Src Flags     : remote
Keepalive Timer Exp: 0d 00:03:12
Up Time            : 0d 00:10:52       Resolved By        : rtable-u

Up JP State        : Joined            Up JP Expiry       : 0d 00:00:07
Up JP Rpt          : Not Joined StarG  Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 192.168.116.2
Incoming Intf      : int-PE-1-P-6a
Outgoing Intf List : system

Curr Fwding Rate   : 0.0 kbps
Forwarded Packets  : 27                Discarded Packets  : 0
Forwarded Octets   : 2106              RPF Mismatches     : 0
Spt threshold      : 0 kbps            ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 2
===============================================================================
*A:PE-1#
```

Conclusion

MVPN Core Diversity allows service providers to provide separation in terms of topology between content providers that use a core network to provide transport between source and receivers in a multicast VPN. This chapter provides the configuration for multiple instances of OSPF which, together with the associated commands and outputs, can be used for verifying and troubleshooting.

Rosen MVPN Inter-AS Option B

This chapter provides information about Rosen MVPN: Inter-AS Option B configurations.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

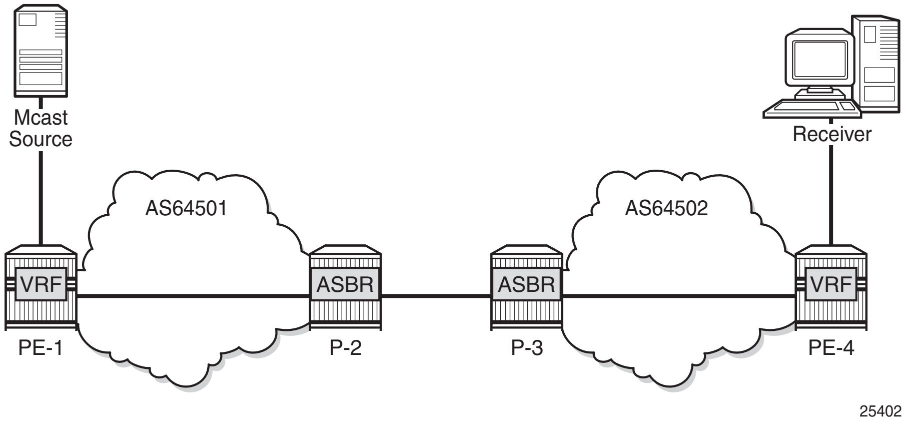
Applicability

This chapter was initially written for SR OS 11.0.R3. The CLI in the current edition is based on SR OS release 15.0.R5. Knowledge of the Nokia multicast and Layer 3 VPNs concepts are assumed throughout this document.

Overview

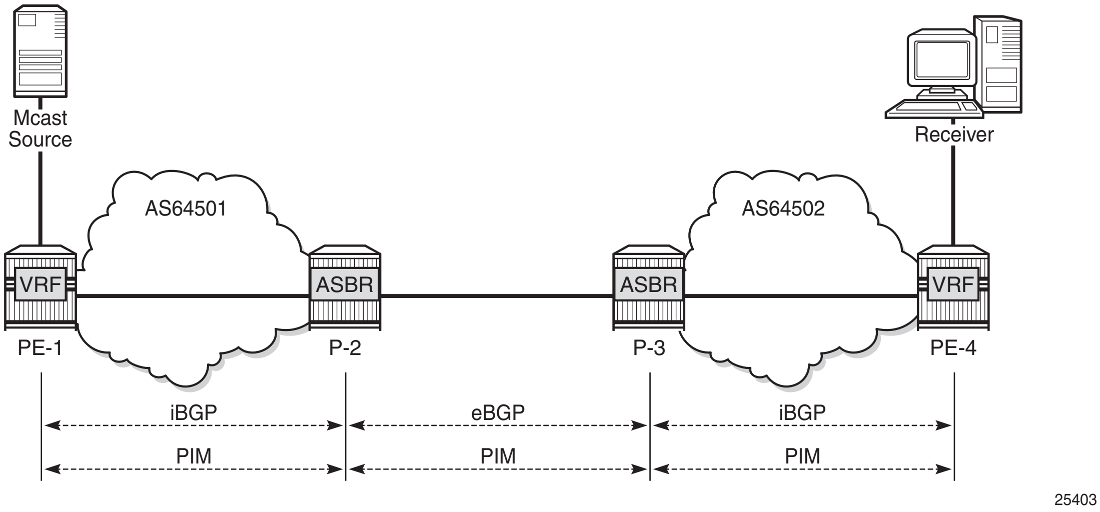
This chapter covers a basic technology overview, the network topology and configuration examples which are used for multicast virtual private network (MVPN) inter-autonomous system (AS) option B.The Inter-AS MVPN feature allows the setup of multicast distribution trees (MDTs) spanning multiple autonomous systems.

Figure 188 **General Topology for Inter-AS MVPN**



This chapter covers Rosen MVPN Inter-AS support (Option-B). Inter-AS Option B is supported for protocol independent multicast (PIM) source-specific multicast (SSM) with Rosen MVPN using multicast distribution tree (MDT) subsequent address family indicator (SAFI), the border gateway protocol (BGP) connector attribute and PIM reverse path forwarding (RPF) vector.

Figure 189 **Protocols Used for Inter-AS MVPN**



The following assumptions are made:

- PE-1 is the sender PE because the multicast source is directly connected to this router.

- PE-4 is the receiver PE because the multicast receiver is directly connected to this router.
- P-2 and P-3 are ASBR routers according to the Inter-AS model.

The multicast receiver and source can be indirectly connected to PE routers via CE routers, but for the core multicast distribution, these variations are conceptually the same. For simplicity, the PE and P router configurations will be provided.

There are several challenges which have to be solved in order to make the complete inter-AS solution operational:

Challenge 1:

In case of Inter-AS MVPN Option B, routing information toward the source PE is not available in a remote AS domain because IGP routes are not exchanged between ASs.

As a result, a PIM-P join would never be sent upstream (from the receiver PE to the sender PE in a different AS). However, the PIM-P join has to be propagated from PE-4 to PE-1. Therefore, a solution is required to issue PIM-P join and perform RPF.

Solution:

Use a PIM reverse path forwarding (RPF) vector (RPFV) to propagate PIM-P over multiple segments. In this example there are three segments:

- PE-4 -> ASBR P-3
- ASBR P-3 -> ASBR P-2
- ASBR P-2 -> PE-1

The RPF vector is added to a PIM join by the PE router when the following option is enabled:

```
*A:PE-4# configure router pim rpfv
  - no rpfv [core] [mvpn]
  - rpfv core mvpn
  - rpfv core
  - rpfv mvpn

 <core>               : Proxy RPF vector for core
 <mvpn>               : Proxy RPF vector for inter-AS rosen mvpn

*A:PE-4#
```

The **mvpn** keyword enables “mvpn RPF vector” processing for Inter-AS Option B MVPN based on RFC 5496 and RFC 6513. If a core RPF vector is received, it will be dropped before a message is processed.

All routers on the multicast traffic transport path must have this option enabled to allow RPF vector processing. If the option is not enabled, the RPF vector is dropped and the PIM join is processed as if the PIM vector is not present.

Details about RPF Vector can be found in the following RFCs: 5496, 5384, 6513.

Challenge 2:

With Inter-AS MVPN Option B, the BGP next-hop is modified by the local and remote ASBRs during re-advertisement of VPN IPv4 routes. When the BGP next-hop is changed, information regarding the originator of the prefix is lost when the advertisement reaches the receiver PE node. Therefore, a solution is required to do a successful RPF check for the VPN source at receiver VPRN.

This challenge does not apply to Model C because in Model C the BGP next-hop for VPN routes is not updated.

Solution:

The transitive BGP connector attribute is added and used to advertise an address of a sender PE node which is carried inside a VPN IPv4 update. The BGP connector attribute allows the sender PE address information to be available to the receiver PE so that a receiver PE is able to associate VPN IPv4 advertisement to the corresponding source PE.

Inter-AS Option B will work when the following criteria are met:

- Rosen MVPN is used with PIM SSM
- BGP MDT-SAFI address family is used
- PIM RPF vector is configured
- BGP connector attribute is used for VPN-IPv4 updates

SR OS inter-AS Option B is designed to be standard compliant based on the following RFCs:

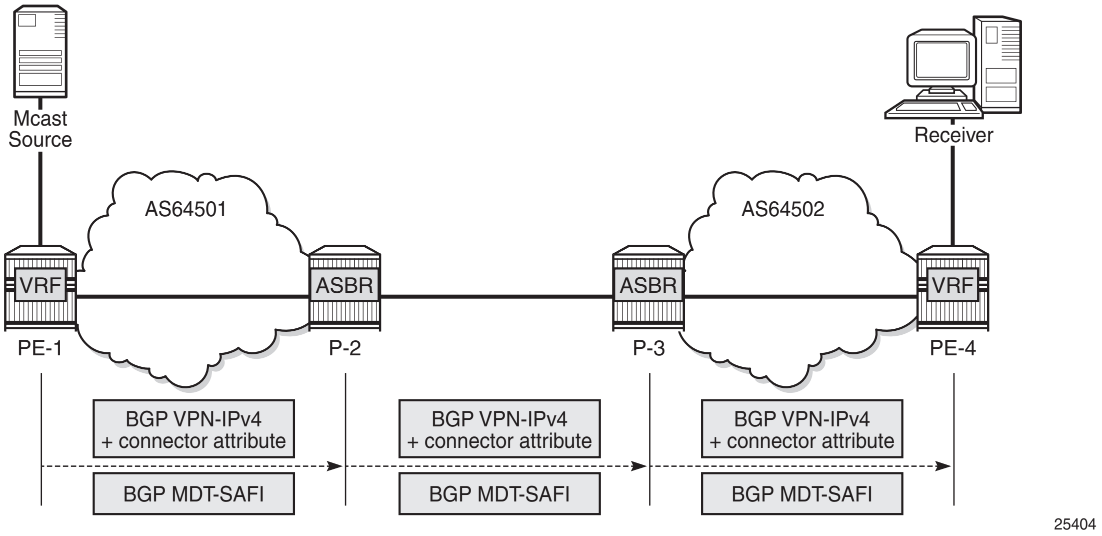
- RFC 5384, *The Protocol Independent Multicast (PIM) Join Attribute Format*
- RFC 5496, *The Reverse Path Forwarding (RPF) Vector TLV*
- RFC 6513, *Multicast in MPLS/BGP IP VPNs*

The following signaling stages can be identified when Inter-AS MVPN is configured:

- Stage 1 - BGP core signaling
- Stage 2 - Core PIM signaling
- Stage 3 - Customer PIM signalling

Stage 1 - BGP core signaling

Figure 190 **BGP Signaling Steps**



The sender PE sends VPN-IPv4 and MDT-SAFI BGP updates for this particular MVPN:

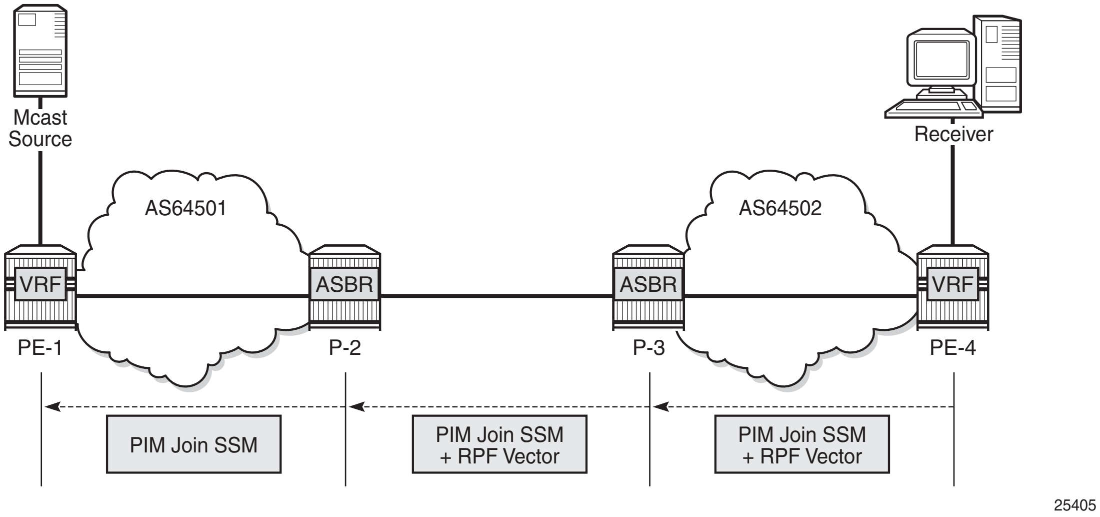
- Every ASBR propagates VPN-IPv4 and MDT-SAFI BGP updates:
  - Next hop (NH) attribute is modified every time
  - Connector attribute stays untouched

When this stage is completed, all routers have information necessary:

- to start PIM signaling in the core network (PIM-P) to prepare the default MDT
- to start PIM signaling of customer multicast streams (PIM-C) inside the VPN

Stage 2 - Core PIM signaling

Figure 191 **PIM-P Signaling Steps for Default MDT**



PE-4 determines the reverse path to the source based on the RPF vector (ASBR P-3 IP address) and not based on the IP address of the multicast source (PE-1) which is unknown to PE-4.

PE-4 inserts an RPF vector and sends a PIM-P join to the immediate next-hop to reach ASBR P-3. Intermediate P-routers (if present) do not change the RPF vector.

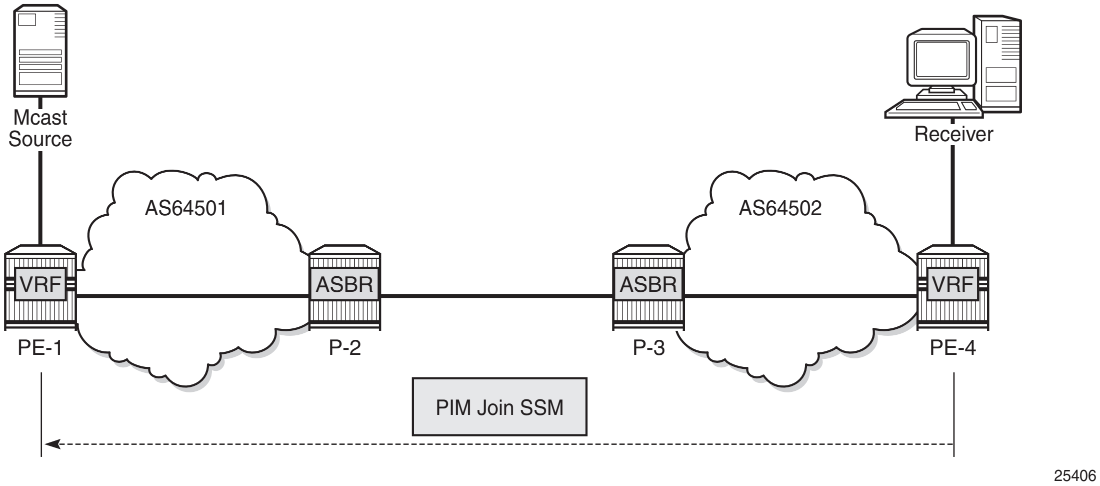
P-3 finds itself in the RPF vector and has to make a decision based on MDT-SAFI BGP table:

- P-3 determines the reverse path to the multicast source based on the RPF vector (ASBR P-2 IP address).
- If the multicast source and the NH do not match, P-3 has to use the RPFV.
- P-3 modifies the PIM-P join received from PE-4 with ASBR P-2’s IP address as the upstream (taken from next hop MDT-SAFI network layer reachability information (NLRI)).
- P-2 can match the source IP with the NH in BGP MDT-SAFI. Therefore, there is no need for the RPF vector to be used.
- P-2 removes the RPF vector and sends a normal PIM-P join toward PE-1.

When this stage is completed, the default MDT is established for this MVPN and PE routers have the necessary information to start PIM signaling inside the VPRN (PIM-C).

Stage 3 - Customer PIM signaling

Figure 192 **PIM-C Signaling**



A PIM-C join is sent to the source PE using the existing tunnel infrastructure to the RPF neighbor PE-1 provided by the BGP connector attribute of the vpn-ipv4 route of the multicast source.

When this stage is completed, the customer multicast flows throughout the network in a default MDT.

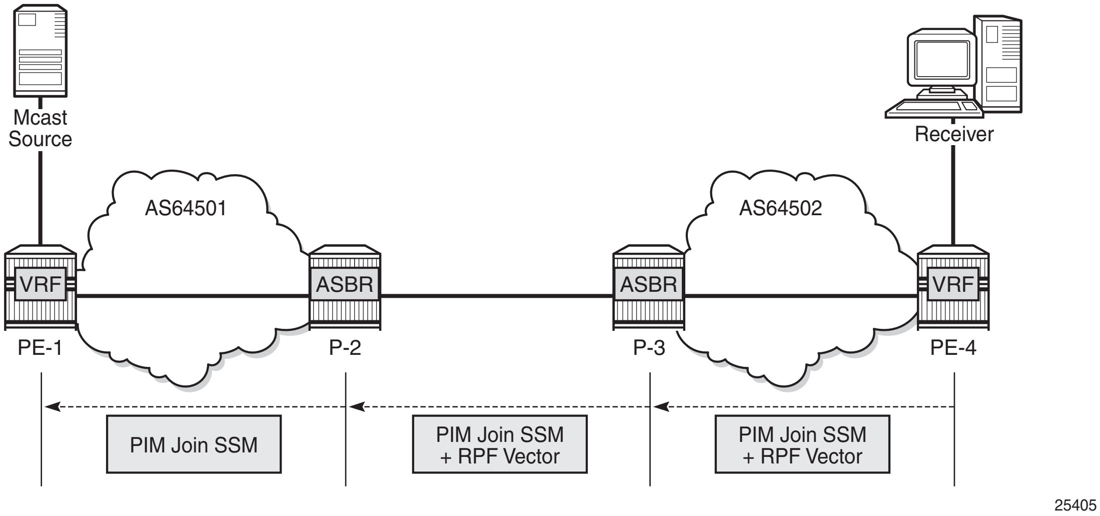
Stage 4 - The multicast stream threshold is reached.

This stage is optional and applicable when S-PMSI instance and S-PMSI threshold are configured.

The process is similar to the default MDT setup:

- PE-4 determines the reverse path to the source based on the RPF vector (ASBR P-3's IP address) and not based on the IP address of the multicast source (PE-1) which is unknown to PE-4.
- PE-4 inserts an RPF vector and sends a PIM-P Join to the immediate next hop to reach ASBR P-3.

Figure 193 **PIM-P Signaling Steps for Data MDT**



- Intermediate P-routers (if present) do not change the RPF vector.
- P-3 finds itself in the RPF vector and has to make a decision based on the MDT-SAFI BGP table:
  - P-3 determines the reverse path to the multicast source based on the RPF Vector (ASBR P-2's IP address).
  - If the multicast source and the NH do not match, P-3 has to use the RPFV.
  - P-3 modifies the PIM-P join received from PE-4 with ASBR P-2's IP address as upstream (taken from next hop MDT-SAFI NLRI).
- P-2 can match the source IP with the NH in the BGP MDT-SAFI. Therefore, there is no need for the RPF vector to be used.
- P-2 removes the RPF vector and sends a normal PIM-P join toward PE-1.

When this optional stage is completed, the customer multicast traffic flows through a dedicated Data MDT.

The SR OS implementation was also designed to interoperate with Cisco routers' Inter-AS implementations that do not fully comply with the RFC 5384 and RFC 5496.

When the following option is enabled:

```
configure router pim rpfv mvpn
```

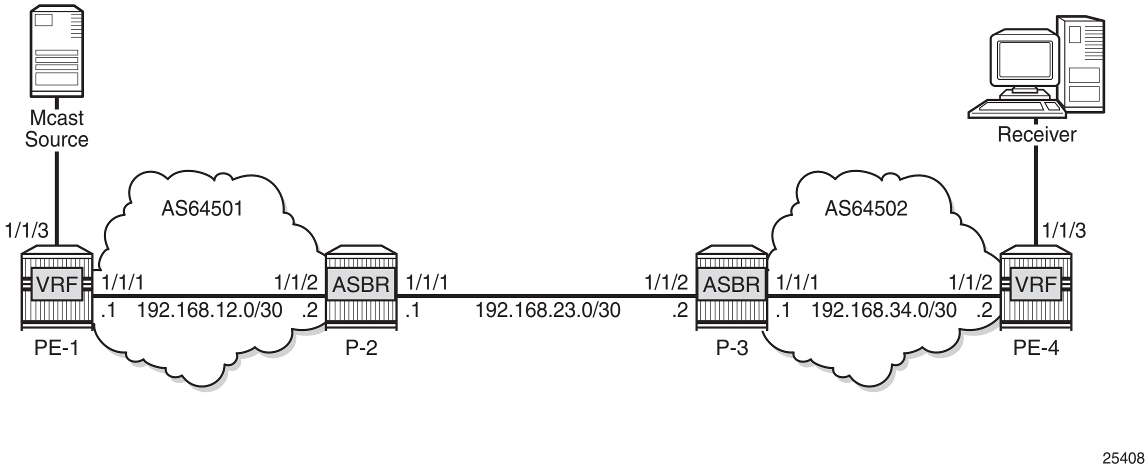
Cisco routers need to be configured to include **RD** in an RPF vector using the following command for interoperability:

```
ip multicast vrf <name> rpf proxy rd vector
```

Configuration

The example topology is shown in Figure 194.

Figure 194 **Example Topology Details**



The following components are used in the example scenario:

- VPRN 1
- Customer multicast group is 232.0.0.0/8
- Default MDT multicast group is 239.255.0.1
- Data MDT multicast group is 239.255.1.0/24
- Multicast source is 172.16.1.1
- PE-x routers have system IP addresses 192.0.2.x
- P-x routers have system IP addresses 192.0.2.x
- Interface between Router A and B has IP address 192.168.AB.x

Global BGP configuration for PE-1 router using the mdt-safi family with an iBGP neighbor to P-2. The system interface IP address is used for the iBGP session.

```
# on PE-1
configure
    router
        bgp
            group "iBGP"
                family vpn-ipv4 mdt-safi
                type internal
                neighbor 192.0.2.2
                    next-hop-self
                exit
            exit
```

The global BGP configuration for P-2 router is using the mdt-safi family with an iBGP neighbor to PE-1 and an eBGP neighbor to P-3. The system interface IP address is used for the iBGP session and the network interface IP address is used for the eBGP session.

```
# on P-2
configure
    router
        bgp
            enable-inter-as-vpn
            group "eBGP"
                family vpn-ipv4 mdt-safi
                neighbor 192.168.23.2
                    type external
                    peer-as 64502
                exit
            exit
            group "iBGP"
                family vpn-ipv4 mdt-safi
                neighbor 192.0.2.1
                    next-hop-self
                    type internal
                exit
            exit
        exit
```

The global BGP configuration for the router P-3 is using the mdt-safi family with an iBGP neighbor to PE-4 and an eBGP neighbor to P-2. The system interface IP address is used for the iBGP session and the network interface IP address is used for the eBGP session.

```
# on P-3
configure
    router
        bgp
            enable-inter-as-vpn
            group "eBGP"
                family vpn-ipv4 mdt-safi
                neighbor 192.168.23.1
                    type external
                    peer-as 64501
                exit
            exit
            group "iBGP"
                family vpn-ipv4 mdt-safi
                neighbor 192.0.2.4
                    next-hop-self
                    type internal
                exit
            exit
        exit
```

The global BGP configuration for router PE-4 is using the mdt-safi family with an iBGP neighbor to P-3 is as follows. The system interface IP address is used for the iBGP session.

```
# on PE-4
configure
    router
        bgp
            group "iBGP"
                family vpn-ipv4 mdt-safi
                type internal
                neighbor 192.0.2.3
                    next-hop-self
                exit
            exit
        exit
```

The global PIM configuration for all routers is as follows:.

```
# on all routers
configure
    router
        pim
            rpf-table both
            apply-to non-ies
            rpfv mvpn
        exit
    exit
```

The VPRN configuration for the PE routers is as follows:

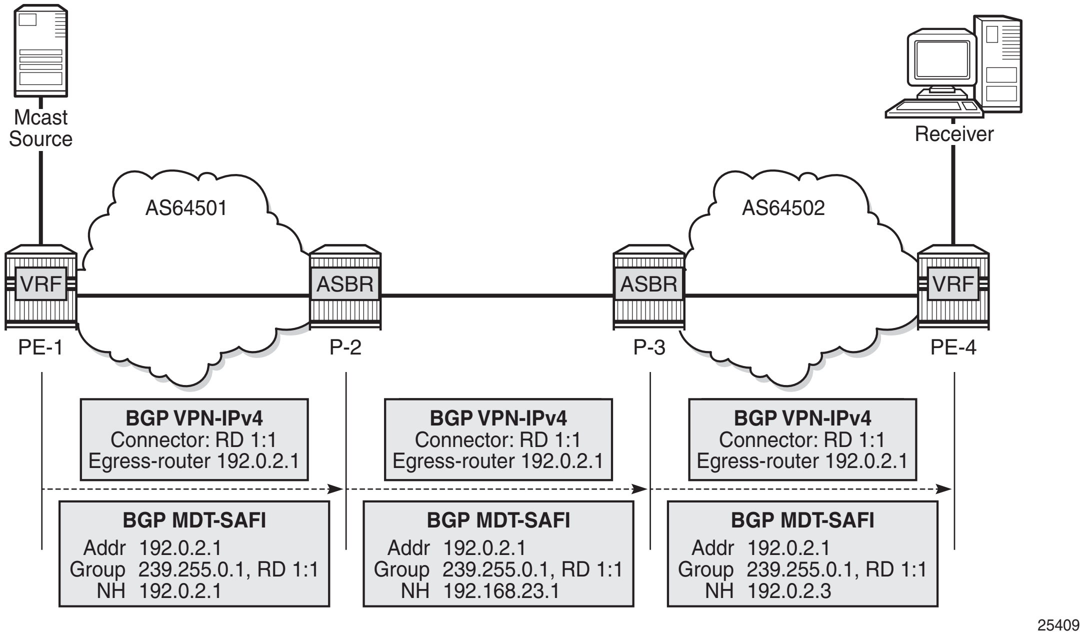
```
# on PE-1
configure
    service
        vprn 1 customer 1 create
            route-distinguisher 1:1
            auto-bind-tunnel
                resolution-filter
                    ldp
                    rsvp
                exit
                resolution filter
            exit
            vrf-target target:1:1
            interface "int-PE-1-S-1" create
                address 172.16.1.2/30
                sap 1/1/3 create
                exit
            exit
            pim
                apply-to all
            exit
            mvpn
                auto-discovery mdt-safi
                provider-tunnel
                    inclusive
```

```
                        pim ssm 239.255.0.1
                        exit
                    exit
                    selective
                        data-threshold 232.0.0.0/8 1
                        pim-ssm 239.255.1.0/24
                    exit
                exit
                vrf-target unicast
                exit
            exit
            no shutdown
        exit

# on PE-4
configure
    service
        vprn 1 customer 1 create
            route-distinguisher 4:1
            auto-bind-tunnel
                resolution-filter
                    ldp
                    rsvp
                exit
                resolution filter
            exit
            vrf-target target:1:1
            interface "int-PE-4-H-4" create
                address 172.16.4.1/30
                sap 1/1/3 create
                exit
            exit
            igmp
                interface "int-PE-4-H-4"
                exit
            exit
            pim
                apply-to all
            exit
            mvpn
                auto-discovery mdt-safi
                provider-tunnel
                    inclusive
                        pim ssm 239.255.0.1
                        exit
                    exit
                    selective
                        data-threshold 232.0.0.0/8 1
                        pim-ssm 239.255.1.0/24
                    exit
                exit
                vrf-target unicast
                exit
            exit
            no shutdown
        exit
```

MVPN Verification and Debugging

BGP Core Signaling

Figure 195 **BGP Signaling Steps**



On PE-1, the **debug router bgp update** output shows the BGP update messages which are sent to P-2. The VPN-IPv4 update contains a connector attribute and the MDT-SAFI update is used for signaling multicast group 239.255.0.1.

```
1 2017/10/17 11:46:31.115 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 79
    Flag: 0x90 Type: 14 Len: 33 Multiprotocol Reachable NLRI:
        Address Family VPN_IPV4
        NextHop len 12 NextHop 192.0.2.1
        172.16.1.0/30 RD 1:1 Label 262141
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:1:1
    Flag: 0xc0 Type: 20 Len: 14 Connector:
        RD 1:1, Egress-router 192.0.2.1
"
2 2017/10/17 11:46:31.115 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
```

```
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 62
    Flag: 0x90 Type: 14 Len: 26 Multiprotocol Reachable NLRI:
        Address Family MDT-SAFI
        NextHop len 4 NextHop 192.0.2.1
        [MDT-SAFI] Addr 192.0.2.1, Group 239.255.0.1, RD 1:1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x80 Type: 4 Len: 4 MED: 0
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:1:1
"
```

On P-2, the **debug router bgp update** output shows the BGP update messages which are sent to P-3. The VPN-IPv4 update contains an unmodified connector attribute and the MDT-SAFI update is used for signaling multicast group 239.255.0.1.

```
3 2017/10/17 11:46:53.793 UTC MINOR: DEBUG #2001 Base Peer 1: 192.168.23.2
"Peer 1: 192.168.23.2: UPDATE
Peer 1: 192.168.23.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 78
    Flag: 0x90 Type: 14 Len: 33 Multiprotocol Reachable NLRI:
        Address Family VPN_IPV4
        NextHop len 12 NextHop 192.168.23.1
        172.16.1.0/30 RD 1:1 Label 262141
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 6 AS Path:
        Type: 2 Len: 1 < 64501 >
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:1:1
    Flag: 0xc0 Type: 20 Len: 14 Connector:
        RD 1:1, Egress-router 192.0.2.1
"
4 2017/10/17 11:46:53.793 UTC MINOR: DEBUG #2001 Base Peer 1: 192.168.23.2
"Peer 1: 192.168.23.2: UPDATE
Peer 1: 192.168.23.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 54
    Flag: 0x90 Type: 14 Len: 26 Multiprotocol Reachable NLRI:
        Address Family MDT-SAFI
        NextHop len 4 NextHop 192.168.23.1
        [MDT-SAFI] Addr 192.0.2.1, Group 239.255.0.1, RD 1:1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 6 AS Path:
        Type: 2 Len: 1 < 64501 >
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:1:1
"
```

On P-3, the **debug router bgp update** output shows the BGP update messages which are sent to PE-4. The VPN-IPv4 update contains an unmodified connector attribute and the MDT-SAFI update is used for signaling multicast group 239.255.0.1.

```
5 2017/10/17 11:47:08.630 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.4
"Peer 1: 192.0.2.4: UPDATE
Peer 1: 192.0.2.4 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 85
    Flag: 0x90 Type: 14 Len: 33 Multiprotocol Reachable NLRI:
        Address Family VPN_IPV4
        NextHop len 12 NextHop 192.0.2.3
        172.16.1.0/30 RD 1:1 Label 262141
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 6 AS Path:
        Type: 2 Len: 1 < 64501 >
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:1:1
    Flag: 0xc0 Type: 20 Len: 14 Connector:
        RD 1:1, Egress-router 192.0.2.1
"
6 2017/10/17 11:47:08.630 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.4
"Peer 1: 192.0.2.4: UPDATE
Peer 1: 192.0.2.4 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 61
    Flag: 0x90 Type: 14 Len: 26 Multiprotocol Reachable NLRI:
        Address Family MDT-SAFI
        NextHop len 4 NextHop 192.0.2.3
        [MDT-SAFI] Addr 192.0.2.1, Group 239.255.0.1, RD 1:1
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 6 AS Path:
        Type: 2 Len: 1 < 64501 >
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        target:1:1
"
```

The BGP tables on PE-1 and PE-4 are updated accordingly. The most interesting aspect here is the MDT-SAFI routes received.

PE-4 has one MDT-SAFI update received from PE-1. The next-hop was modified according to the Option-B model.

```
*A:PE-4# show router bgp neighbor 192.0.2.3 received-routes mdt-safi
===============================================================================
 BGP Router ID:192.0.2.4        AS:64502       Local AS:64502
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MDT-SAFI Routes
===============================================================================
Flag  Network                                           LocalPref   MED
      Nexthop                Group-Addr                             Label
      As-Path
-------------------------------------------------------------------------------
```

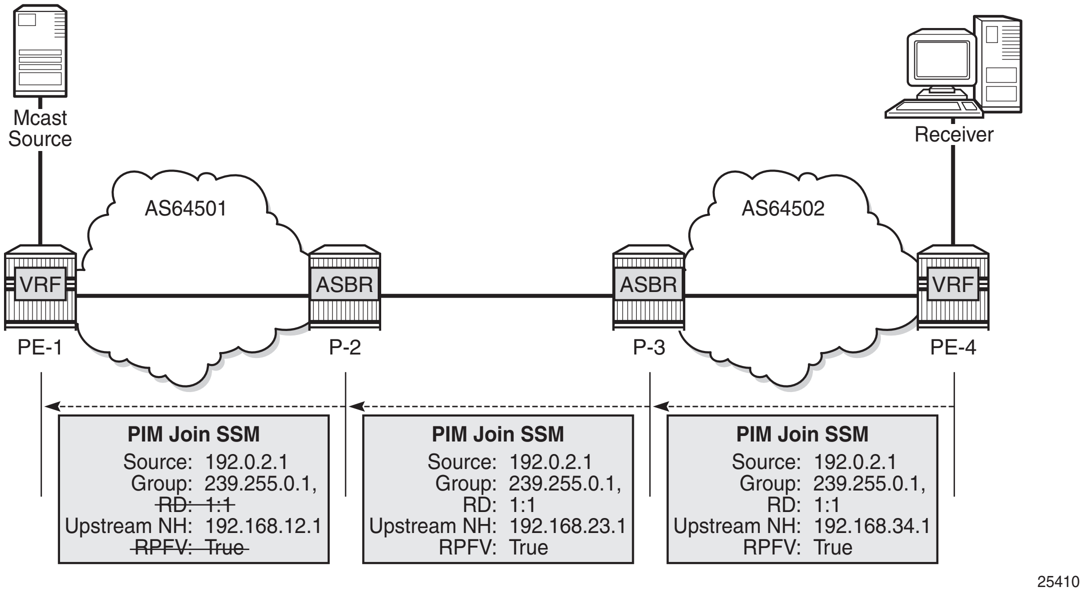
```
u*>i  1:1:192.0.2.1                                      100         None
      192.0.2.3                   239.255.0.1                        -
      64501
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-4#
```

PE-1 has one MDT-SAFI update received from PE-4. The next-hop was modified according to the Option B model.

```
*A:PE-1# show router bgp neighbor 192.0.2.2 received-routes mdt-safi
===============================================================================
 BGP Router ID:192.0.2.1        AS:64501       Local AS:64501
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP MDT-SAFI Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop                     Group-Addr                         Label
      As-Path
-------------------------------------------------------------------------------
u*>i  4:1:192.0.2.4                                      100         None
      192.0.2.2                   239.255.0.1                        -
      64502
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-1#
```

Core PIM Signaling

Figure 196 **PIM-P Signaling Steps for Default MDT**



On PE-4, the **debug router pim packet jp** output shows the PIM join/prune message which is sent to P-3. This message contains the original source of the multicast traffic (PE-1: 192.0.2.1) and the RPF Vector (P-3: 192.0.2.3).

```
5 2017/10/17 12:03:18.425 UTC MINOR: DEBUG #2001 Base PIM[Instance 1 Base]
"PIM[Instance 1 Base]: Join/Prune
[000 00:27:18.430] PIM-TX ifId 2 ifName int-PE-4-P-3 0.0.0.0 -> 224.0.0.13 Length: 48
PIM Version: 2 Msg Type: Join/Prune Checksum: 0x3d2c
Upstream Nbr IP : 192.168.34.1 Resvd: 0x0, Num Groups 1, HoldTime 210
Group: 239.255.0.1/32 Num Joined Srcs: 1, Num Pruned Srcs: 0
Joined Srcs:
192.0.2.1/32 Flag S  <S,G> JA={rpfvMvpn 192.0.2.3 1:1}
"
```

On P-3, **the debug router pim packet jp** output shows the PIM join/prune message which is propagated to P-2. The source of multicast traffic is untouched while the RPF Vector is modified for Inter-AS propagation.

```
14 2017/10/17 12:03:16.352 UTC MINOR: DEBUG #2001 Base PIM[Instance 1 Base]
"PIM[Instance 1 Base]: Join/Prune
[000 00:27:27.450] PIM-TX ifId 2 ifName int-P-3-P-2 0.0.0.0 -> 224.0.0.13 Length: 48
PIM Version: 2 Msg Type: Join/Prune Checksum: 0x3286
Upstream Nbr IP : 192.168.23.1 Resvd: 0x0, Num Groups 1, HoldTime 210
Group: 239.255.0.1/32 Num Joined Srcs: 1, Num Pruned Srcs: 0
Joined Srcs:
192.0.2.1/32 Flag S  <S,G> JA={rpfvMvpn 192.168.23.1 1:1}
"
```

On P-2, the **debug router pim packet jp** output shows the PIM join/prune message which is propagated to P-1. The source of the multicast traffic is untouched while the RPF Vector is not present anymore.

```
16 2017/10/17 12:03:16.346 UTC MINOR: DEBUG #2001 Base PIM[Instance 1 Base]
"PIM[Instance 1 Base]: Join/Prune
[000 00:27:35.920] PIM-TX ifId 2 ifName int-P-2-PE-1 0.0.0.0 -> 224.0.0.13 Length: 34
PIM Version: 2 Msg Type: Join/Prune Checksum: 0x563f
Upstream Nbr IP : 192.168.12.1 Resvd: 0x0, Num Groups 1, HoldTime 210
Group: 239.255.0.1/32 Num Joined Srcs: 1, Num Pruned Srcs: 0
Joined Srcs:
192.0.2.1/32 Flag S  <S,G>
"
```

As a result of this signaling, the default MDT is established between the two ASs. This can be checked with the **show router pim group** command.

The following PE-1 output shows the active multicast groups which are used as default MDT.

```
*A:PE-1# show router pim group
===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
239.255.0.1                 (S,G)               spt      system         2
   192.0.2.1
239.255.0.1                 (S,G)               spt      int-PE-1-P-2   1
   192.0.2.4
-------------------------------------------------------------------------------
Groups : 2
===============================================================================
*A:PE-1#
```

The following PE-4 output shows the active multicast groups which are used as default MDT:

```
*A:PE-4# show router pim group
===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
239.255.0.1                 (S,G)               spt      int-PE-4-P-3   1
   192.0.2.1
239.255.0.1                 (S,G)               spt      system         2
   192.0.2.4
-------------------------------------------------------------------------------
```

```
Groups : 2
===============================================================================
*A:PE-4#
```

The detailed information about the PIM-P group shows that the default MDT is used to deliver traffic. Key parameters such as the incoming/outgoing interfaces and non-zero traffic counters allow this conclusion to be made.

PE-4 has the incoming interface “int-PE-4-P-3”, and outgoing interface “system”, as follows:

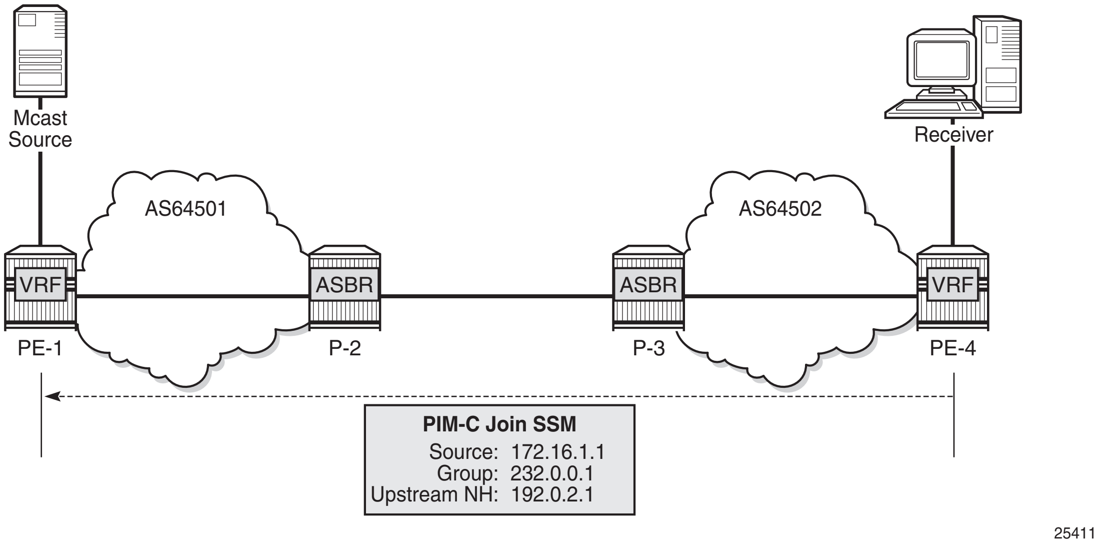
```
*A:PE-4# show router pim group detail
===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 239.255.0.1
Source Address     : 192.0.2.1
RP Address         : 0
Advt Router        : 192.0.2.3

Upstream RPFV Nbr  : 192.168.34.1
RPFV Type          : Mvpn 1:1          RPFV Proxy         : 192.0.2.3

Flags              : spt               Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 192.168.34.1
MRIB Src Flags     : remote
Keepalive Timer Exp: 0d 00:03:06
Up Time            : 0d 00:01:50       Resolved By        : rtable-u

Up JP State        : Joined            Up JP Expiry       : 0d 00:00:09
Up JP Rpt          : Not Joined StarG  Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 192.168.34.1
Incoming Intf      : int-PE-4-P-3
Outgoing Intf List : system

Curr Fwding Rate   : 0.0 kbps
Forwarded Packets  : 4                 Discarded Packets  : 0
Forwarded Octets   : 312               RPF Mismatches     : 0
Spt threshold      : 0 kbps            ECMP opt threshold : 7
Admin bandwidth    : 1 kbps

--- snipped ---

-------------------------------------------------------------------------------
Groups : 2
===============================================================================
*A:PE-4#
```

PE-1 has incoming the interface “system”, and outgoing interfaces “system, int-PE-1-P-2”, as follows:

```
*A:PE-1# show router pim group detail
===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 239.255.0.1
Source Address     : 192.0.2.1
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              : spt                Type              : (S,G)
Mode               : sparse
MRIB Next Hop      :
MRIB Src Flags     : self
Keepalive Timer Exp: 0d 00:03:14
Up Time            : 0d 00:01:50        Resolved By       : rtable-m

Up JP State        : Joined             Up JP Expiry      : 0d 00:00:10
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       :
Incoming Intf      : system
Outgoing Intf List : system, int-PE-1-P-2

Curr Fwding Rate   : 0.0 kbps
Forwarded Packets  : 8                  Discarded Packets  : 0
Forwarded Octets   : 624                RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps

--- snipped ---

-------------------------------------------------------------------------------
Groups : 2
===============================================================================
*A:PE-1#
```

Customer PIM Signaling

Figure 197 **PIM-C Signaling**



The PIM-C Join is sent to the sender PE using the existing tunnel infrastructure.

On PE-4, the **debug router 1 pim packet jp** output shows the PIM join/prune message which is sent to PE-1 using PMSI interface “1-mt-239.255.0.1” inside VPRN 1. All of this information and more can be found in the output of the **debug** command.

```
29 2017/10/17 12:15:26.477 UTC MINOR: DEBUG #2001 vprn1 PIM[Instance 2 vprn1]
"PIM[Instance 2 vprn1]: Join/Prune
[000 00:39:26.480] PIM-TX ifId 16385 ifName 1-mt-239.255.0.1 0.0.0.0 -> 224.0.0.13
 Length: 34
PIM Version: 2 Msg Type: Join/Prune Checksum: 0x7dd6
Upstream Nbr IP : 192.0.2.1 Resvd: 0x0, Num Groups 1, HoldTime 210
Group: 232.0.0.1/32 Num Joined Srcs: 1, Num Pruned Srcs: 0
Joined Srcs:
172.16.1.1/32 Flag S  <S,G>
"
```

The detailed information about the PIM-C group for a particular VPRN shows that the default MDT is used to deliver traffic. For this purpose, the **show router 1 pim group detail** command is used. Key parameters such as the correct multicast group, correct incoming/outgoing interfaces and non-zero flow rate allow this conclusion to be made.

PE-1 has the incoming interface “int-PE-1-S-1”, and outgoing interface “1-mt-239.255.0.1”. If the threshold hasn’t been reached to set up a selective provider tunnel, only one outgoing interface is listed.In order to generate this output, the data threshold for the selective provider tunnel was temporarily raised to 100000 kbps in VPRN 1.

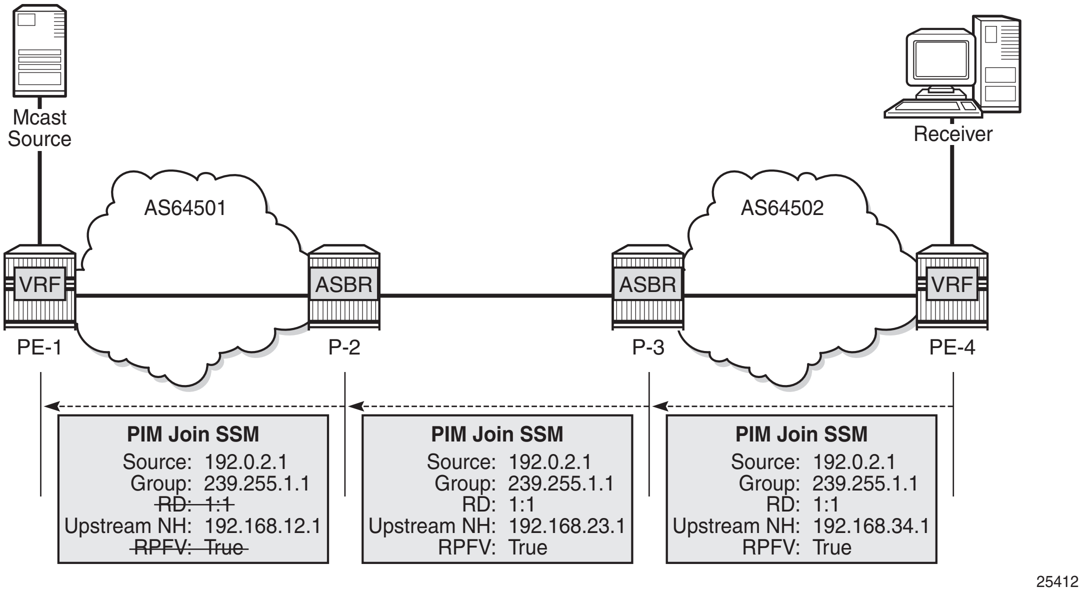
```
*A:PE-1# show router 1 pim group detail
===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 232.0.0.1
Source Address     : 172.16.1.1
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 172.16.1.1
MRIB Src Flags     : direct
Keepalive Timer    : Not Running
Up Time            : 0d 00:01:04        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:00
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 172.16.1.1
Incoming Intf      : int-PE-1-S-1
Outgoing Intf List : 1-mt-239.255.0.1

Curr Fwding Rate   : 509.3 kbps
Forwarded Packets  : 2684               Discarded Packets  : 0
Forwarded Octets   : 4020632            RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-1#
```

PE-4 has the incoming interface “1-mt-239.255.0.1”, and outgoing interface “int-PE-4-H-4” to the receiving host. As long as there is no S-PMSI, the following output can be seen.

```
*A:PE-4# show router 1 pim group detail
===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 232.0.0.1
Source Address     : 172.16.1.1
RP Address         : 0
Advt Router        : 192.0.2.3
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 192.0.2.1
MRIB Src Flags     : remote
Keepalive Timer    : Not Running
Up Time            : 0d 00:01:10        Resolved By         : rtable-u

Up JP State        : Joined             Up JP Expiry        : 0d 00:00:49
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override  : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 192.0.2.1
Incoming Intf      : 1-mt-239.255.0.1
Outgoing Intf List : int-PE-4-H-4

Curr Fwding Rate   : 509.3 kbps
Forwarded Packets  : 3006               Discarded Packets   : 0
Forwarded Octets   : 4502988            RPF Mismatches      : 0
Spt threshold      : 0 kbps             ECMP opt threshold  : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-4#
```

When Multicast Stream Threshold is Reached

Figure 198 **PIM-P Signaling Steps for Data MDT**



On PE-4, the **debug router pim packet jp** output shows the PIM join/prune message which is sent to P-3. This message contains the original source of the multicast traffic (PE-1: 192.0.2.1) and the RPF Vector (P-3: 192.0.2.3).

A new multicast group (239.255.1.1) is signaled for purposes of establishing the data MDT.

```
19 2017/10/17 12:19:19.174 UTC MINOR: DEBUG #2001 Base PIM[Instance 1 Base]
"PIM[Instance 1 Base]: Join/Prune
[000 00:43:19.180] PIM-TX ifId 2 ifName int-PE-4-P-3 0.0.0.0 -> 224.0.0.13 Length: 48
PIM Version: 2 Msg Type: Join/Prune Checksum: 0x3d2c
Upstream Nbr IP : 192.168.34.1 Resvd: 0x0, Num Groups 1, HoldTime 210
Group: 239.255.1.1/32 Num Joined Srcs: 1, Num Pruned Srcs: 0
Joined Srcs:
192.0.2.1/32 Flag S  <S,G> JA={rpfvMvpn 192.0.2.3 1:1}
"
```

On P-3, the **debug router pim packet jp** output shows the PIM join/prune message which is propagated to P-2. The source of multicast traffic is untouched while the RPF Vector is modified for Inter-AS propagation.

```
29 2017/10/17 12:19:19.030 UTC MINOR: DEBUG #2001 Base PIM[Instance 1 Base]
"PIM[Instance 1 Base]: Join/Prune
[000 00:43:30.130] PIM-TX ifId 2 ifName int-P-3-P-2 0.0.0.0 -> 224.0.0.13 Length: 48
PIM Version: 2 Msg Type: Join/Prune Checksum: 0x3286
Upstream Nbr IP : 192.168.23.1 Resvd: 0x0, Num Groups 1, HoldTime 210
Group: 239.255.1.1/32 Num Joined Srcs: 1, Num Pruned Srcs: 0
Joined Srcs:
192.0.2.1/32 Flag S  <S,G> JA={rpfvMvpn 192.168.23.1 1:1}
"
```

On P-2, the **debug router pim packet jp** output shows the PIM join/prune message which is propagated to PE-1. The source of multicast traffic is untouched while the RPF Vector is not present anymore.

```
29 2017/10/17 12:19:17.592 UTC MINOR: DEBUG #2001 Base PIM[Instance 1 Base]
"PIM[Instance 1 Base]: Join/Prune
[000 00:43:37.170] PIM-TX ifId 2 ifName int-P-2-PE-1 0.0.0.0 -> 224.0.0.13 Length: 34
PIM Version: 2 Msg Type: Join/Prune Checksum: 0x563f
Upstream Nbr IP : 192.168.12.1 Resvd: 0x0, Num Groups 1, HoldTime 210
Group: 239.255.1.1/32 Num Joined Srcs: 1, Num Pruned Srcs: 0
Joined Srcs:
192.0.2.1/32 Flag S  <S,G>
"
```

As a result of this signaling, the Data MDT is established between the two ASs. This can be checked with **show router pim group** command.

The PE-1 output shows an additional multicast group (239.255.1.1), which was created in the global routing table (GRT).

```
*A:PE-1# show router pim group
===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
```

```
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
239.255.0.1                 (S,G)               spt      system         2
   192.0.2.1
239.255.0.1                 (S,G)               spt      int-PE-1-P-2   1
   192.0.2.4
239.255.1.1                 (S,G)                        system         1
   192.0.2.1
-------------------------------------------------------------------------------
Groups : 3
===============================================================================
*A:PE-1#
```

The PE-4 output shows an additional multicast group (239.255.1.1), which was created in the GRT.

```
A:PE-4# show router pim group
===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
239.255.0.1                 (S,G)               spt      int-PE-4-P-3   1
   192.0.2.1
239.255.0.1                 (S,G)               spt      system         2
   192.0.2.4
239.255.1.1                 (S,G)                        int-PE-4-P-3   1
   192.0.2.1
-------------------------------------------------------------------------------
Groups : 3
===============================================================================
A:PE-4#
```

The detailed information about the PIM group in a VPRN shows that the data MDT is used to receive traffic instead of the default MDT.

The PE-4 output for multicast groups in a VPRN 1 has slightly changed: a new line “Incoming SPMSI Intf” was added. This indicates that the S-PMSI instance and dedicated Data MDT are used for this particular multicast group. The non-zero rate for the multicast flow is also an indication that multicast traffic is forwarded.

```
*A:PE-4# show router 1 pim group detail
===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 232.0.0.1
Source Address     : 172.16.1.1
RP Address         : 0
Advt Router        : 192.0.2.3
Flags              :                   Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 192.0.2.1
MRIB Src Flags     : remote
Keepalive Timer    : Not Running
Up Time            : 0d 00:02:49       Resolved By        : rtable-u

Up JP State        : Joined            Up JP Expiry       : 0d 00:00:10
Up JP Rpt          : Not Joined StarG  Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 192.0.2.1
Incoming Intf      : 1-mt-239.255.0.1
Incoming SPMSI Intf: 1-mt-239.255.0.1*
Outgoing Intf List : int-PE-4-H-4

Curr Fwding Rate   : 509.3 kbps
Forwarded Packets  : 7210              Discarded Packets  : 0
Forwarded Octets   : 10800580          RPF Mismatches     : 0
Spt threshold      : 0 kbps            ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-4#
```

The **show router 1 pim s-pmsi detail** command can also be used to verify existence of the S-PMSI instance for the VPRN 1. The output includes the multicast group inside the VPRN, the multicast source IP, the multicast group which is used for S-PMSI tunneling and the current forwarding rate.

```
*A:PE-4# show router 1 pim s-pmsi detail
===============================================================================
PIM Selective provider tunnels
===============================================================================
Md Source Address  : 192.0.2.1         Md Group Address   : 239.255.1.1
Number of VPN SGs  : 1                 Uptime             : 0d 00:03:15
MT IfIndex         : 24576             Egress Fwding Rate : 509.3 kbps

VPN Group Address  : 232.0.0.1
VPN Source Address : 172.16.1.1
State              : RX Joined
Expiry Timer       : 0d 00:01:57
===============================================================================
PIM Selective provider tunnels Interfaces : 1
===============================================================================
*A:PE-4#
```

Conclusion

Inter-AS MVPN offers flexibility for the operators who can use it to provide additional value added services to their customers. Before implementing this feature in the network the following are required:

- The RPF vector must be enabled on every router for inter-AS MVPN.
- Can be used only with Rosen MVPN with PIM SSM and MDT SAFI.

Spoke Termination for IPv6-6VPE

This chapter provides information about spoke termination for IPv6-6VPE.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SR-OS 9.0. The CLI in the current edition corresponds to 14.0.R5.

Overview

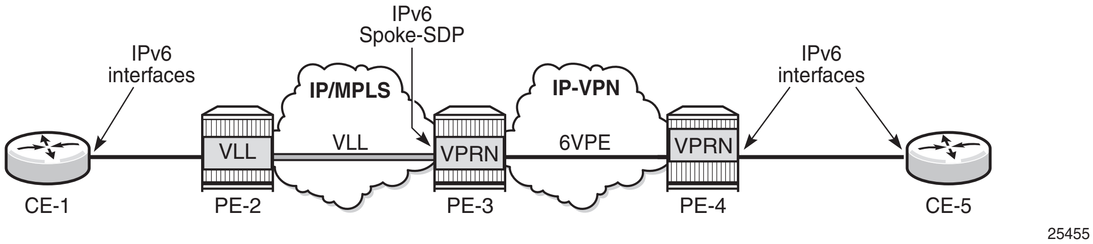
RFC 4659, *BGP-MPLS IP Virtual Private Network (VPN) Extension for IPv6 VPN*, standardized the use of an IPv6 over IPv4 tunneling scheme. SR OS supports the standardized IPv6 over IPv4 tunneling scheme for Virtual Private Routed Network (VPRN) services using Multi-Protocol Border Gateway Protocol (MP-BGP), also known as 6VPE. The 7750 SR also supports pseudowire termination by a VPRN from an Epipe Virtual Leased Line (VLL) or VPLS spoke SDP where the pseudowire can be given IPv6 addresses and run IPv6 protocols. In the example used in this chapter, any advertisements across the Multi-Protocol Labeled Switching (MPLS) network between Virtual Private Routed Network (VPRN) Provider Edge (PE) devices will use 6VPE. The goal of this section is to list configuration guidelines for IPv6 spoke termination to a VPRN over an Epipe VLL and transporting IPv6 packets over 6VPE tunnels between PE devices.

This solution is to be used where a service provider is providing VPRN services built on a transport network whose Interior Gateway Protocol (IGP) is using IPv4 addressing on the network interfaces. The customer's CE and the service provider's PE must support IPv6 pseudowires, IPv6 interfaces and in addition, the service provider also must be able to support the advertisements of IPv6 prefixes between CE-PE peerings and between the transport PE routers using MP-BGP. The advertisement of IPv6 prefixes across the MPLS network and the transport of IPv6 traffic is tunneled using 6VPE.

The VPRN PE has the ability to support spoke termination of Epipe VLL services on access with IPv6 addressing between the CE and VPRN PE. The IPv6 spoke termination on VPRN services has the same functionality as VPRN IPv4 spoke termination.

The example in Figure 199 illustrates a CE device that connects to a VPRN PE on an IPv6 interface addressing using spoke termination.

Figure 199 **Spoke Termination for IPv6**



CE-1 is connected to the VPRN service on PE-3, using IPv6 interfaces. CE-1 reaches PE-3 by connecting to PE-2. PE-2 uses an Epipe VLL for transport to the VPRN on the PE-3. The connectivity between the VLL service on the VPRN service on PE-3 is using spoke termination with IPv6 addressing on the spoke-service distribution point (spoke SDP) interface on PE-3.

Figure 200 **IPv6 Addressing and IPv6 Prefixes**

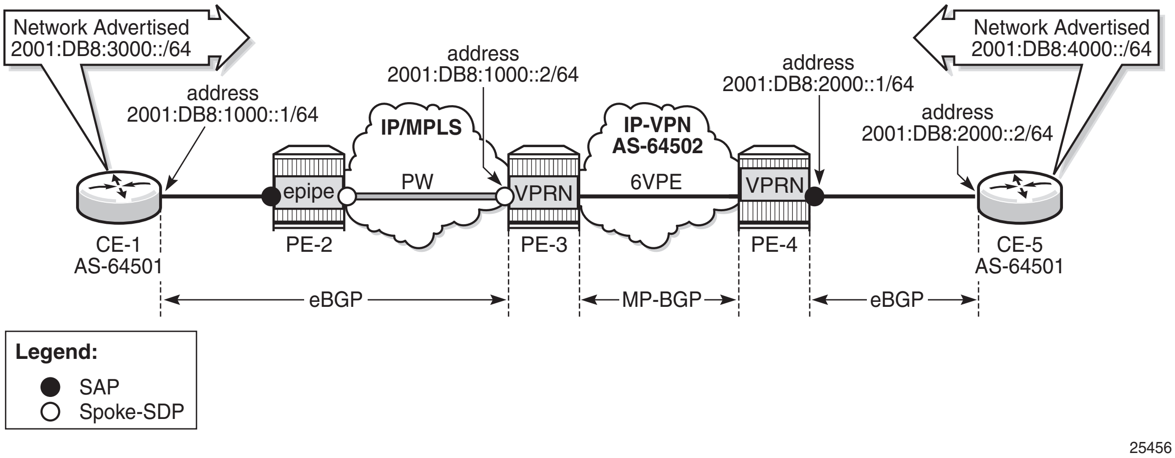


Figure 200 shows the overall IPv6 addressing from interfaces to prefixes advertised from CE-1 and CE-5 across the VPRN network.

- Link between CE-1 and PE-3: 2001:DB8:1000::/64
- Link between CE-5 and PE-4: 2001:DB8:2000::/64
- Advertised Prefix from CE-1: 2001:DB8:3000::/64
- Advertised Prefix from CE-5: 2001:DB8:4000::/64

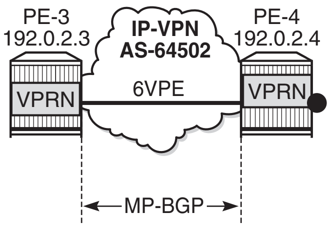
PE-3 has an MP-eBGP session with CE-1 to receive and advertise IPv6 routes. PE-3 also has an MP-iBGP peering session with PE-4 to use 6VPE to tunnel IPv6 routes and traffic to and from PE-4. PE-4 has an IPv6 SAP interface to CE-5 and uses MP-eBGP to advertise and receive routes to/from CE-5 (no spoke-termination). The configuration of PE-3 will be included to provide examples of the end-to-end VPRN service using a 6VPE model.

This network topology will illustrate the use of spoke termination using IPv6 interfaces and the tunneling of IPv6 traffic over a 6VPE MPLS network.

Configuration

First an MPLS network is established where the VPRN service can use 6VPE to tunnel traffic across the IPv4 IGP.

Figure 201 **MP-BGP VPN IPv6**



25457

PE-3 and PE-4 in Figure 201 are edge routers running VPRN Services on access with IPv6 Interfaces. The MPLS network is configured using IPv4 link addressing. Interior Border Gateway Protocol (iBGP) peerings need to be established with MP-BGP for VPN-IPv6 address families between PE-3 and PE-4.

```
*A:PE-3# configure
    router
        autonomous-system 64502
        bgp
            group "iBGP"
                description "iBGP peering in AS 64502"
                family vpn-ipv6
                type internal
                neighbor 192.0.2.4
                    description "PE-4"
                exit
            exit
        exit
    exit
exit
```

```
*A:PE-4# configure
    router
        autonomous-system 64502
        bgp
            group "iBGP"
                description "iBGP peering in AS 64502"
                family vpn-ipv6
                type internal
                neighbor 192.0.2.3
                    description "PE-3"
                exit
            exit
        exit
    exit
exit
```

Configuring address family vpn-ipv6 between VPRN PE edge routers in BGP enables the functionality of MP-BGP for the Layer 3 VPNs supporting the customer’s IPv6 Addressing (6VPE).

Verify BGP sessions for VPN-IPv6 address families between PE-3 and PE-4.

```
*A:PE-3# show router bgp neighbor 192.0.2.4
===============================================================================
BGP Neighbor
===============================================================================
-------------------------------------------------------------------------------
Peer                 : 192.0.2.4
Description          : PE-4
Group                : iBGP
-------------------------------------------------------------------------------
Peer AS              : 64502            Peer Port            : 50222
Peer Address         : 192.0.2.4
Local AS             : 64502            Local Port           : 179
Local Address        : 192.0.2.3
Peer Type            : Internal         Dynamic Peer         : No
State                : Established      Last State           : Established
Last Event           : recvKeepAlive
Last Error           : Cease (Connection Collision Resolution)
Local Family         : VPN-IPv6
Remote Family        : VPN-IPv6
---snip---
```

```
*A:PE-4# show router bgp neighbor 192.0.2.3
===============================================================================
BGP Neighbor
===============================================================================
-------------------------------------------------------------------------------
Peer                 : 192.0.2.3
Description          : PE-3
Group                : iBGP
-------------------------------------------------------------------------------
Peer AS              : 64502            Peer Port            : 179
Peer Address         : 192.0.2.3
Local AS             : 64502            Local Port           : 50222
Local Address        : 192.0.2.4
Peer Type            : Internal         Dynamic Peer         : No
State                : Established      Last State           : Active
Last Event           : recvKeepAlive
Last Error           : Unrecognized Error
Local Family         : VPN-IPv6
Remote Family        : VPN-IPv6
---snip---
```

After the MP-BGP sessions are established for VPN-IPv6 address-family, 6VPE tunnel support is provided between PE-3 and PE-4.

Figure 202 **Spoke Termination for IPv6 Addressing**

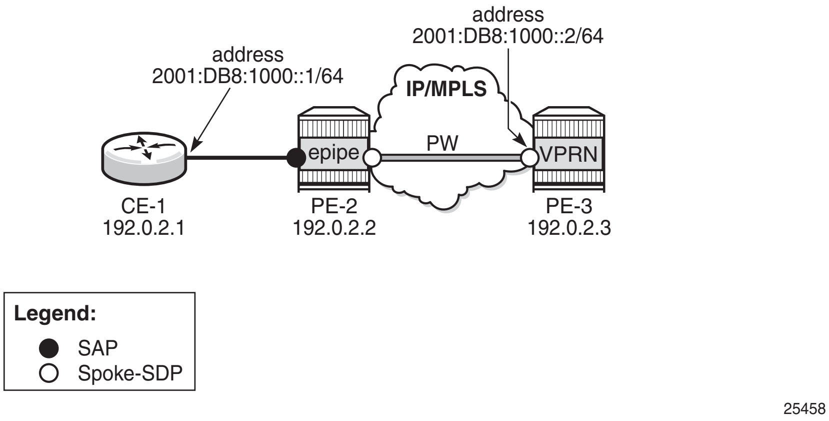


Figure 202 illustrates the model for spoke termination for IPv6 using VPRN services. CE-1 is configured with IPv6 addressing on the access interface facing the VPRN service. CE-1’s access is backhauled to the VPRN service on PE-3 using Epipe VLL with spoke termination. Only Epipe VLL is supported for IPv6 spoke termination within the VPRN in R8.0. The configuration of the Epipe VLL on PE-2 is as follows:

```
*A:PE-2# configure
    service
        sdp 231 mpls create
            far-end 192.0.2.3
            ldp
            no shutdown
        exit
        epipe 1 customer 1 create
            sap 1/1/2 create
            exit
            spoke-sdp 231:1 create
            exit
            no shutdown
        exit
```

The example is taken from PE-2 which has been configured using the Epipe VLL service with a SAP interface facing the customer and a spoke SDP facing PE-3. The spoke SDP is terminated into the customer’s VPRN service on PE-3.

PE-3 is now ready to be configured for the IPv6 spoke SDP. Review the possible IPv6 options for spoke SDP interfaces on the CLI for VPRN Services (compliant with RFC 4213, *Basic Transition Mechanisms for IPv6 Hosts and Routers <draft-ietf-v6ops-mech-v2-07.txt>)*:

- Interface spoke SDP (IPv6 options only)

```
*A:PE-3# configure service vprn 1 interface "int-PE-3-PE-2"
---snip---
```

```
[no] ipv6            + Enables/Configures IPv6 for a VPRN interface
---snip---

*A:PE-3# configure service vprn 1 interface "int-PE-3-PE-2" ipv6
  - ipv6
  - no ipv6

 [no] address         - Assigns an IPv6 address to the VPRN interface.
 [no] bfd             - Configure BFD parameters
 [no] dad-disable     - Disable Duplicate Address Detection
 [no] dhcp6-relay     + Configure DHCPv6 relay parameters for the VPRN interface
 [no] dhcp6-server    + Configure DHCPv6 server parameters for the VPRN interface
      icmp6           + Configure ICMPv6 parameters for the VPRN interface
 [no] link-local-add* - Configure link-local address
 [no] local-dhcp-ser* - Assign a DHCP server to the interface
 [no] local-proxy-nd  - Enable/disable local proxy Neighbor Discovery on the VPRN
                        interface
 [no] neighbor        - Configure IPv6-to-MAC address mapping on the VPRN interface
 [no] neighbor-limit  - Configures the maximum amount of IPv6 neighbor entries
 [no] proxy-nd-policy - Configure a proxy Neighbor Discovery policy for the VPRN
                        interface
 [no] qos-route-look* - Enable/Disable Qos route lookup for the interface
 [no] reachable-time  - Configure neighbor reachability detection timer
 [no] secure-nd       + Configure Secure Neighbor Discovery (SEND) parameters for
                        the interface
 [no] stale-time      - Configure the time a neighbor discovery cache entry can
                        remain stale before being removed
 [no] tcp-mss         - Configure TCP maximum segment size for the interface
 [no] urpf-check      + Enables/Configures unicast RPF check for an interface
 [no] vrrp            + Context to create and configure VRRP virtual router instance
                        on the interface
```

- IPv6 address

```
*A:PE-3# configure service vprn 1 interface "int-PE-3-PE-2" ipv6 address
  - address <ipv6-address/prefix-length> [eui-64] [preferred]
            [track-srrp <srrp-instance>] [modifier <cga-modifier>]
  - no address <ipv6-address/prefix-length>

 <ipv6-address/pref*> : ipv6-address  x:x:x:x:x:x:x:x   (eight 16-bit pieces)
                                      x:x:x:x:x:x:d.d.d.d
                                      x [0..FFFF]H
                                      d [0..255]D
                                      (no multicast address)
                        prefix-length [1..128]
 <eui-64>             : keyword
 <preferred>          : keyword
 <srrp-instance>      : [1..4294967295]
 <cga-modifier>       : [0x0..0xFFFFFFFF...(32 hex nibbles)]
```

- DHCPv6 relay parameters for the VPRN service (default settings)

```
*A:PE-3>config>service>vprn>if>ipv6>dhcp6-relay# info detail
----------------------------------------------
                        shutdown
                        no description
                        no lease-populate
                        no neighbor-resolution
                        option
                            no interface-id
                            no remote-id
                        exit
                        no source-address
                        no link-address
                        no user-db
                        no python-policy
                        no server
```

- DHCPv6 server parameters for the VPRN service (default)

```
*A:PE-3>config>service>vprn>if>ipv6>dhcp6-server# info detail
----------------------------------------------
                        prefix-delegation
                            shutdown
                        exit
                        max-nbr-of-leases 8000
```

- ICMPv6 (default)

```
*A:PE-3>config>service>vprn>if>ipv6>icmp6# info detail
----------------------------------------------
                        packet-too-big 100 10
                        param-problem 100 10
                        redirects 100 10
                        time-exceeded 100 10
                        unreachables 100 10
```

- Link-local-addressing, for the VPRN interface. By default, link-local addressing is assigned dynamically. Use this command if you would like to add a static link-local-address.

```
*A:PE-3# configure service vprn 1 interface "int-PE-3-PE-2" ipv6 link-local-address
  - link-local-address <ipv6-address> [preferred]
  - no link-local-address

 <ipv6-address>       : ipv6-address   - x:x:x:x:x:x:x:x
                                         x:x:x:x:x:x:d.d.d.d
                                         x [0..FFFF]H
                                         d [0..255]D
 <preferred>          : keyword
```

- Neighbor

  * IPv6 to MAC address mapping on the VRPN Interface

```
*A:PE-3# configure service vprn 1 interface "int-PE-3-PE-2" ipv6 neighbor
  - neighbor <ipv6-address> <mac-address>
  - no neighbor <ipv6-address>

 <ipv6-address>       : x:x:x:x:x:x:x:x   (eight 16-bit pieces)
                                      x:x:x:x:x:x:d.d.d.d
                                      x [0..FFFF]H
                                      d [0..255]D
                        prefix-length [1..128]
 <mac-address>        : xx:xx:xx:xx:xx:xx or xx-xx-xx-xx-xx-xx
```

- Enabling Local Proxy Neighbor Discovery

```
*A:PE-3# configure service vprn 1 interface "int-PE-3-PE-2" ipv6 local-proxy-nd
  - local-proxy-nd
  - no local-proxy-nd
```

- VRRP

```
*A:PE-3# configure service vprn 1 interface "int-PE-3-PE-2" ipv6 vrrp
  - no vrrp <virtual-router-id>
  - vrrp <virtual-router-id> [owner]

 <virtual-router-id>  : [1..255]
 <owner>              : keyword

 [no] backup          - Configure virtual router IP addresses for the interface
 [no] bfd-enable      - Configure a BFD interface
 [no] init-delay      - Configure VRRP initialization delay timer
 [no] mac             - Configure a Virtual MAC address to use in Neighbor Discovery
 [no] master-int-inh* - Allow/disallow the master instance to dictate the master
                        down timer (non-owner context only)
 [no] message-interv* - Configure the interval for sending VRRP Advertisement
                        messages
 [no] ping-reply      - Allow/disallow non-owner master to reply to ICMP Echo
                        requests (non-owner context only)
 [no] policy          - Associate a VRRP Priority Control Policy with the virtual
                        router instance (non-owner context only)
 [no] preempt         - Allow/disallow the virtual router instance to override an
                        existing non-owner master (non-owner context only)
 [no] priority        - Configure the base priority for the virtual router
                        instance (non-owner context only)
 [no] shutdown        - Administratively enable/disable the virtual router
                        instance (non-owner context only)
 [no] standby-forwar* - Allow/disallow the forwarding of packets by a standby router
 [no] telnet-reply    - Allow/disallow non-owner master to reply to Telnet requests
                        (non-owner context only)
 [no] traceroute-rep* - Allow/disallow non-owner master to reply to traceroute
                        requests (non-owner context only)
```

The VPRN on PE-3 will export IPv6 routes (IPv6 route on CE-5) to CE-1. A route policy needs to be configured.

```
*A:PE-3# configure
    router
        policy-options
            begin
            prefix-list "PE-3-CE-1"
                prefix 2001:db8:4000::/64 exact
            exit
            policy-statement "PE-3-BGP-CE-1"
                entry 10
                    from
                        prefix-list "PE-3-CE-1"
                    exit
                    action accept
                    exit
                exit
                default-action drop
                exit
            exit
            commit
        exit
    exit
exit
```

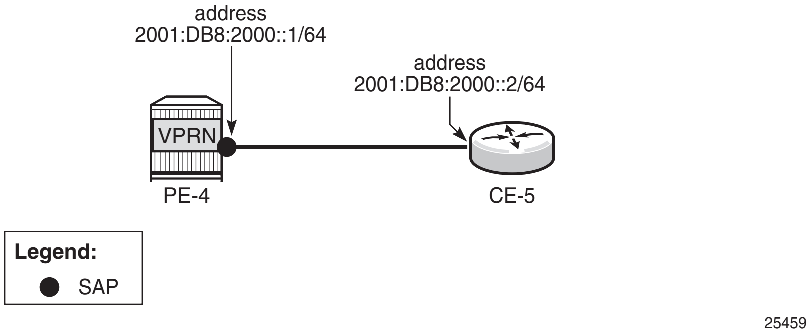
The configuration for the VPRN service on PE-3 with IPv6 interface (spoke SDP) in reference to Figure 202:

```
*A:PE-3# configure
    service
        sdp 321 mpls create
            far-end 192.0.2.2
            ldp
            no shutdown
        exit
        vprn 1 customer 1 create
            router-id 192.0.2.6
            autonomous-system 64502
            route-distinguisher 64502:1
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64502:1
            interface "loopback" create
                address 192.0.2.6/32
                loopback
            exit
            interface "int-PE-3-PE-2" create
                description "Spoke SDP"
                ipv6
                    address 2001:db8:1000::2/64
                exit
                spoke-sdp 321:1 create
                exit
            exit
            bgp
                router-id 192.0.2.6
                group "Spoke-CE-1-PE-3"
                    family ipv6
                    peer-as 65501
                    local-address 2001:db8:1000::2
                    neighbor 2001:db8:1000::1
                        as-override
                        type external
                        export "PE-3-BGP-CE-1"
                    exit
                exit
                no shutdown
            exit
            no shutdown
        exit
    exit
exit
```

In the preceding configuration example, PE-3 has been configured with an IPv6 spoke SDP (spoke termination) with spoke interface int-PE-3-PE-2. The VPRN configuration has also been set up for MP-eBGP peering to CE-1 through the IPv6 spoke interface. The MP-eBGP peering will be receiving and advertising IPv6 prefixes from/to CE-1. Route policy configuration has been included to show how IPv6 routes are advertised to CE-1 from PE-3 (policy-statement PE-3-BGP-CE-1).

The configuration on PE-4 is similar, but there is a SAP interface to CE-5 instead of a spoke-SDP.

Figure 203 **PE-4 VPRN with SAP to CE-5**



The IPv6 configuration options for the SAP interface (int-PE-4-CE-5) are similar to those in the preceding example for the spoke SDP on PE-3. The PE-4 BGP export policy (PE-4-BGP-CE-5) is also similar to the example for PE-3 in advertising the learned IPv6 route to CE-5.

```
*A:PE-4# configure
    router
        policy-options
            begin
            prefix-list "PE-4-CE-5"
                prefix 2001:db8:3000::/64 exact
            exit
            policy-statement "PE-4-BGP-CE-5"
                entry 10
                    from
                        prefix-list "PE-4-CE-5"
                    exit
                    action accept
                    exit
                exit
                default-action drop
                exit
            exit
            commit
        exit
    exit
exit
```

```
*A:PE-4# configure
    service
        vprn 1 customer 1 create
            router-id 192.0.2.7
            autonomous-system 64502
            route-distinguisher 64502:1
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64502:1
            interface "loopback" create
                address 192.0.2.7/32
                loopback
            exit
            interface "int-PE-4-CE-5" create
                ipv6
                    address 2001:db8:2000::1/64
                exit
                sap 1/1/1 create
                exit
            exit
            bgp
                router-id 192.0.2.7
                group "CE-5-PE-4"
                    family ipv6
                    peer-as 64501
                    local-address 2001:db8:2000::1
                    neighbor 2001:db8:2000::2
                        as-override
                        type external
                        export "PE-4-BGP-CE-5"
                    exit
                exit
                no shutdown
            exit
            no shutdown
        exit
    exit
exit
```

In this setup, the configuration on CE-1 is as follows.

```
*A:CE-1# configure
    router
        static-route-entry 2001:db8:3000::/64
            black-hole
                no shutdown
            exit
        exit
        policy-options
            begin
            prefix-list "CE-1-192.0.2.1"
                prefix 2001:db8:3000::/64 exact
            exit
            policy-statement "CE-1-sys-to-eBGP"
```

```
                entry 10
                    from
                        prefix-list "CE-1-192.0.2.1"
                    exit
                    action accept
                    exit
                exit
                default-action drop
                exit
            exit
            commit
        exit
        bgp
            router-id 192.0.2.1
            group "eBGP_to_64502"
                description "eBGP_to_PE-3_AS64502"
                family ipv6
                type external
                peer-as 64502
                local-address 2001:db8:1000::1
                neighbor 2001:db8:1000::2
                    export "CE-1-sys-to-eBGP"
                exit
            exit
            no shutdown
        exit
    exit
exit
```

```
*A:CE-1# configure
    service
        ies 1 customer 1 create
            interface "int-CE-1-PE-2" create
                description "SAP_toward_VPRN_Service"
                ipv6
                    address 2001:db8:1000::1/64
                exit
                sap 1/1/1 create
                exit
            exit
            no shutdown
        exit
    exit
exit
```

The configuration on CE-5 is similar.

The following command on PE-2 shows that the Epipe VLL is established with the SAP facing CE-1 and spoke SDP facing VPRN on PE-3.

```
*A:PE-2# show service id 1 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 1                   Vpn Id            : 0
Service Type      : Epipe
```

```
Name              : (Not Specified)
Description       : (Not Specified)
Customer Id       : 1                   Creation Origin   : manual
Last Status Change: 10/20/2016 12:54:38
Last Mgmt Change  : 10/20/2016 12:54:11
Test Service      : No
Admin State       : Up                  Oper State        : Up
MTU               : 9190
Vc Switching      : False
SAP Count         : 1                   SDP Bind Count    : 1
Per Svc Hashing   : Disabled
Force QTag Fwd    : Disabled

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/2                                null         9212    9212    Up   Up
sdp:231:1 S(192.0.2.3)                   Spok         0       9190    Up   Up
===============================================================================
*A:PE-2#
```

The same command can be launched on PE-3 to verify that the VPRN service is up and that the spoke SDP is up (admin state up/oper state up).

```
*A:PE-3# show service id 1 base
===============================================================================
Service Basic Information
===============================================================================
Service Id        : 1                   Vpn Id            : 0
Service Type      : VPRN
Name              : (Not Specified)
Description       : (Not Specified)
Customer Id       : 1                   Creation Origin   : manual
Last Status Change: 10/20/2016 12:56:13
Last Mgmt Change  : 10/20/2016 12:56:13
Admin State       : Up                  Oper State        : Up

Route Dist.       : 64502:1             VPRN Type         : regular
Oper Route Dist   : 64502:1
Oper RD Type      : configured
AS Number         : 64502               Router Id         : 192.0.2.6
ECMP              : Enabled             ECMP Max Routes   : 1
Max IPv4 Routes   : No Limit

Auto Bind Tunnel
Resolution        : filter
Filter Protocol   : ldp

Max IPv6 Routes   : No Limit
Ignore NH Metric  : Disabled
Hash Label        : Disabled
Entropy Label     : Disabled
Vrf Target        : target:64502:1
Vrf Import        : None
Vrf Export        : None
MVPN Vrf Target   : None
MVPN Vrf Import   : None
MVPN Vrf Export   : None
Car. Sup C-VPN    : Disabled
Label mode        : vrf
BGP VPN Backup    : Disabled
BGP Export Inacti*: Disabled

SAP Count         : 0                   SDP Bind Count    : 1
VSD Domain        : <none>

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                              Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sdp:321:1 S(192.0.2.2)                  TLDP         0       9190    Up   Up
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-3#
```

The following command shows that the IPv6 interface is established and its IPv6 address is preferred (2001:DB8:1000::2/64). The IPv6 link local address has been dynamically assigned and preferred (FE80::48C6:FFFF:FE00:0/64).

```
*A:PE-3# show service id 1 interface
===============================================================================
Interface Table
===============================================================================
Interface-Name                  Adm         Opr(v4/v6)  Type    Port/SapId
   IP-Address                                                   PfxState
-------------------------------------------------------------------------------
loopback                        Up          Up/Down     VPRN    loopback
   192.0.2.6/32                                                 n/a
int-PE-3-PE-2                   Up          Down/Up     VPRN    spoke-321:1
   2001:db8:1000::2/64                                          PREFERRED
   fe80::48c6:ffff:fe00:0/64                                    PREFERRED
-------------------------------------------------------------------------------
Interfaces : 2
===============================================================================
*A:PE-3#
```

With the following command an extensive list of parameters is displayed, including IPv6-related fields that can be checked if configured: DHCP6-relay, DHCP6-server, and so on. It is possible to use filters to reduce the output.

```
*A:PE-3# show service id 1 all
```

After verification of the services (Epipe, VPRN), the MP-eBGP peering connectivity (through IPv6 interfaces) on the VPRN between PE-3 and CE-1 can be verified as follows:

```
*A:PE-3# show router 1 bgp neighbor
===============================================================================
BGP Neighbor
===============================================================================
-------------------------------------------------------------------------------
Peer                 : 2001:db8:1000::1
Description          : (Not Specified)
Group                : Spoke-CE-1-PE-3
-------------------------------------------------------------------------------
Peer AS              : 64501            Peer Port            : 179
Peer Address         : 2001:db8:1000::1
Local AS             : 64502            Local Port           : 49218
Local Address        : 2001:db8:1000::2
Peer Type            : External         Dynamic Peer         : No
State                : Established      Last State           : Active
Last Event           : recvKeepAlive
Last Error           : Unrecognized Error
Local Family         : IPv6
Remote Family        : IPv6
---snip---
IPv6 Recd. Prefixes  : 1                IPv6 Active Prefixes : 1
---snip---
Local Capability     : RtRefresh MPBGP 4byte ASN
Remote Capability    : RtRefresh MPBGP 4byte ASN
---snip---
-------------------------------------------------------------------------------
Neighbors : 1
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-3#
```

Not only is the MP-eBGP session on the VPRN established, but the MP-BGP capabilities are also supported (locally and remotely). PE-3 and its BGP peer CE-1 have advertised and received an IPv6 prefix.

The same commands can be launched on PE-4. The status of the VPRN service on PE-4 and its interface to CE-5 can be verified as follows:

```
*A:PE-4# show service id 1 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 1                    Vpn Id            : 0
Service Type      : VPRN
Name              : (Not Specified)
Description       : (Not Specified)
Customer Id       : 1                    Creation Origin   : manual
Last Status Change: 10/20/2016 12:55:47
Last Mgmt Change  : 10/20/2016 12:55:47
Admin State       : Up                   Oper State        : Up

Route Dist.       : 64502:1              VPRN Type         : regular
Oper Route Dist   : 64502:1
Oper RD Type      : configured
AS Number         : 64502                Router Id         : 192.0.2.7
ECMP              : Enabled              ECMP Max Routes   : 1
Max IPv4 Routes   : No Limit

Auto Bind Tunnel
Resolution        : filter
```

```
Filter Protocol   : ldp

Max IPv6 Routes   : No Limit
Ignore NH Metric  : Disabled
Hash Label        : Disabled
Entropy Label     : Disabled
Vrf Target        : target:64502:1
Vrf Import        : None
Vrf Export        : None
MVPN Vrf Target   : None
MVPN Vrf Import   : None
MVPN Vrf Export   : None
Car. Sup C-VPN    : Disabled
Label mode        : vrf
BGP VPN Backup    : Disabled
BGP Export Inacti*: Disabled

SAP Count         : 1                   SDP Bind Count    : 0
VSD Domain        : <none>

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/1                                null         9212    9212    Up   Up
===============================================================================
*A:PE-4#
```

The VPRN service is up and the SAP is up.

The following output shows that the IPv6 interface is established on PE-4 and its IPv6 address is preferred.

```
*A:PE-4# show service id 1 interface
===============================================================================
Interface Table
===============================================================================
Interface-Name                   Adm         Opr(v4/v6)  Type    Port/SapId
   IP-Address                                                    PfxState
-------------------------------------------------------------------------------
loopback                         Up          Up/Down     VPRN    loopback
   192.0.2.7/32                                                  n/a
int-PE-4-CE-5                    Up          Down/Up     VPRN    1/1/1
   2001:db8:2000::1/64                                           PREFERRED
   fe80::48c7:ffff:fe00:0/64                                     PREFERRED
-------------------------------------------------------------------------------
Interfaces : 2
===============================================================================
*A:PE-4#
```

MP-iBGP, providing 6VPE, has been configured and built between PE-3 and PE-4 across the MPLS network. Now, propagate the advertisements of the IPv6 prefixes learned on PE-3 from CE-1 (2001:DB8:3000::/64) and on PE-4 from CE-5 (2001:DB8:4000::/64) across the MPLS network using MP-iBGP (6VPE).

CE-1 advertises IPv6 prefix 2001:DB8:3000::/64 and CE-5 advertises IPv6 prefix2001:DB8:4000::/64.

Verify on PE-3 whether VPN-IPv6 routes were received from and advertised to its iBGP peer PE-4, as follows:

```
*A:PE-3# show router bgp summary
===============================================================================
 BGP Router ID:192.0.2.3        AS:64502       Local AS:64502
===============================================================================
BGP Admin State         : Up          BGP Oper State              : Up
Total Peer Groups       : 1           Total Peers                 : 1
Total BGP Paths         : 13          Total Path Memory           : 2456
---snip---
Total VPN Peer Groups   : 1           Total VPN Peers             : 1
Total VPN Local Rts     : 4
Total VPN-IPv4 Rem. Rts : 0           Total VPN-IPv4 Rem. Act. Rts: 0
Total VPN-IPv6 Rem. Rts : 2           Total VPN-IPv6 Rem. Act. Rts: 2
Total VPN-IPv4 Bkup Rts : 0           Total VPN-IPv6 Bkup Rts     : 0

Total VPN Supp. Rts     : 0           Total VPN Hist. Rts         : 0
Total VPN Decay Rts     : 0
---snip---
===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
                   AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.4
PE-4
                64502      52    0 00h23m09s 2/2/2 (VpnIPv6)
                           54    0
-------------------------------------------------------------------------------
*A:PE-3#
```

PE-3 has received and learned a valid and best route for IPv6 prefix 2001:DB8:3000::/64 with a BGP next hop of 2001:DB8:1000:: (CE-1), as follows:

```
*A:PE-3# show router 1 bgp routes ipv6
===============================================================================
 BGP Router ID:192.0.2.6        AS:64502       Local AS:64502
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv6 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>?  2001:db8:3000::/64                                 None        None
      2001:db8:1000::1                                   None        -
      64501
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-3#
```

The following output shows the advertised IPv6 prefix of 2001:DB8:4000::/64 advertised by PE-3 to eBGP peer CE-1.

```
*A:PE-3# show router 1 bgp neighbor 2001:db8:1000::1 advertised-routes ipv6
===============================================================================
 BGP Router ID:192.0.2.6        AS:64502       Local AS:64502
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv6 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
?     2001:db8:4000::/64                                 n/a         None
      2001:db8:1000::2                                   None        -
      64502 64502
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-3#
```

The IPv6 route 2001:DB8:4000::/64 originates from CE-5 and was advertised from CE-5 to its eBGP peer PE-4, then from PE-4 as VPN-IPv6 route to its iBGP peer PE-3 with next-hop PE-4, as follows:

```
*A:PE-3# show router bgp routes vpn-ipv6
===============================================================================
 BGP Router ID:192.0.2.3        AS:64502       Local AS:64502
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv6 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
```

```
-------------------------------------------------------------------------------
u*>i  64502:1:2001:db8:2000::/64                         100        None
      ::ffff:192.0.2.4                                   None       262140
      No As-Path
u*>?  64502:1:2001:db8:4000::/64                         100        None
      ::ffff:192.0.2.4                                   None       262140
      64501
-------------------------------------------------------------------------------
Routes : 2
```

PE-3 is advertising the VPN-IPv6 route of 2001:DB8:3000::/64 to its MP-iBGP peer PE-4, as follows. The IPv6 prefix 2001:DB8:3000::/64 was learned from CE-1 in an MP-eBGP session.

```
*A:PE-3# show router bgp neighbor 192.0.2.4 advertised-routes vpn-ipv6
===============================================================================
 BGP Router ID:192.0.2.3        AS:64502       Local AS:64502
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP VPN-IPv6 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
i     64502:1:2001:db8:1000::/64                         100         None
      ::ffff:192.0.2.3                                   None        262140
      No As-Path
?     64502:1:2001:db8:3000::/64                         100         None
      ::ffff:192.0.2.3                                   None        262140
      64501
-------------------------------------------------------------------------------
Routes : 2
```

The list of VPN-IPv6 routes on PE-4 includes the VPN-IPv6 route that is being learned from PE-3: 2001:DB8:3000::/64, as follows:

```
*A:PE-4# show router bgp routes vpn-ipv6
===============================================================================
 BGP Router ID:192.0.2.4        AS:64502       Local AS:64502
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP VPN-IPv6 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
```

```
u*>i  64502:1:2001:db8:1000::/64                           100         None
      ::ffff:192.0.2.3                                     None        262140
      No As-Path
u*>?  64502:1:2001:db8:3000::/64                           100         None
      ::ffff:192.0.2.3                                     None        262140
      64501
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-4#
```

The following output shows the advertised VPN-IPv6 route of 2001:DB8:4000::/64 from PE-4 to PE-3.

```
*A:PE-4# show router bgp neighbor 192.0.2.3 advertised-routes vpn-ipv6
===============================================================================
 BGP Router ID:192.0.2.4        AS:64502       Local AS:64502
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv6 Routes
===============================================================================
Flag  Network                                              LocalPref   MED
      Nexthop (Router)                                     Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
i     64502:1:2001:db8:2000::/64                           100         None
      ::ffff:192.0.2.4                                     None        262140
      No As-Path
?     64502:1:2001:db8:4000::/64                           100         None
      ::ffff:192.0.2.4                                     None        262140
      64501
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-4#
```

The following output from PE-4 shows the IPv6 prefix 2001:DB8:4000::/64 learned from CE-5.

```
*A:PE-4# show router 1 bgp routes ipv6
===============================================================================
 BGP Router ID:192.0.2.7        AS:64502       Local AS:64502
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv6 Routes
===============================================================================
Flag  Network                                              LocalPref   MED
```

```
      Nexthop (Router)                                    Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>?  2001:db8:4000::/64                                  None        None
      2001:db8:2000::2                                    None        -
      64501
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-4#
```

The following output verifies the advertisement of IPv6 prefix 2001:DB8:3000::/64 to CE-5.

```
*A:PE-4# show router 1 bgp neighbor 2001:db8:2000::2 advertised-routes ipv6
===============================================================================
 BGP Router ID:192.0.2.7        AS:64502       Local AS:64502
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv6 Routes
===============================================================================
Flag  Network                                             LocalPref   MED
      Nexthop (Router)                                    Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
?     2001:db8:3000::/64                                  n/a         None
      2001:db8:2000::1                                    None        -
      64502 64502
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-4#
```

The final verification of CE-1 and CE-5 shows that IPv6 routes for AS 64501 have been received and are valid across the VPRN service, as follows:

```
*A:CE-1# show router bgp routes ipv6
===============================================================================
 BGP Router ID:192.0.2.1        AS:64501       Local AS:64501
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv6 Routes
===============================================================================
Flag  Network                                             LocalPref   MED
      Nexthop (Router)                                    Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
```

```
u*>?  2001:db8:4000::/64                                  None        None
      2001:db8:1000::2                                    None        -
      64502 64502
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:CE-1#


*A:CE-5# show router bgp routes ipv6
===============================================================================
 BGP Router ID:192.0.2.5        AS:64501       Local AS:64501
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv6 Routes
===============================================================================
Flag  Network                                             LocalPref   MED
      Nexthop (Router)                                    Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>?  2001:db8:3000::/64                                  None        None
      2001:db8:2000::1                                    None        -
      64502 64502
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:CE-5#
```

Conclusion

Spoke termination for IPv6-6VPE extends the use of spoke terminated interfaces from an Epipe VLL into a VPRN service using IPv6 interfaces on the access. Supporting the requirement of IPv6 interfaces, routing of IPv6 prefixes and the use of 6VPE for IPv6 tunneling over an IPv4 network allows SR OS to provide capabilities to support the growth of IPv6 architectures. This chapter provides examples of this feature with **show** commands for guidance.

Traffic Leaking from VPRN to GRT

This chapter provides information about Traffic Leaking from VPRN to GRT.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

The information and configuration in this chapter was originally written for SR OS release 14.0 R4. The CLI is updated to SR OS release 15.0.R4.

Overview

RFC 4364 *BGP/MPLS IP Virtual Private Networks (VPNs)* describes a method of distributing routing information using BGP and MPLS forwarding data to provide a Layer 3 VPN service to end users. Each virtual private routed network (VPRN) consists of a set of customer sites connected to one or more PE routers. Each associated PE router maintains a separate IP forwarding table for each VPRN. Additionally, the PE routers exchange the routing information configured or learned from all customer sites via multi-protocol border gateway protocol (MP-BGP) peering. Each route exchanged via the MP-BGP protocol includes a route distinguisher (RD), which identifies the VPRN association and resolves any IP address overlap.

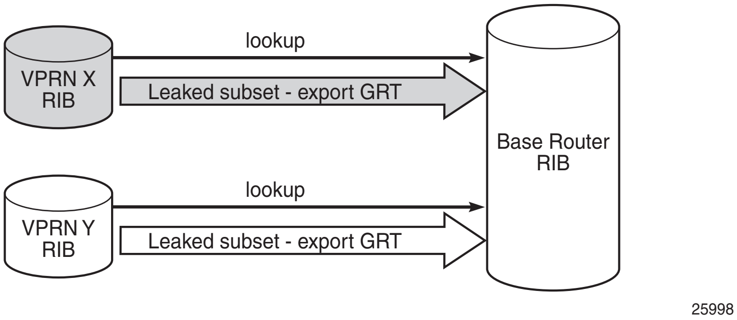
It has always been possible to exchange traffic from one VPRN to another, using scenarios such as "extranet", "hub and spoke" and so on, using the vrf-import and vrf-export policies for BGP VPN-IPv4 route distribution.

Traffic leaking to the global route table (GRT) allows service providers to offer VPRN and Internet services over a single virtual routing and forwarding VRF interface. Packets entering a VRF interface can have route processing results derived from the VRF or the GRT. The leaking and preferred lookup settings are configured on a per-VPRN basis.

To allow data flowing from a VPRN to the base router, routing information from the base router must be made available for lookup by the VPRN. The GRT lookup can be general (for example, any lookup miss in the virtual routing and forwarding (VRF) table can be resolved in the GRT), or specific (for example, specific routes should only be searched for in the GRT and ignored by the VPRN).

To enable the GRT lookup from the VPRN, the **enable-grt** command is used. This only provides part of the solution, because packets can now be forwarded from the VPRN to the GRT, but not in the opposite direction. The GRT needs to learn specific destination prefixes from the VPRN and this is achieved by route leaking from the VPRN to the GRT, using policies (**export-grt** command). The maximum number of routes leaked from a VPRN to the GRT is five by default, but this maximum can be modified or even removed. Prefixes should be globally unique within the service provider network and if these are propagated outside the provider's network, they must be from the public IP space and globally unique.

Figure 204 **VPRN to GRT leak process**

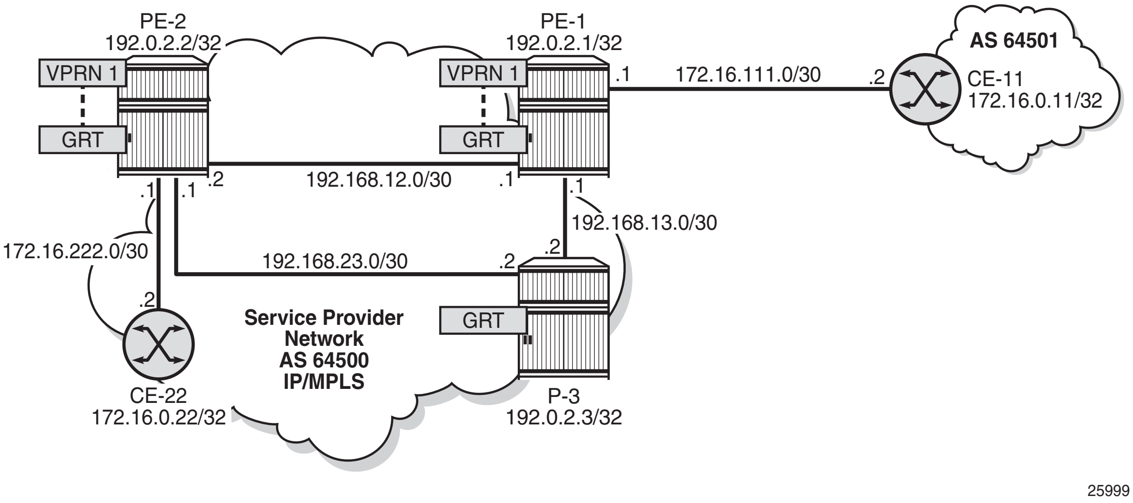


The method described in this chapter allows the network administrator to leak specific or all routes that are inside a VPRN to the GRT. Route leaking from VPRN to GRT is protocol-independent and can be applied for BGP, OSPF(v3), IS-IS, static, local routes, and so on. For BGP routes, there is an improved route leaking mechanism that allows leaking routes preserving all BGP attributes; see chapter *BGP Route Leaking*.

Configuration

Figure 205 shows the example topology used in this chapter, including the IPv4 addresses. The interfaces also have IPv6 addresses, which will be shown in Figure 207.

Figure 205 **Example Topology with IPv4 Addresses**



Initial Configuration

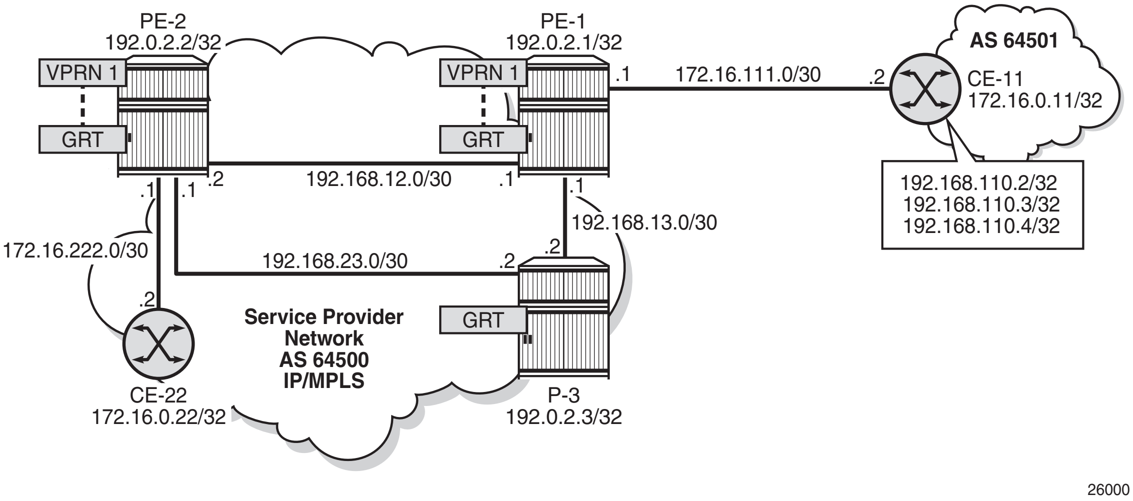
The nodes in the example topology have the following initial configuration:

- Cards, MDAs, ports
- Router interfaces
- IGP (IS-IS or OSPF) between the PEs
- LDP between the PEs
- VPRN 1 on PE-1 and PE-2
- BGP (IBGP between the PEs and between PE-2 and CE-22; EBGP between PE-1 and CE-11)
    - On the PEs, BGP is configured in the base router and in the VPRNs.
- Loopback addresses on CE-11, such as 192.168.110.2/32.
- Export policies on CE-11 to export routes from direct with certain prefixes.

Protocol-independent IPv4 Route Leaking from VPRN to GRT

Figure 206 shows the topology with the IP addresses for this example. Route leaking from VPRN to GRT is protocol independent and in this example, VPRN 1 on PE-1 will leak local routes, static routes, and imported BGP routes to the GRT. VPRN 1 on PE-2 will leak local routes and IS-IS routes. OSPF routes can also be leaked, but that is not shown here.

Figure 206 **IPv4 VPRN to GRT route leaking for ISIS**



GRT-leak is by default disabled. The routing table for VPRN 1 on PE-1 contains local routes, static routes, and BGP routes that are learned from CE-11, as follows:

```
*A:PE-1# show router 1 route-table

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.1.1/32                                 Local   Local     00h06m18s  0
       system                                                        0
172.16.111.0/30                               Local   Local     00h06m18s  0
       int-PE-1-CE-11                                                0
192.168.110.2/32                              Remote  BGP       00h03m00s  170
       172.16.111.2                                                  0
192.168.110.3/32                              Remote  BGP       00h03m00s  170
       172.16.111.2                                                  0
192.168.110.4/32                              Remote  BGP       00h03m00s  170
       172.16.111.2                                                  0
192.168.120.0/24                              Remote  Static    00h00m14s  5
```

```
       172.16.111.2                                                 1
-------------------------------------------------------------------------------
No. of Routes: 6
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

By default, the GRT is not learning the VPRN routes, as follows:

```
*A:PE-1# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Local   Local     00h06m18s  0
       system                                                       0
192.0.2.2/32                                  Remote  ISIS      00h05m51s  15
       192.168.12.2                                                 10
192.0.2.3/32                                  Remote  ISIS      00h05m40s  15
       192.168.13.2                                                 10
192.168.12.0/30                               Local   Local     00h06m18s  0
       int-PE-1-PE-2                                                0
192.168.13.0/30                               Local   Local     00h06m18s  0
       int-PE-1-P-3                                                 0
192.168.23.0/30                               Remote  ISIS      00h05m51s  15
       192.168.12.2                                                 20
-------------------------------------------------------------------------------
No. of Routes: 6
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

On PE-2, the routing table for VPRN 1 contains local routes and IS-IS routes, as follows:

```
*A:PE-2# show router 1 route-table

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.1.2/32                                 Local   Local     00h06m02s  0
       system                                                       0
172.16.2.22/32                                Remote  ISIS      00h05m10s  15
       172.16.222.2                                                 10
172.16.222.0/30                               Local   Local     00h06m02s  0
```

```
       int-PE-2-CE-22                                                0
--------------------------------------------------------------------------------
No. of Routes: 3
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
================================================================================
*A:PE-2#
```

By default, none of these VPRN routes are leaked to the GRT, as follows:

```
*A:PE-2# show router route-table

================================================================================
Route Table (Router: Base)
================================================================================
Dest Prefix[Flags]                             Type    Proto     Age        Pref
      Next Hop[Interface Name]                                     Metric
--------------------------------------------------------------------------------
192.0.2.1/32                                   Remote  ISIS      00h05m57s  15
       192.168.12.1                                                  10
192.0.2.2/32                                   Local   Local     00h06m03s  0
       system                                                        0
192.0.2.3/32                                   Remote  ISIS      00h05m45s  15
       192.168.23.2                                                  10
192.168.12.0/30                                Local   Local     00h06m03s  0
       int-PE-2-PE-1                                                 0
192.168.13.0/30                                Remote  ISIS      00h05m57s  15
       192.168.12.1                                                  20
192.168.23.0/30                                Local   Local     00h06m03s  0
       int-PE-2-P-3                                                  0
--------------------------------------------------------------------------------
No. of Routes: 6
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
================================================================================
*A:PE-2#
```

To enable VPRN to GRT leaking, the following route policy is configured in VPRN 1, on PE-1:

```
configure
    router
        policy-options
            begin
            policy-statement "LeakVPRNtoGRT_pref8"
                entry 10
                    action accept
                        preference 8
                    exit
                exit
            exit
            commit
        exit
```

This policy allows leaking all routes from a VPRN to the base router, without any match criteria. However, when routes are leaked from VPRNs to the GRT, they need to be unique and only routes that need to be known in the GRT should be leaked. By default, the preference for a leaked route is 180. The preference can be manually configured to a lower value, such as 8, to avoid network inconsistencies between the IGP and the RT on the router where the routes are leaked.

On PE-2, a similar policy is applied, but without the preference configuration, as follows. This implies that the routes leaked on PE-2 will have default preference 180.

```
configure
    router
        policy-options
            begin
            policy-statement "LeakVPRNtoGRT"
                entry 10
                    action accept
                    exit
                exit
            exit
            commit
        exit
```

The policy must be applied to VPRN 1, as follows:

```
configure
    service
        vprn 1
            grt-lookup
                enable-grt
                exit
                export-grt "Leak-VPRN-toGRT"
            exit
        exit
```

When **enable-grt** is configured, any lookup miss in the VRF table will be resolved in the GRT, if available. This only works from VPRN to GRT and does not require route leaking. However, the base router needs to be able to route packets back to the VPRN and it cannot perform a lookup in the routing table of the VPRN. Therefore, route leaking from VPRN to GRT is required, and **export-grt** is configured. Prefixes in the VPRN must be leaked to the GRT through a policy. Prefixes leaked from any VPRN should never conflict with prefixes leaked from any other VPRN or existing prefixes in the GRT.

This configuration is protocol-independent. Route leaking from VPRN to GRT is applicable for all kinds of learned routes, such as static routes, local routes, IS-IS, OSPF, BGP, and so on.

After routes are leaked from the VPRN to the GRT, the routing table of the base router includes the leaked routes, with protocol "VPN Leak". For PE-1, the routing table contains the following routes:

```
*A:PE-1# show router route-table
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.1.1/32                                 Remote  VPN Leak  00h02m40s  8
       system                                                       0
172.16.111.0/30                               Remote  VPN Leak  00h02m40s  8
       int-PE-1-CE-11                                               0
192.0.2.1/32                                  Local   Local     00h15m01s  0
       system                                                       0
192.0.2.2/32                                  Remote  ISIS      00h14m33s  15
       192.168.12.2                                                 10
192.0.2.3/32                                  Remote  ISIS      00h14m22s  15
       192.168.13.2                                                 10
192.168.12.0/30                               Local   Local     00h15m01s  0
       int-PE-1-PE-2                                                0
192.168.13.0/30                               Local   Local     00h15m01s  0
       int-PE-1-P-3                                                 0
192.168.23.0/30                               Remote  ISIS      00h14m33s  15
       192.168.12.2                                                 20
192.168.110.3/32                              Remote  VPN Leak  00h02m40s  8
       172.16.111.2                                                 0
192.168.110.4/32                              Remote  VPN Leak  00h02m40s  8
       172.16.111.2                                                 0
192.168.120.0/24                              Remote  VPN Leak  00h02m40s  8
       172.16.111.2                                                 0
-------------------------------------------------------------------------------
No. of Routes: 11
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

Regardless the preference of the original routes in VPRN 1, all the leaked routes in the GRT have preference 8, as configured. By default, only five routes are leaked. This export limit can be overruled, as follows:

```
*A:PE-1# configure service vprn 1 grt-lookup export-limit 10
```

The following command shows only the routes leaked from any VPRN to GRT on PE-1:

```
*A:PE-1# show router route-table protocol vpn-leak all

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                         Active     Metric
-------------------------------------------------------------------------------
172.16.1.1/32                                 Remote  VPN Leak  00h00m21s  8
```

```
       system                                          Y            0
172.16.111.0/30                               Remote  VPN Leak  00h00m21s  8
       int-PE-1-CE-11                                  Y            0
192.168.110.2/32                              Remote  VPN Leak  00h00m21s  8
       172.16.111.2                                    Y            0
192.168.110.3/32                              Remote  VPN Leak  00h00m21s  8
       172.16.111.2                                    Y            0
192.168.110.4/32                              Remote  VPN Leak  00h00m21s  8
       172.16.111.2                                    Y            0
192.168.120.0/24                              Remote  VPN Leak  00h00m21s  8
       172.16.111.2                                    Y            0
-------------------------------------------------------------------------------
No. of Routes: 6
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
       E = Inactive best-external BGP route
===============================================================================
*A:PE-1#
```

On PE-2, the following routes are leaked from VPRN to GRT:

```
*A:PE-2# show router route-table protocol vpn-leak all

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                         Active     Metric
-------------------------------------------------------------------------------
172.16.1.2/32                                 Remote  VPN Leak  00h03m16s   180
       system                                          Y            0
172.16.2.22/32                                Remote  VPN Leak  00h03m16s   180
       172.16.222.2                                    Y            0
172.16.222.0/30                               Remote  VPN Leak  00h03m16s   180
       int-PE-2-CE-22                                  Y            0
-------------------------------------------------------------------------------
No. of Routes: 3
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
       E = Inactive best-external BGP route
===============================================================================
*A:PE-2#
```

Different types of routes are leaked to the GRT with protocol type "VPN Leak" and all of them get the same preference, configured or default. The detailed output for any leaked route in the preceding list for PE-1 shows protocol VPN_LEAK and preference 8, as follows:

```
*A:PE-1# show router route-table protocol vpn-leak 192.168.110.2/32 extensive

===============================================================================
Route Table (Router: Base)
```

```
===============================================================================
Dest Prefix             : 192.168.110.2/32
  Protocol              : VPN_LEAK
  Age                   : 00h01m16s
  Preference            : 8
  Indirect Next-Hop     : 172.16.111.2
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : N/A
    Resolving Next-Hop  : 172.16.111.2
      Interface         : int-PE-1-CE-11 (VPRN 1)
      Metric            : 0
      ECMP-Weight       : N/A
-------------------------------------------------------------------------------
No. of Destinations: 1
===============================================================================
*A:PE-1#
```

Export IPv4 VPN-Leak Routes to Routing Protocols

Until now, the VPN-leak routes are leaked locally to the GRT, but they are not advertised in IS-IS, OSPF, or BGP. Router P-3 has not learned any of the leaked routes, as follows:

```
*A:P-3# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  ISIS      00h16m53s  15
       192.168.13.1                                                 10
192.0.2.2/32                                  Remote  ISIS      00h16m53s  15
       192.168.23.1                                                 10
192.0.2.3/32                                  Local   Local     00h17m00s  0
       system                                                       0
192.168.12.0/30                               Remote  ISIS      00h16m53s  15
       192.168.13.1                                                 20
192.168.13.0/30                               Local   Local     00h17m00s  0
       int-P-3-PE-1                                                 0
192.168.23.0/30                               Local   Local     00h17m00s  0
       int-P-3-PE-2                                                 0
-------------------------------------------------------------------------------
No. of Routes: 6
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:P-3#
```

To reduce the number of routes to be exported, the export will only be configured on PE-1 and a match criterion will be added for the routes to be leaked, as follows:

```
configure
    router
        policy-options
            begin
            prefix-list "192.168.110.0"
                prefix 192.168.110.0/24 longer
            exit
            policy-statement "LeakVPRNtoGRT_pref8_110"
                entry 10
                    from
                        prefix-list "192.168.110.0"
                    exit
                    action accept
                        preference 8
                    exit
                exit
            exit
            commit

configure
    service
        vprn 1
            grt-lookup
                export-grt "LeakVPRNtoGRT_pref8_110"
            exit
        exit
```

VPN-leak routes can be exported to any routing protocol. Prefix lists can be used to filter routes, but that is not configured in this example. The following export policy is configured on PE-1 to export the VPN-leak routes:

```
configure
    router
        policy-options
            begin
            policy-statement "export-vpn-leak"
                entry 10
                    from
                        protocol vpn-leak
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

The same export policy will be used for export to IS-IS, OSPF, and BGP.

Export IPv4 VPN-Leak Routes to IS-IS

The export policy is applied in the IS-IS context on PE-1, as follows:

```
configure
    router
        isis
            export "export-vpn-leak"
        exit
    exit
exit
```

The leaked routes are now advertised via IS-IS and appear as IS-IS routes with default preference for IS-IS routes on PE-2 and P-3. The route table on P-3 looks as follows:

```
*A:P-3# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  ISIS      00h20m56s  15
       192.168.13.1                                                 10
192.0.2.2/32                                  Remote  ISIS      00h20m56s  15
       192.168.23.1                                                 10
192.0.2.3/32                                  Local   Local     00h21m03s  0
       system                                                       0
192.168.12.0/30                               Remote  ISIS      00h20m56s  15
       192.168.13.1                                                 20
192.168.13.0/30                               Local   Local     00h21m03s  0
       int-P-3-PE-1                                                 0
192.168.23.0/30                               Local   Local     00h21m03s  0
       int-P-3-PE-2                                                 0
192.168.110.2/32                              Remote  ISIS      00h00m49s  15
       192.168.13.1                                                 10
192.168.110.3/32                              Remote  ISIS      00h00m49s  15
       192.168.13.1                                                 10
192.168.110.4/32                              Remote  ISIS      00h00m49s  15
       192.168.13.1                                                 10
-------------------------------------------------------------------------------
No. of Routes: 9
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:P-3#
```

The export policy is removed from the IS-IS context on PE-1, as follows:

```
*A:PE-1# configure router isis no export
```

Export IPv4 VPN-Leak Routes to OSPF

When OSPF is used instead of IS-IS, the behavior is similar. The export policy is applied in the OSPF context on PE-1, as follows:

```
configure
    router
        ospf
            export "export-vpn-leak"
        exit
    exit
exit
```

To export routes into OSPF using a policy, the router must be configured as ASBR, as follows:

```
*A:PE-1# configure router ospf asbr
```

The routes with protocol VPN-leak on PE-1 are now exported in OSPF to PE-2 and P-3. The default preference for external OSPF routes is 150. On PE-3, the routing table contains the following OSPF routes:

```
*A:P-3# show router route-table protocol ospf

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  OSPF      00h03m30s  10
       192.168.13.1                                                 10
192.0.2.2/32                                  Remote  OSPF      00h03m30s  10
       192.168.23.1                                                 10
192.168.12.0/30                               Remote  OSPF      00h03m30s  10
       192.168.13.1                                                 20
192.168.110.2/32                              Remote  OSPF      00h00m28s  150
       192.168.13.1                                                 1
192.168.110.3/32                              Remote  OSPF      00h00m28s  150
       192.168.13.1                                                 1
192.168.110.4/32                              Remote  OSPF      00h00m28s  150
       192.168.13.1                                                 1
-------------------------------------------------------------------------------
No. of Routes: 6
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:P-3#
```

The export policy is removed from the OSPF context on PE-1 as follows:

```
*A:PE-1# configure router ospf no export
```

Export IPv4 VPN-Leak Routes to BGP

The export policy is applied in the general BGP context of PE-1, as follows:

```
configure
    router
        bgp
            export "export-vpn-leak"
        exit
    exit
exit
```

The VPN-leak routes from PE-1 will be advertised as BGP routes to BGP neighbors PE-2 and PE-3, and the routing tables will contain BGP routes with preference 170. P-3 has the following BGP routes:

```
*A:P-3# show router route-table protocol bgp

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.168.110.2/32                              Remote  BGP        00h00m40s  170
       192.168.13.1                                                 0
192.168.110.3/32                              Remote  BGP        00h00m40s  170
       192.168.13.1                                                 0
192.168.110.4/32                              Remote  BGP        00h00m40s  170
       192.168.13.1                                                 0
-------------------------------------------------------------------------------
No. of Routes: 3
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:P-3#
```

If it is required to preserve the BGP path attributes in the leaking process, you must use the BGP Route Leaking process described in chapter *BGP Route Leaking*. However, with this protocol-independent route leaking mechanism, it is possible to leak non-BGP routes to the GRT that will be advertised as BGP routes.

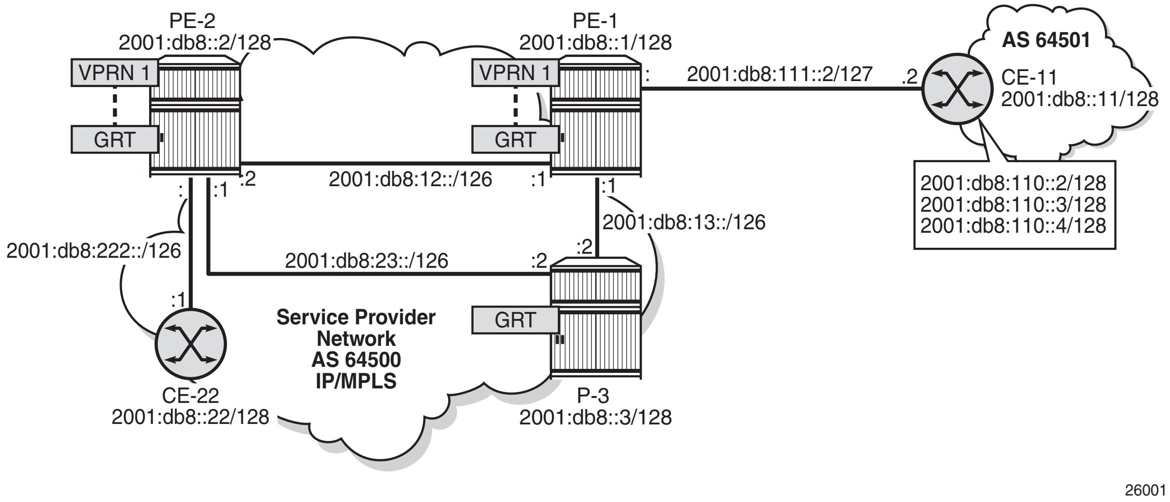
The export policy is removed from the BGP context, as follows:

```
*A:PE-1# configure router bgp no export
```

Protocol-independent IPv6 Route Leaking from VPRN to GRT

Figure 207 shows the topology and the IP addresses used for IPv6. CE-11 exports routes such as 2001:db8:110::2/128 to VPRN 1on PE-1. On PE-1, local routes, static routes, and BGP routes will be leaked to the GRT. On PE-2, IS-IS routes and local routes are leaked to the GRT.

Figure 207 **Example Topology with IPv6 Addresses**



The IPv6 routing table for VPRN 1 on PE-1 includes local addresses, a static route, and three BGP routes exported by CE-11, as follows:

```
*A:PE-1# show router 1 route-table ipv6

===============================================================================
IPv6 Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
2001:db8::1:1/128                             Local   Local     00h31m16s  0
       system                                                       0
2001:db8:110::2/128                           Remote  BGP       00h27m59s  170
       2001:db8:111::1                                              0
2001:db8:110::3/128                           Remote  BGP       00h27m59s  170
       2001:db8:111::1                                              0
2001:db8:110::4/128                           Remote  BGP       00h27m59s  170
       2001:db8:111::1                                              0
2001:db8:111::/127                            Local   Local     00h31m17s  0
       int-PE-1-CE-11                                               0
2001:db8:120::/120                            Remote  Static    00h00m30s  5
       2001:db8:111::1                                              1
-------------------------------------------------------------------------------
```

```
No. of Routes: 6
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

By default, route leaking is disabled and the IPv6 GRT on PE-1 does not contain any of the IPv6 routes in VPRN 1, as follows:

```
*A:PE-1# show router route-table ipv6

===============================================================================
IPv6 Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
2001:db8::1/128                               Local   Local     00h31m18s  0
       system                                                       0
2001:db8::2/128                               Remote  OSPF3     00h30m56s  10
       fe80::b:1ff:fe01:1-"int-PE-1-PE-2"                           10
2001:db8::3/128                               Remote  OSPF3     00h30m35s  10
       fe80::d:1ff:fe01:2-"int-PE-1-P-3"                            10
2001:db8:12::/126                             Local   Local     00h31m17s  0
       int-PE-1-PE-2                                                0
2001:db8:13::/126                             Local   Local     00h31m17s  0
       int-PE-1-P-3                                                 0
2001:db8:23::/126                             Remote  OSPF3     00h30m35s  10
       fe80::d:1ff:fe01:2-"int-PE-1-P-3"                            20
-------------------------------------------------------------------------------
No. of Routes: 6
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

The VPN-leak route policy is the same as for IPv4 routes, and is applied in the VPRN context in the same way as for IPv4 routes, as follows:

```
configure
    router
        policy-options
            begin
            policy-statement "LeakVPRNtoGRT_pref8"
                entry 10
                    action accept
                        preference 8
                    exit
                exit
            exit
            commit
```

```
configure
    service
        vprn 1
            grt-lookup
                enable-grt
                exit
                export-grt "LeakVPRNtoGRT_pref8"
            exit
        exit
    exit
exit
```

On PE-1, the VPN leak policy sets the preference to 8, and on PE-2, the default preference is used, as in the preceding example for IPv4.

On PE-2, the IPv6 routing table for VPRN1 contains two local routes and one IS-IS route, as follows:

```
*A:PE-2# show router 1 route-table ipv6

===============================================================================
IPv6 Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
2001:db8:1::2/128                             Local   Local     00h35m08s  0
       system                                                        0
2001:db8:2::22/128                            Remote  ISIS      00h34m17s  15
       fe80::11:1ff:fe01:1-"int-PE-2-CE-22"                          10
2001:db8:222::/127                            Local   Local     00h35m10s  0
       int-PE-2-CE-22                                                0
-------------------------------------------------------------------------------
No. of Routes: 3
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-2#
```

All of these routes are leaked to the GRT on PE-2 with default preference 180, as follows:

```
*A:PE-2# show router route-table ipv6 protocol vpn-leak all

===============================================================================
IPv6 Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                         Active     Metric
-------------------------------------------------------------------------------
2001:db8:1::2/128                             Remote  VPN Leak  00h03m45s  180
       system                                          Y             0
2001:db8:2::22/128                            Remote  VPN Leak  00h03m45s  180
       fe80::11:1ff:fe01:1-"int-PE-2-CE-22"            Y             0
```

```
2001:db8:222::/127                            Remote  VPN Leak  00h03m45s  180
       int-PE-2-CE-22                                  Y            0
-------------------------------------------------------------------------------
No. of Routes: 3
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
       E = Inactive best-external BGP route
===============================================================================
*A:PE-2#
```

On PE-1, the IPv6 routing table for VPRN 1 contains six routes, but by default, a maximum of five routes are leaked, as follows:

```
*A:PE-1# show router route-table ipv6 protocol vpn-leak all

===============================================================================
IPv6 Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                         Active     Metric
-------------------------------------------------------------------------------
2001:db8::1:1/128                             Remote  VPN Leak  00h03m52s  8
       system                                          Y            0
2001:db8:110::2/128                           Remote  VPN Leak  00h03m52s  8
       2001:db8:111::1                                 Y            0
2001:db8:110::4/128                           Remote  VPN Leak  00h03m52s  8
       2001:db8:111::1                                 Y            0
2001:db8:111::/127                            Remote  VPN Leak  00h03m52s  8
       int-PE-1-CE-11                                  Y            0
2001:db8:120::/120                            Remote  VPN Leak  00h03m52s  8
       2001:db8:111::1                                 Y            0
-------------------------------------------------------------------------------
No. of Routes: 5
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
       E = Inactive best-external BGP route
===============================================================================
*A:PE-1#
```

The export limit for IPv6 routes is removed, as follows:

```
*A:PE-1# configure service vprn 1 grt-lookup export-v6-limit 0
```

As a result, there is no limit to the number of leaked IPv6 routes, and all six IPv6 routes are leaked from VPRN 1 to the GRT with the configured preference 8, as follows:

```
*A:PE-1# show router route-table ipv6 protocol vpn-leak

===============================================================================
IPv6 Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
2001:db8::1:1/128                             Remote  VPN Leak  00h00m23s  8
       system                                                       0
2001:db8:110::2/128                           Remote  VPN Leak  00h00m23s  8
       2001:db8:111::1                                              0
2001:db8:110::3/128                           Remote  VPN Leak  00h00m23s  8
       2001:db8:111::1                                              0
2001:db8:110::4/128                           Remote  VPN Leak  00h00m23s  8
       2001:db8:111::1                                              0
2001:db8:111::/127                            Remote  VPN Leak  00h00m23s  8
       int-PE-1-CE-11                                               0
2001:db8:120::/120                            Remote  VPN Leak  00h00m23s  8
       2001:db8:111::1                                              0
-------------------------------------------------------------------------------
No. of Routes: 6
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

The details for any of the routes shows that the protocol is VPN-leak and the preference is 8, as follows:

```
*A:PE-1# show router route-table protocol vpn-leak 2001:db8:110::2/128 extensive

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix             : 2001:db8:110::2/128
  Protocol              : VPN_LEAK
  Age                   : 00h00m51s
  Preference            : 8
  Indirect Next-Hop     : 2001:db8:111::1
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : N/A
    Resolving Next-Hop  : 2001:db8:111::1
      Interface         : int-PE-1-CE-11 (VPRN 1)
      Metric            : 0
      ECMP-Weight       : N/A
-------------------------------------------------------------------------------
No. of Destinations: 1
===============================================================================
*A:PE-1#
```

Export IPv6 VPN-Leak Routes to Routing Protocols

Until now, the IPv6 VPN-leak routes are leaked locally to the GRT, but they are not advertised in IS-IS, OSPFv3, or BGP. Router P-3 has not learned any of the leaked IPv6 routes, as follows:

```
*A:P-3# show router route-table ipv6

===============================================================================
IPv6 Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
2001:db8::1/128                               Remote  OSPF3     00h54m09s  10
       fe80::9:1ff:fe01:1-"int-P-3-PE-1"                            10
2001:db8::2/128                               Remote  OSPF3     00h54m15s  10
       fe80::b:1ff:fe01:2-"int-P-3-PE-2"                            10
2001:db8::3/128                               Local   Local     00h54m16s  0
       system                                                       0
2001:db8:12::/126                             Remote  OSPF3     00h54m09s  10
       fe80::9:1ff:fe01:1-"int-P-3-PE-1"                            20
2001:db8:13::/126                             Local   Local     00h54m14s  0
       int-P-3-PE-1                                                 0
2001:db8:23::/126                             Local   Local     00h54m15s  0
       int-P-3-PE-2                                                 0
-------------------------------------------------------------------------------
No. of Routes: 6
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:P-3#
```

To reduce the number of VPN-leak routes, a match criterion is added to the route policy on PE-1, as follows:

```
configure
    router
        policy-options
            begin
            prefix-list "2001:db8:110::"
                prefix 2001:db8:110::/125 longer
            exit
            policy-statement "LeakVPRNtoGRT_pref8_110"
                entry 20
                    from
                        prefix-list "2001:db8:110::"
                    exit
                    action accept
                        preference 8
                    exit
                exit
            exit
            commit
        exit


configure
    service
        vprn 1
            grt-lookup
                enable-grt
                exit
                export-grt "LeakVPRNtoGRT_pref8_110"
            exit
        exit
    exit
exit
```

The following IPv6 routes are leaked from VPRN 1 to GRT on PE-1:

```
*A:PE-1# show router route-table ipv6 protocol vpn-leak

===============================================================================
IPv6 Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
2001:db8:110::2/128                           Remote  VPN Leak  00h00m21s  8
       2001:db8:111::1                                              0
2001:db8:110::3/128                           Remote  VPN Leak  00h00m21s  8
       2001:db8:111::1                                              0
2001:db8:110::4/128                           Remote  VPN Leak  00h00m21s  8
       2001:db8:111::1                                              0
-------------------------------------------------------------------------------
No. of Routes: 3
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

IPv6 VPN-leak routes can be exported to routing protocols IS-IS, OSPFv3, and BGP.

The export policy on PE-1 is the same as in all the preceding examples for IPv4, as follows:

```
configure
    router
        policy-options
            begin
            policy-statement "export-vpn-leak"
                entry 10
                    from
                        protocol vpn-leak
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

Export IPv6 VPN-Leak Routes to IS-IS

The export policy for IPv6 routes of protocol VPN-leak is applied for IS-IS, as follows:

```
configure
    router
        isis
            export "export-vpn-leak"
        exit
    exit
exit
```

The three IPv6 VPN-leak routes from PE-1 are now advertised by IS-IS to PE-2 and P-3. The routing table on P-3 contains the following IPv6 IS-IS routes:

```
*A:P-3# show router route-table ipv6 protocol isis

===============================================================================
IPv6 Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
2001:db8:110::2/128                           Remote  ISIS      00h06m53s  15
       fe80::9:1ff:fe01:1-"int-P-3-PE-1"                            10
2001:db8:110::3/128                           Remote  ISIS      00h06m53s  15
       fe80::9:1ff:fe01:1-"int-P-3-PE-1"                            10
2001:db8:110::4/128                           Remote  ISIS      00h06m53s  15
       fe80::9:1ff:fe01:1-"int-P-3-PE-1"                            10
-------------------------------------------------------------------------------
No. of Routes: 3
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:P-3#
```

The export policy is removed for IS-IS, as follows:

```
*A:PE-1# configure router isis no export
```

Export IPv6 VPN-Leak Routes to OSPFv3

The export policy for IPv6 routes of protocol VPN-leak is applied for OSPFv3, as follows:

```
configure
    router
        ospf3
            export "export-vpn-leak"
        exit
    exit
exit
```

Routes can only be exported to OSPFv3 if the router is configured as ASBR, as follows:

```
*A:PE-1# configure router ospf3 asbr
```

The IPv6 VPN-leak routes from PE-1 are now advertised by OSPFv3 to PE-2 and P-3. The routing table on P-3 contains the following IPv6 OSPFv3 routes:

```
*A:P-3# show router route-table ipv6 protocol ospf3

===============================================================================
IPv6 Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
2001:db8::1/128                               Remote  OSPF3     01h18m10s  10
       fe80::9:1ff:fe01:1-"int-P-3-PE-1"                            10
2001:db8::2/128                               Remote  OSPF3     01h18m15s  10
       fe80::b:1ff:fe01:2-"int-P-3-PE-2"                            10
2001:db8:12::/126                             Remote  OSPF3     01h18m10s  10
       fe80::9:1ff:fe01:1-"int-P-3-PE-1"                            20
2001:db8:110::2/128                           Remote  OSPF3     00h13m27s  150
       fe80::9:1ff:fe01:1-"int-P-3-PE-1"                            1
2001:db8:110::3/128                           Remote  OSPF3     00h13m27s  150
       fe80::9:1ff:fe01:1-"int-P-3-PE-1"                            1
2001:db8:110::4/128                           Remote  OSPF3     00h13m27s  150
       fe80::9:1ff:fe01:1-"int-P-3-PE-1"                            1
-------------------------------------------------------------------------------
No. of Routes: 6
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:P-3#
```

The preference for remote OSPFv3 routes is by default 150. The export policy is removed for OSPFv3, as follows:

```
*A:PE-1# configure router ospf3 no export
```

Export IPv6 VPN-Leak Routes to BGP

The export policy for IPv6 routes of protocol VPN-leak is applied for BGP, as follows:

```
configure
    router
        bgp
            export "export-vpn-leak"
        exit
    exit
exit
```

The three IPv6 VPN-leak routes from PE-1 are now advertised by BGP to PE-2 and P-3. The routing table on P-3 contains the following IPv6 BGP routes:

```
*A:P-3# show router route-table ipv6 protocol bgp

===============================================================================
IPv6 Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
2001:db8:110::2/128                           Remote  BGP       00h02m44s  170
       fe80::9:1ff:fe01:1-"int-P-3-PE-1"                            0
2001:db8:110::3/128                           Remote  BGP       00h02m44s  170
       fe80::9:1ff:fe01:1-"int-P-3-PE-1"                            0
2001:db8:110::4/128                           Remote  BGP       00h02m44s  170
       fe80::9:1ff:fe01:1-"int-P-3-PE-1"                            0
-------------------------------------------------------------------------------
No. of Routes: 3
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:P-3#
```

The export policy is removed for BGP, as follows:

```
*A:PE-1# configure router bgp no export
```

In this example, BGP leaked IPv6 routes are advertised by BGP. For scenarios with only BGP routes, a dedicated BGP route leaking mechanism that preserves all attributes is preferred, as described in chapter *BGP Route Leaking*. However, with the same configuration as in this chapter, it is possible to leak non-BGP routes and advertise them using BGP.

Conclusion

Routes learned in a VPRN can be leaked to the base router and advertised using routing protocols. The mechanism described in this chapter is protocol-independent: all kinds of routes can be leaked from a VRF to the GRT: local, static, IS-IS, OSPF, BGP routes, and so on. In some cases, it might be useful to leak the routes from a VPRN to the entire network using the routing protocol, in order to access the resources defined inside the VRF. Routes that are leaked from VPRNs to the GRT must be unique in the network where they will be advertised. For BGP routes, the protocol-independent route leaking mechanism described here does not preserve the attributes, unlike the dedicated BGP route leaking feature.

VPRN Inter-AS VPN Model C

This chapter provides information about virtual private routed network (VPRN) inter-autonomous system (AS) virtual private network (VPN) model C.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for release 7.0. The CLI in the current edition corresponds to 15.0.R4. The address family label-ipv4 is supported in SR OS release 14.0.R4, and later, see chapter *Separate BGP RIBs for Labeled Routes*. There are no prerequisites for this configuration.

Overview

Introduction

Section 10 of RFC 4364, *BGP/MPLS IP Virtual Private Networks (VPNs)*, describes three potential methods for service providers to interconnect their IP-VPN (Internet Protocol — Virtual Private Network) backbones in order to provide an end-to-end MPLS-VPN where one or more sites of the VPN are connected to different service provider autonomous systems. The purpose of this chapter is to describe the configuration and troubleshooting for inter-AS VPN model C.

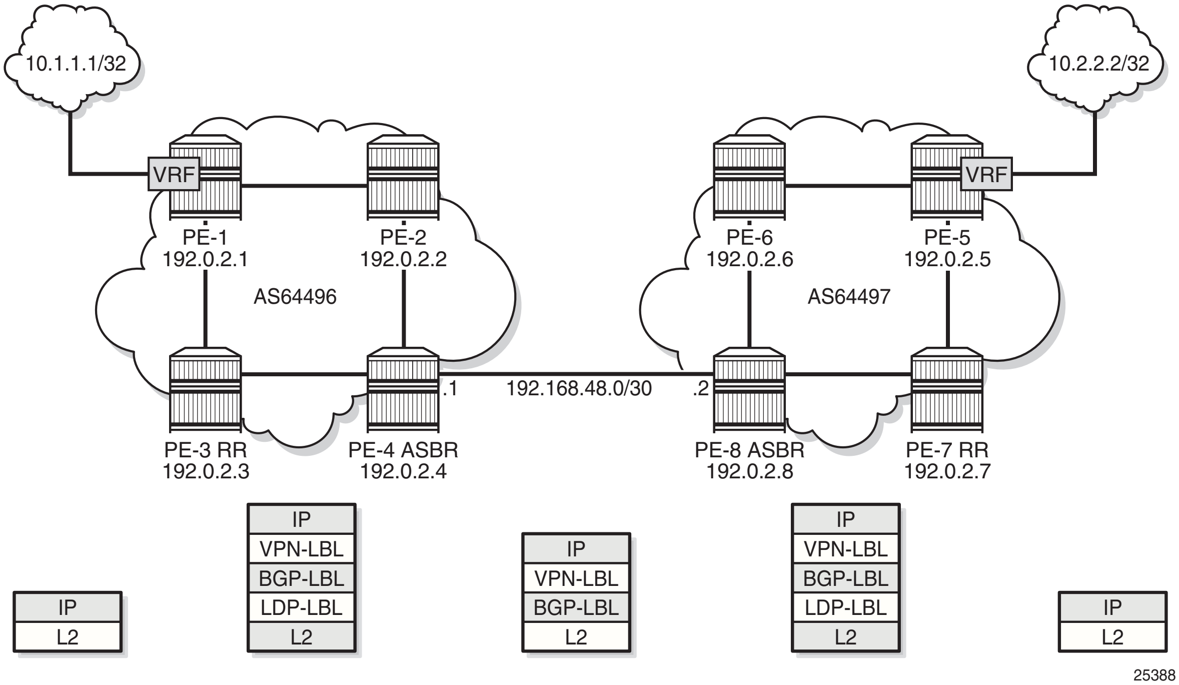
In this architecture, VPN prefixes are neither held, nor re-advertised by the Autonomous System Border Router — Provider Edge (ASBR-PE) routers, which makes Model C more scalable than Model B (where the only prefixes exchanged between ASs are VPN-IPv4). In Model C, the only MPLS data plane resources consumed in the ASBRs are for infrastructure addresses of PEs and RRs rather than VPN prefixes. In this example, an export policy is configured to ensure that the nodes advertise their system IP addresses (IPv4 /32 addresses) in labeled BGP to all their

BGP peers within the AS. Therefore, the ASBR-PE maintains labeled IPv4 /32 BGP routes to other PE routers within its own AS. These BGP routes are inactive, because for each destination within the AS, an IGP route exists which is preferred to BGP routes. The ASBR redistributes these inactive /32 IPv4 prefixes in external Border Gateway Protocol (eBGP) to the ASBR-PE in other service providers ASs, because the parameter advertise-inactive is configured in eBGP. No export policy is required in eBGP. At the same time, the ASBR programs a label switch for the received and advertised BGP labels. The receiving ASBR advertises the received IP system prefixes to its iBGP peers (in this case, a Route Reflector (RR)) within their AS, and eventually, all PEs in the AS learn the system IP prefixes of the peer AS. However, there is no need to learn the system IP address of the ASBRs in peer ASs, because they do not exchange customer VPN prefixes. After the system IP addresses have been learned in the peer AS, it is possible for PE routers in different ASs to establish multi-hop Multi Protocol — external Border Gateway Protocol (MP-eBGP) sessions for address family VPN-IPv4 to each other in order to exchange customer VPN prefixes over those connections. The multihop sessions can be established between the RR in the ASs, but these RRs should not modify the next-hop attribute of the BGP update across the eBGP session.

A three-level label stack is imposed on the ingress PE. The bottom-level label would be assigned by the egress PE (advertised in multi-hop MP-eBGP without next-hop override) and is commonly referred to as the VPN-label. The middle label would be assigned by the local ASBR-PE and would correspond to the /32 route of the egress PE (in a different AS) using BGP-LBL (RFC 3107, *Carrying Label Information in BGP-4*). The top level label would then be the label assigned by the local ASBR-PE(s) /32 loopback address, which would be assigned by the IGP next-hop of the ingress PE. This label is referred to as the LDP-LBL. Figure 208 reflects this mechanism. The VPN-LBL is assigned by PE-5, the BGP-LBL is assigned by PE-4 and the LDP-LBL is also assigned by PE-4. The BGP-LBL is swapped in both ASBRs. The label stack contains three labels in each AS: VPN-LBL, BGP-LBL, and LDP-LBL) and two labels on the eBGP link between the ASs: VPN-LBL and BGP-LBL.

Note: This configuration that uses advertise-inactive is preferred to a configuration where the BGP routes are not exchanged within their AS and the ASBRs use an export policy with a prefix list for all local system prefixes to be advertised to the peer ASs. The routes for those prefixes will be taken from the RTM, where these routes are not known via BGP, but via IS-IS. In that case, IS-IS routes are effectively redistributed into labeled BGP (which most operators do not want) and as a result, the ASBR is not programming a label switch for the BGP label. Furthermore, the label stack will be asymmetrical: three labels in the originating AS (VPN-LBL, BGP-LBL, and LDP-LBL) and only two labels in the target AS (VPN-LBL, LDP-LBL), because the local routes are not known via labeled BGP in this scenario. This scenario is not explained in this chapter; only the preferred scenario with local labeled BGP routes in each AS is explained.

Figure 208 **Inter-AS VPN Model C**



The VPN connectivity is established using Labeled VPN route exchange using MP-eBGP without next-hop override. The PE connectivity will be established as follows.

EBGP PE /32 loopback leaking routing exchange using eBGP LBL (RFC 3107) at the ASBR-PE. The /32 PE routes learned from the other AS through the ASBR-PE are further distributed into the local AS using iBGP and optionally through Route Reflectors (RRs). This model uses a three label stack and is referred to as Model C. Resilience for ASBR-PE failures depends on BGP.

Figure 209 **Protocol Overview**

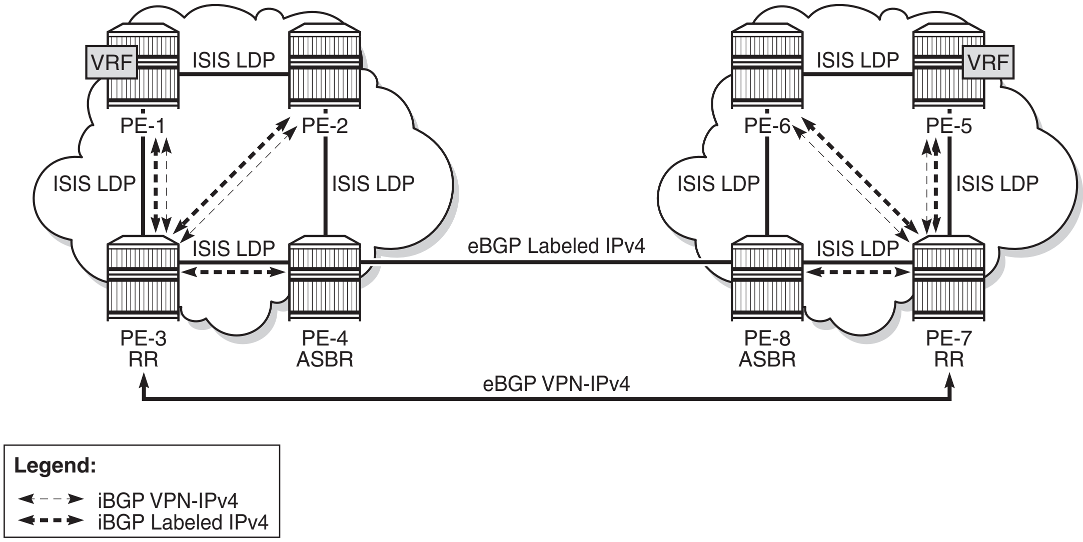


Figure 209 gives an overview of all protocols used when implementing Inter-AS Model C. Inside each AS, there is an IS-IS adjacency and a link LDP session between each pair of adjacent nodes. As an alternative, OSPF can be used as IGP. There is also an iBGP session between each PE and the RR. The address family is both VPN-IPv4 for the exchange of customer VPN prefixes and Labeled IPv4 for the exchange of labeled IPv4 prefixes. Between the RR and the ASBR, only Labeled IPv4 is required, because the ASBR will not exchange any customer VPN prefixes. When no RR is used, a full mesh of iBGP sessions can be established in each AS.

Between the ASBRs, there is an eBGP session for the exchange of labeled IPv4 prefixes. The ASBRs will override the next-hop for those prefixes. Between the RRs in the different ASs, there is a multihop eBGP session for the exchange of VPN-IPv4 customer prefixes. The RRs will not override the next-hop for those prefixes.

The main advantage of this model is that no VPN routes need to be held on the ASBR-PEs and therefore, it scales the best among all the three Inter-AS IP-VPN models. However, leaking /32 PE addresses between service providers raises some security concerns. Therefore, we see Model C typically deployed within a service provider network.

The example topology is displayed in Figure 208 and consists of two times four (2 x 4) SR OS nodes located in different autonomous systems. There is an AS interconnection from ASBR PE-4 to ASBR PE-8. PE-3 and PE-7 will serve as RRs for their AS. It is assumed that an IP-VPN is already configured in each AS. Following configuration tasks should be done first:

- IS-IS or OSPF on all interfaces within each of the ASs.
- LDP on all interfaces within each of the ASs.
- MP-iBGP sessions between the PE routers and the RRs in each of the ASs, as shown in the following section.
- IP-VPN on PE-1 and on PE-5 with identical route targets.
- A loopback interface in the VRF on PE-1 and PE-5.

Configuration

The first step is to configure an MP-iBGP session between the PEs in both ASs. An export policy is configured to export the system prefixes from the PEs in labeled BGP. PE-3 and PE-7 serve as RR in the ASs. In AS 64496, PE-1 and PE-2 are peered with RR PE-3 for the labeled IPv4 and VPN-IPv4 address families; ASBR PE-4 is peered with RR PE-3 for the labeled IPv4 address family only. In AS 64497, PE-5 and PE-6 are peered with RR PE-7 for the labeled IPv4 and VPN-IPv4 address families; ASBR PE-8 is peered with RR PE-7 for the labeled IPv4 address family only. Address family **label-ipv4** is required to advertise labeled IPv4 routes toward each neighbor PE. Address family **vpn-ipv4** is required to advertise IPv4 customer VPN routes within the AS. The configuration for RR PE-3 is as follows:

```
configure
    router
        autonomous-system 64496
        bgp
            split-horizon
            group "IBGP"
                cluster 192.0.2.3
                export "export-bgp"
                peer-as 64496
                neighbor 192.0.2.1
                    family vpn-ipv4 label-ipv4
                    advertise-inactive
                exit
                neighbor 192.0.2.2
                    family vpn-ipv4 label-ipv4
                    advertise-inactive
                exit
                neighbor 192.0.2.4
                    family label-ipv4
                    advertise-inactive
                exit
            exit
```

The export policy is defined as follows:

```
configure
    router
```

```
        policy-options
            begin
            prefix-list "PE-sys"
                prefix 192.0.2.0/28 longer
            exit
            policy-statement "export-bgp"
                entry 10
                    from
                        protocol direct
                        prefix-list "PE-sys"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

On the ASBRs in both ASs, eBGP and iBGP need to be configured. The eBGP session is configured with parameter **advertise-inactive** and will be used to redistribute labeled IPv4 routes for the /32 system IP addresses between the ASs, even if those routes are not the most preferred routes within the system for a certain destination.

The configuration for ASBR PE-4 is as follows. The address family **label-ipv4** is required to enable the advertising of labeled IPv4 routes. This address family is also required on the RR neighbor in order to propagate the labeled IPv4 routes toward the other PEs in the AS.

```
configure
    router
        autonomous-system 64496
        bgp
            split-horizon
            group "EBGP"
                neighbor 192.168.48.2
                    family label-ipv4
                    peer-as 64497
                    advertise-inactive
                exit
            exit
            group "IBGP"
                export "export-bgp"
                peer-as 64496
                neighbor 192.0.2.3
                    family label-ipv4
                exit
            exit
       exit
    exit
exit
```

On the remaining PE nodes in AS 64496, PE-1 and PE-2, the address families **label-ipv4** and **vpn-ipv4** must be enabled, as follows:

```
configure
```

```
    router
        autonomous-system 64496
        bgp
            split-horizon
            group "IBGP"
                export "export-bgp"
                peer-as 64496
                neighbor 192.0.2.3
                    family vpn-ipv4 label-ipv4
                exit
            exit
        exit
    exit
exit
```

The configuration for the nodes in AS 64497 is very similar. The IP addresses can be derived from Figure 208.

On ASBR PE-4, verify that the EBGP and IBGP sessions are up, as follows:

```
*A:PE-4# show router bgp summary
===============================================================================
 BGP Router ID:192.0.2.4        AS:64496       Local AS:64496
===============================================================================
BGP Admin State         : Up          BGP Oper State              : Up
Total Peer Groups       : 2           Total Peers                 : 2
Total VPN Peer Groups   : 0           Total VPN Peers             : 0
Total BGP Paths         : 17          Total Path Memory           : 4504

Total IPv4 Remote Rts   : 0           Total IPv4 Rem. Active Rts  : 0
Total IPv6 Remote Rts   : 0           Total IPv6 Rem. Active Rts  : 0
Total IPv4 Backup Rts   : 0           Total IPv6 Backup Rts       : 0
Total LblIpv4 Rem Rts   : 6           Total LblIpv4 Rem. Act Rts  : 3
Total LblIpv6 Rem Rts   : 0           Total LblIpv6 Rem. Act Rts  : 0
Total LblIpv4 Bkp Rts   : 0           Total LblIpv6 Bkp Rts       : 0
Total Supressed Rts     : 0           Total Hist. Rts             : 0
Total Decay Rts         : 0

Total VPN-IPv4 Rem. Rts : 0           Total VPN-IPv4 Rem. Act. Rts: 0
Total VPN-IPv6 Rem. Rts : 0           Total VPN-IPv6 Rem. Act. Rts: 0
Total VPN-IPv4 Bkup Rts : 0           Total VPN-IPv6 Bkup Rts     : 0
Total VPN Local Rts     : 0           Total VPN Supp. Rts         : 0
Total VPN Hist. Rts     : 0           Total VPN Decay Rts         : 0

Total MVPN-IPv4 Rem Rts : 0           Total MVPN-IPv4 Rem Act Rts : 0
Total MVPN-IPv6 Rem Rts : 0           Total MVPN-IPv6 Rem Act Rts : 0
Total MDT-SAFI Rem Rts  : 0           Total MDT-SAFI Rem Act Rts  : 0
Total McIPv4 Remote Rts : 0           Total McIPv4 Rem. Active Rts: 0
Total McIPv6 Remote Rts : 0           Total McIPv6 Rem. Active Rts: 0
Total McVpnIPv4 Rem Rts : 0           Total McVpnIPv4 Rem Act Rts : 0
Total McVpnIPv6 Rem Rts : 0           Total McVpnIPv6 Rem Act Rts : 0

Total EVPN Rem Rts      : 0           Total EVPN Rem Act Rts      : 0
Total L2-VPN Rem. Rts   : 0           Total L2VPN Rem. Act. Rts   : 0
Total MSPW Rem Rts      : 0           Total MSPW Rem Act Rts      : 0
Total RouteTgt Rem Rts  : 0           Total RouteTgt Rem Act Rts  : 0
Total FlowIpv4 Rem Rts  : 0           Total FlowIpv4 Rem Act Rts  : 0
Total FlowIpv6 Rem Rts  : 0           Total FlowIpv6 Rem Act Rts  : 0
Total Link State Rem Rts: 0           Total Link State Rem Act Rts: 0

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
                   AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.3
               64496       32    0 00h13m17s 3/0/4 (Lbl-IPv4)
                           33    0
192.168.48.2
               64497       29    0 00h12m27s 3/3/3 (Lbl-IPv4)
                           28    0
-------------------------------------------------------------------------------
*A:PE-4#
```

On ASBR PE-4, three inactive labeled IPv4 routes have been received from the iBGP peers and three active labeled IPv4 routes have been received via eBGP, as follows:

```
*A:PE-4# show router bgp routes label-ipv4
===============================================================================
 BGP Router ID:192.0.2.4        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
*i    192.0.2.1/32                                       100         None
      192.0.2.1                                          None        262139
      No As-Path
*i    192.0.2.2/32                                       100         None
      192.0.2.2                                          None        262139
      No As-Path
*i    192.0.2.3/32                                       100         None
      192.0.2.3                                          None        262139
      No As-Path
u*>i  192.0.2.5/32                                       None        None
      192.168.48.2                                       None        262138
      64497
u*>i  192.0.2.6/32                                       None        None
      192.168.48.2                                       None        262133
      64497
u*>i  192.0.2.7/32                                       None        None
      192.168.48.2                                       None        262134
      64497
```

```
-------------------------------------------------------------------------------
Routes : 6
===============================================================================
*A:PE-4#
```

The following three routes have been received from eBGP peer PE-8: one for each system IP address in the remote AS, except for the ASBR itself:

```
*A:PE-4# show router bgp neighbor 192.168.48.2 received-routes label-ipv4
===============================================================================
 BGP Router ID:192.0.2.4        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  192.0.2.5/32                                       n/a         None
      192.168.48.2                                       None        262138
      64497
u*>i  192.0.2.6/32                                       n/a         None
      192.168.48.2                                       None        262133
      64497
u*>i  192.0.2.7/32                                       n/a         None
      192.168.48.2                                       None        262134
      64497
-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:PE-4#
```

In this example, the IP prefix for PE-8 itself is not included. The prefix of the ASBR need not be advertised in labeled BGP to the remote AS, because ASBRs will not advertise VPN-IPv4 prefixes.

More detailed information about the advertised route from PE-5 can be seen with following command on PE-4:

```
*A:PE-4# show router bgp routes 192.0.2.5/32 label-ipv4 hunt
===============================================================================
 BGP Router ID:192.0.2.4        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP Routes
```

```
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : 192.0.2.5/32
Nexthop        : 192.168.48.2
Path Id        : None
From           : 192.168.48.2
Res. Nexthop   : 192.168.48.2
Local Pref.    : None                   Interface Name : int-PE-4-PE-8
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : No Community Members
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.8
Fwd Class      : None                   Priority       : None
IPv4 Label     : 262138                 Label Type     : SWAP
Flags          : Used  Valid  Best  IGP
Route Source   : External
AS-Path        : 64497
Route Tag      : 0
Neighbor-AS    : 64497
Orig Validation: NotFound
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h22m14s

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
Network        : 192.0.2.5/32
Nexthop        : 192.0.2.4
Path Id        : None
To             : 192.0.2.3
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : No Community Members
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.3
IPv4 Label     : 262138                 Label Type     : SWAP
Origin         : IGP
AS-Path        : 64497
Route Tag      : 0
Neighbor-AS    : 64497
Orig Validation: NotFound
Source Class   : 0                      Dest Class     : 0

-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-4#
```

In the RIB In entries, the received label from PE-8 can be seen (262138). In the RIB Out entries, the locally assigned (Advertised) label for this prefix can be seen (262138). These labels need not match. The ASBR PE-4 swaps BGP labels, according to the following label mapping:

```
*A:PE-4# show router bgp inter-as-label

===============================================================================
BGP Inter-AS labels
Flags: B - entry has backup, P - entry is promoted
===============================================================================
NextHop                       Received       Advertised     Label
                              Label          Label          Origin
-------------------------------------------------------------------------------
0.0.0.0                       0              262139         Edge
192.0.2.1                     262139         262137         Internal
192.0.2.2                     262139         262136         Internal
192.168.48.2                  262133         262134         External
192.168.48.2                  262134         262133         External
192.168.48.2                  262138         262138         External
192.0.2.3                     262139         262135         Internal
-------------------------------------------------------------------------------
Total Labels allocated:   7
===============================================================================
*A:PE-4#
```

The route from PE-1 toward PE-5 uses received label 262138 and advertised label 262138, as indicated on the sixth row in the table. The BGP label in the label stack sent by PE-1 will contain BGP label 262138 toward ASBR PE-4, where it will be swapped to BGP label 262138 toward ASBR PE-8.

ASBR PE-8 swaps BGP label 262138 to BGP label 262139 toward PE-5, as follows:

```
*A:PE-8# show router bgp inter-as-label

===============================================================================
BGP Inter-AS labels
Flags: B - entry has backup, P - entry is promoted
===============================================================================
NextHop                       Received       Advertised     Label
                              Label          Label          Origin
-------------------------------------------------------------------------------
0.0.0.0                       0              262139         Edge
192.168.48.1                  262135         262137         External
192.168.48.1                  262136         262136         External
192.168.48.1                  262137         262135         External
192.0.2.5                     262139         262138         Internal
192.0.2.6                     262139         262133         Internal
192.0.2.7                     262139         262134         Internal
-------------------------------------------------------------------------------
Total Labels allocated:   7
===============================================================================
*A:PE-8#
```

On ASBR PE-4, the routes toward PE-5, PE-6, and PE-7 in the remote AS have been installed in the routing table, as follows:

```
*A:PE-4# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  ISIS      00h49m21s  18
       192.168.24.1                                                 20
192.0.2.2/32                                  Remote  ISIS      00h49m22s  18
       192.168.24.1                                                 10
192.0.2.3/32                                  Remote  ISIS      00h49m22s  18
       192.168.34.1                                                 10
192.0.2.4/32                                  Local   Local     00h49m23s  0
       system                                                       0
192.0.2.5/32                                  Remote  BGP_LABEL 00h42m19s  170
       192.168.48.2                                                 0
192.0.2.6/32                                  Remote  BGP_LABEL 00h41m51s  170
       192.168.48.2                                                 0
192.0.2.7/32                                  Remote  BGP_LABEL 00h41m51s  170
       192.168.48.2                                                 0
192.168.24.0/30                               Local   Local     00h49m23s  0
       int-PE-4-PE-2                                                0
192.168.34.0/30                               Local   Local     00h49m23s  0
       int-PE-4-PE-3                                                0
192.168.48.0/30                               Local   Local     00h49m23s  0
       int-PE-4-PE-8                                                0
-------------------------------------------------------------------------------
No. of Routes: 10
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-4#
```

The BGP labeled routes for the remote PE system prefixes are further advertised toward all the PEs in the AS (through the RR) and are installed in the routing table on all PEs.

At this point, all PEs in one AS have the /32 system IPs of the remote PEs in their routing table, for example for PE-1:

```
*A:PE-1# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Local   Local     00h48m48s  0
```

```
       system                                                      0
192.0.2.2/32                                  Remote  ISIS      00h48m30s  18
       192.168.12.2                                                10
192.0.2.3/32                                  Remote  ISIS      00h48m02s  18
       192.168.13.2                                                10
192.0.2.4/32                                  Remote  ISIS      00h47m48s  18
       192.168.12.2                                                20
192.0.2.5/32                                  Remote  BGP_LABEL 00h40m28s  170
       192.0.2.4 (tunneled)                                        20
192.0.2.6/32                                  Remote  BGP_LABEL 00h39m57s  170
       192.0.2.4 (tunneled)                                        20
192.0.2.7/32                                  Remote  BGP_LABEL 00h39m57s  170
       192.0.2.4 (tunneled)                                        20
192.168.12.0/30                               Local   Local     00h48m48s  0
       int-PE-1-PE-2                                               0
192.168.13.0/30                               Local   Local     00h48m48s  0
       int-PE-1-PE-3                                               0
-------------------------------------------------------------------------------
No. of Routes: 9
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

All PEs in one AS have also received labels for all /32 system IPs of the remote PEs. Therefore, an MP-eBGP session can be created between the RRs in the different ASs to exchange VPN-IPv4 routes.

The configuration for RR PE-3 is as follows. The configuration for RR PE-7 is very similar. The IP addresses can be derived from Figure 209.

```
configure
    router
        bgp
            group "peer-AS-RR"
                family vpn-ipv4
                peer-as 64497
                local-address 192.0.2.3
                neighbor 192.0.2.7
                    multihop 10
                    vpn-apply-export
                    export "EBGP-VPN-IPv4"
                exit
            exit
        exit
```

Policies can be applied on the peering session using the **export** command followed by a policy name, together with the **vpn-apply-export** command necessary to enforce base BGP instance policy on VPN-IPv4 prefixes.

On the RRs, the MP-eBGP session is up, as follows:

```
*A:PE-3# show router bgp neighbor 192.0.2.7
```

```
===============================================================================
BGP Neighbor
===============================================================================
-------------------------------------------------------------------------------
Peer                 : 192.0.2.7
Description          : (Not Specified)
Group                : peer-AS-RR
-------------------------------------------------------------------------------
Peer AS              : 64497            Peer Port            : 179
Peer Address         : 192.0.2.7
Local AS             : 64496            Local Port           : 50423
Local Address        : 192.0.2.3
Peer Type            : External         Dynamic Peer         : No
State                : Established      Last State           : Active
Last Event           : recvKeepAlive
Last Error           : Unrecognized Error
Local Family         : VPN-IPv4
Remote Family        : VPN-IPv4
Hold Time            : 90               Keep Alive           : 30
Min Hold Time        : 0
Active Hold Time     : 90               Active Keep Alive    : 30
--- snipped ---
*A:PE-3#
```

The EBGP session between the two RRs is established.

The VPRNs on PE-1 in AS 64496 and PE-5 in AS 64497 are now interconnected.The routing table for VPRN 1 shows that the remote PE can be reached via a BGP tunnel, as follows:

```
*A:PE-1# show router 1 route-table

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.1.1.1/32                                   Local   Local     00h48m00s  0
       loopback                                                     0
10.2.2.2/32                                   Remote  BGP VPN   00h05m31s  170
       192.0.2.5 (tunneled:BGP)                                     0
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

Packets originating in AS 64496 with a destination in AS 64497 will have 3 labels in AS 64496 (and in AS 64497). Originate a VPRN ping on PE-1 toward the VPRN loopback IP address on PE-5:

```
*A:PE-1# ping router 1 10.2.2.2
PING 10.2.2.2 56 data bytes
64 bytes from 10.2.2.2: icmp_seq=1 ttl=64 time=4.95ms.
64 bytes from 10.2.2.2: icmp_seq=2 ttl=64 time=4.76ms.
64 bytes from 10.2.2.2: icmp_seq=3 ttl=64 time=4.88ms.
64 bytes from 10.2.2.2: icmp_seq=4 ttl=64 time=4.77ms.
64 bytes from 10.2.2.2: icmp_seq=5 ttl=64 time=2.93ms.

---- 10.2.2.2 PING Statistics ----
5 packets transmitted, 5 packets received, 0.00% packet loss
round-trip min = 2.93ms, avg = 4.46ms, max = 4.95ms, stddev = 0.768ms
*A:PE-1#
```

The top label is the LDP label to reach the exit point of the AS (PE-4). This label can be seen with following command on PE-1:

```
*A:PE-1# show router ldp bindings active prefixes prefix 192.0.2.4/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.4/32                                Push              --      262140
192.168.12.2                                1/1/1

192.0.2.4/32                                Swap            262140    262140
192.168.12.2                                1/1/1

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:PE-1#
```

This LDP label will be popped by ASBR PE-4. No LDP label is used between the ASBRs. ASBR PE-8 will push another LDP label.

To reach a PE in the remote AS, a BGP transport label is required, which will be the middle label in the stack. The tunnel table on PE-1 shows a BGP tunnel toward PE-5, as follows:

```
*A:PE-1# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.2/32      ldp       MPLS  65537     9        192.168.12.2   10
192.0.2.3/32      ldp       MPLS  65538     9        192.168.13.2   10
192.0.2.4/32      ldp       MPLS  65539     9        192.168.12.2   20
192.0.2.5/32      bgp       MPLS  262145    12       192.0.2.4      1000
192.0.2.6/32      bgp       MPLS  262146    12       192.0.2.4      1000
192.0.2.7/32      bgp       MPLS  262147    12       192.0.2.4      1000
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-1#
```

The BGP label is assigned by the next hop, in this case by the local ASBR PE-4. This IPv4 label can be seen with following command on PE-1:

```
*A:PE-1# show router bgp routes 192.0.2.5/32 label-ipv4 hunt
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : 192.0.2.5/32
Nexthop        : 192.0.2.4
Path Id        : None
From           : 192.0.2.3
Res. Nexthop   : 192.168.12.2 (LDP)
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : No Community Members
Cluster        : 192.0.2.3
Originator Id  : 192.0.2.4              Peer Router Id : 192.0.2.3
Fwd Class      : None                   Priority       : None
IPv4 Label     : 262138                 Label Type     : SWAP
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : 64497
Route Tag      : 0
Neighbor-AS    : 64497
Orig Validation: NotFound
```

```
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 01h00m25s

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-1#
```

This BGP label will be swapped by ASBR PE-4 in AS 64496 and by ASBR PE-8 in AS 64497.

The bottom label is the VPN label assigned by the remote PE in the remote AS for the destination network. This label is retrieved on PE-1, as follows:

```
*A:PE-1# show router bgp routes 10.2.2.2/32 vpn-ipv4 hunt
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : 10.2.2.2/32
Nexthop        : 192.0.2.5
Route Dist.    : 64497:1                VPN Label      : 262138
Path Id        : None
From           : 192.0.2.3
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : target:64496:1
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.3
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : 64497
Route Tag      : 0
Neighbor-AS    : 64497
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h14m05s
```

```
VPRN Imported  :  1

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-1#
```

Conclusion

Inter-AS option C allows the delivery of Layer 3 VPN services to customers who have sites connected in different ASs. This example shows the configuration of inter-AS option C (specific to this feature) together with the associated show output which can be used for verification and troubleshooting.

Quality of Service

In This Section

This section provides configuration information for the following topics:

- Class Fair Hierarchical Policing for SAPs
- FP and Port Queue Groups
- High Scale QoS IOM: QoS, Service, and Network Configuration
- Pseudowire QoS
- QoS Architecture and Basic Operation

Class Fair Hierarchical Policing for SAPs

This chapter provides information to configure Class Fair Hierarchical Policing for SAPs.

Topics in this chapter include:

- Applicability
- Summary
- Overview
- Configuration
- Conclusion

Applicability

The information in this chapter is applicable to all of the Nokia 7x50 platforms and is focused on the FP2 chipset. The configuration was tested on release 9.0.R1. There are no specific pre-requisites for this configuration.

Summary

The Quality of Service (QoS) features of the 7x50 platforms provide traffic control with both shaping and policing.

Shaping is achieved using a queue; packets are placed on the queue and a scheduler removes packets from the queue at a given rate. This provides an upper bound to the traffic rate sent, thereby protecting down stream devices from bursts. However, shaping can introduce latency and jitter as packets are delayed in the queue. Packets can be dropped when the queue is full or statistically when weighted random early discard is applied. Configuration of shaping on the 7x50 is described in QoS Architecture and Basic Operation.

Policing is another mechanism for controlling traffic rates but it does not introduce latency/jitter. This is achieved using a token bucket mechanism which drops certain packets from the traffic. A common disadvantage of policing implementations is that they are usually applicable to a single level of traffic priority and have no way to fairly share capacity between multiple streams at the same priority level. Nokia's Class Fair Hierarchical Policing (CFHP) addresses these problems by implementing a four level prioritized policing hierarchy which also provides weighted fairness for traffic at a given priority.

Regardless of whether shaping or policing is being used, the preceding QoS classification and subsequent packet marking functionality is similar for both and is covered in more detail in QoS Architecture and Basic Operation.

This note describes the configuration and operation of CFHP when applied to Service Access Points (SAPs). It is also possible to use CFHP for subscribers in a Triple Play Service Delivery Architecture (TPSDA) environment but it is beyond the scope of this note.

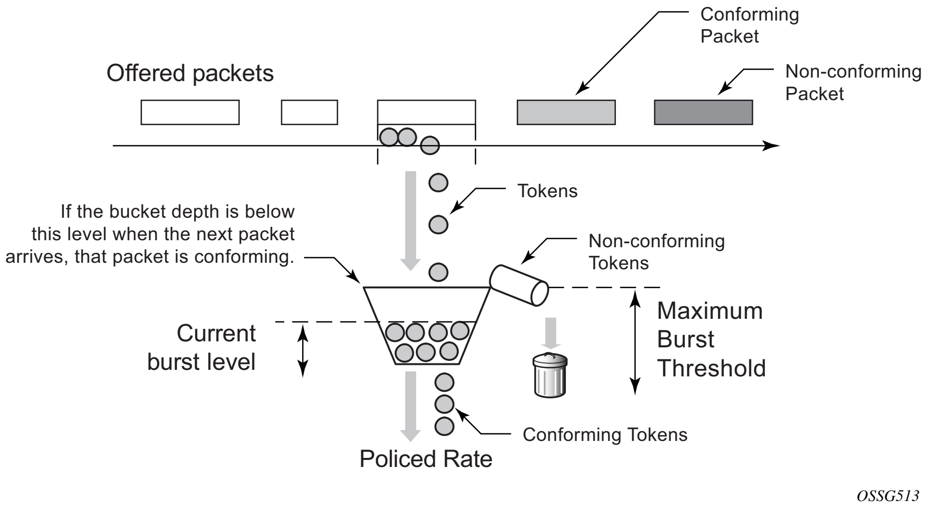
Overview

Policers

CFHP can be used both for ingress and egress QoS. The basic element is a policer which can apply both a committed information rate (CIR) and peak information rate (PIR) to a traffic flow (determined by the ingress classification). Traffic is directed to a policer by assigning a forwarding class (FC) to the policer.

To describe the operation of a policer we will use a token bucket model, this is shown in Figure 210.

Figure 210 **Policer Token Bucket Model**



The policer is modeled by a bucket being filling with tokens which represent the bytes in the packets passing through the policer. The bucket drains at a given rate (the policed rate) and if the token (byte) arrival rate exceeds the drain rate then the bucket will fill. The bucket has a maximum depth, defined by a maximum burst threshold. If tokens for a packet arrive in the bucket when the current burst level of tokens is below the maximum burst threshold then the packet is considered to be conforming and all of its tokens are accepted into the bucket. If a packet's tokens arrive when the current burst level has exceeded the maximum burst threshold then none its tokens are accepted into the bucket and the packet is considered to be non-conforming (in the representation, these tokens over-flow into a waste bin).

Table 18 shows an example of the two possibilities.

Table 18 **Burst Levels**

Maximum burst threshold = 2000 tokens (bytes) Policed rate 2 Mbps = 250000 bytes/sec (250 tokens/ms)				
Arrival Time	**Packet Size**	**Current Burst Level**	**Conforming Packet**	**New Burst Level**
T0	1024	1500	Yes	1500 + 1024 = 2524
T0 + 1ms	128	2524 - 250 = 2274	No	2274

When the first packet arrives the current burst level is below the maximum burst threshold so it is conforming, however, when the second packet arrives the current burst level is above the maximum burst threshold so it is non-conforming.

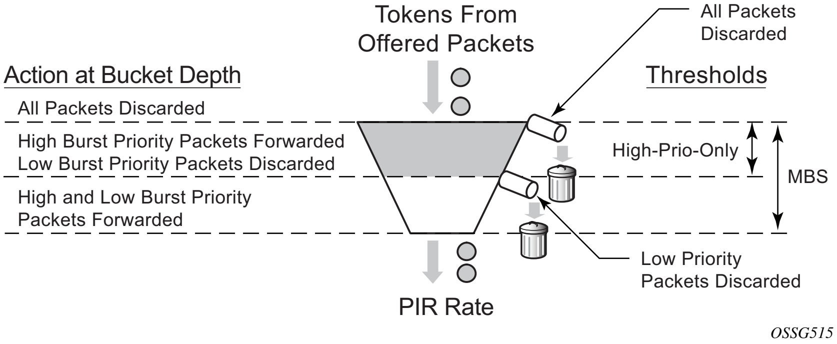
An important aspect of the implementation of hierarchical policing is the ability of a policer bucket to have multiple burst thresholds. The tokens for each arriving packet are only compared against a single threshold relating to the characteristics of packet. These burst thresholds allow specific granular QoS control.

Policer Buckets

A policer uses up to 3 buckets depending on its configuration. A PIR bucket to control the traffic rate which is always used though its rate could be max, there can be an optional CIR bucket if a CIR rate is defined for dynamically profiling (in-profile/out-of-profile) packets, finally there may be a fair information rate (FIR) bucket used to maintain traffic fairness in a hierarchical policing scenario when multiple child policers are configured at the same parent priority level.

The PIR bucket is drained at the PIR rate and has two burst thresholds, one for high burst priority traffic (defined by the maximum burst size (MBS)) and a second for low burst priority traffic (defined by the MBS minus high-prio-only), see Figure 211. The traffic burst priority is determined at ingress by the configured priority of either high or low, and at the egress by the profile state of the packets (in-profile=high, out-of-profile=low). Note that by default all FCs are low burst priority. If a packet conforms at the PIR bucket (its tokens enter the bucket) then the packet is forwarded, otherwise the packet is discarded. Discarding logically results in the packet's tokens not being placed into the CIR, FIR or parent policer buckets.

Figure 211 **Peak Information Rate (PIR) Bucket**

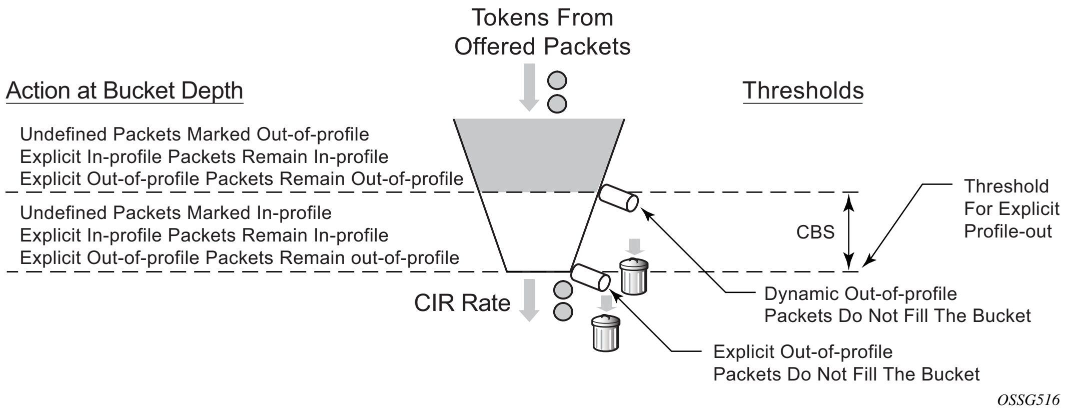


The CIR bucket is drained at the CIR rate and has one configurable burst threshold (defined by the committed burst size (CBS)). At the ingress, if the bucket level is below this threshold traffic is determined to be in-profile so the only action of the CIR bucket is to set the state of dynamically profiled packets to be either in-profile or out-of-profile. At the egress, re-profiling only affects Dot1P and DEI (Layer 2) egress marking (if the frame is double tagged, only the outer VLAN tag is remarked).

The CBS threshold is used when operating in color-blind mode, the profile of incoming packets is undefined and dynamically set based on the current burst level in the CIR bucket compared to the CBS threshold. It is also possible to operate (simultaneously) in color-aware mode, where the classification of incoming packets is used to explicitly determine whether a packet is in-profile or out-of-profile. For color-aware mode, the CIR bucket does not change the packet profile state.

In order to ensure that the overall amount of in-profile traffic takes into account both the explicit and dynamic in-profile packets, tokens from the explicit in-profile packets are allowed to fill the bucket above the CBS threshold. By doing this, dynamically profiled packets are only marked as in-profile after the token level representing dynamically in-profile and explicit in-profile packets have fallen below the CBS threshold (as the bucket drains). Note that explicitly marked out-of-profile packets remain out-of-profile, so the bottom of the bucket can be considered to be an implicit burst threshold for these packets. This is shown in Figure 212.

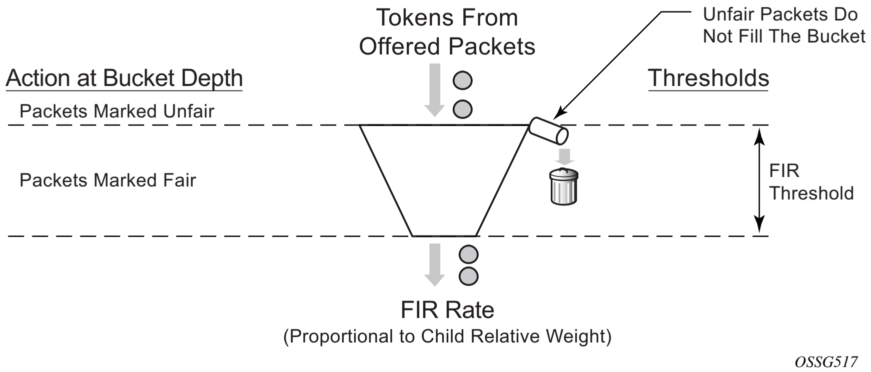
***Figure 212* Committed Information Rate (CIR) Bucket**


Tokens From
Offered Packets
Action at Bucket Depth
Thresholds
Undefined Packets Marked Out-of-profile
Explicit In-profile Packets Remain In-profile
Explicit Out-of-profile Packets Remain Out-of-profile
Undefined Packets Marked In-profile
Explicit In-profile Packets Remain In-profile
Explicit Out-of-profile Packets Remain out-of-profile
CBS
Threshold
For Explicit
Profile-out
CIR Rate
Dynamic Out-of-profile
Packets Do Not Fill The Bucket
Explicit Out-of-profile
Packets Do Not Fill The Bucket
OSSG516

As the depths of the PIR and CIR buckets (MBS and CBS, respectively) are configured independently it is possible to have, for example, the CBS to be larger than the MBS (which is not possible for a queue). This could result in traffic being discarded because it is non-conforming at the PIR bucket but would have been conforming at the CIR bucket. Conversely, if the CBS is smaller than the MBS and the PIR=CIR traffic can be forwarded as out-of-profile, which would not be the case with a queue.

The FIR bucket is controlled by the system and is only used in hierarchical policing scenarios to determine a child’s fair access to the available capacity at a parent priority level relative to other children at the same level. This bucket is only used when there is more than one child policer assigned to a given parent policer priority level. The drain rate of the FIR bucket is dynamically set proportionally to the weight configured for the child. This is shown in Figure 213.

Figure 213 **Fair Information Rate (FIR) Bucket**



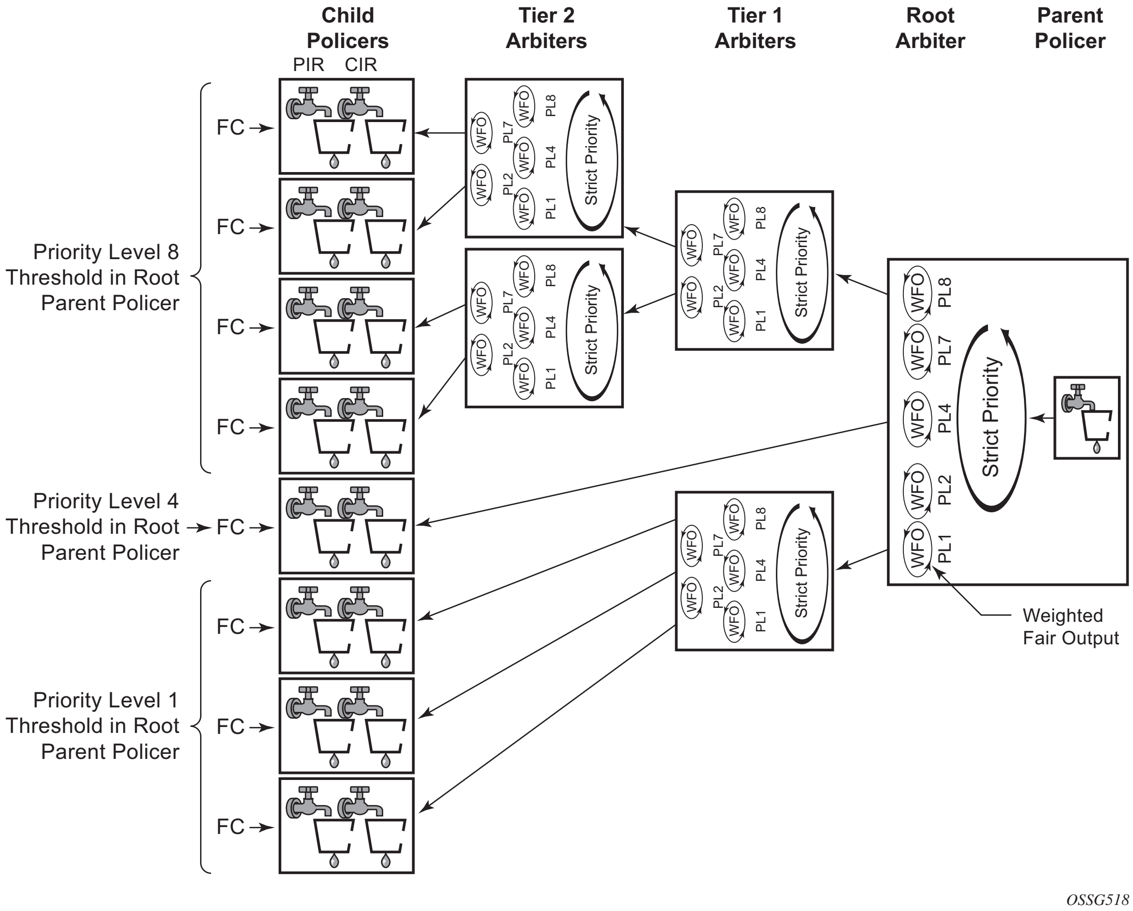
Hierarchical Policing

Policers can be used standalone or with a parent policer to provide hierarchical policing. Up to four stages can be configured in the hierarchy: the child policer, tier 1 and 2 intermediate arbiters, and a root arbiter (which is associated with the parent policer). The arbiters are logical entities that distribute bandwidth at a particular tier to their children in a priority level order, see Figure 214.

This may result in the drain rates for the child policer buckets being modified, so each child policer PIR and CIR bucket has an administrative rate value (what it is configured to) and an operational rate value (the current operating rate) based on the bandwidth distribution by the parent arbiters.

Each stage in the hierarchy connects to its parent at a priority level and a weight. There are eight available priorities which are serviced in a strict order (8 to 1, highest to lowest, respectively). The weight is used to define relative fairness when multiple children are configured in the same priority level. Note that the child access to parent policer burst capacity is governed by the level at which the child ultimately connects into the root arbiter, not by its connection level at any intermediate arbiters.

Figure 214 **Policer and Arbiter Hierarchy**


Child Policers
PIR
CIR
Tier 2 Arbiters
Tier 1 Arbiters
Root Arbiter
Parent Policer
Priority Level 8 Threshold in Root Parent Policer
Priority Level 4 Threshold in Root Parent Policer
Priority Level 1 Threshold in Root Parent Policer
FC
WFO
PL1
PL2
PL4
PL7
PL8
Strict Priority
Weighted Fair Output
OSSG518

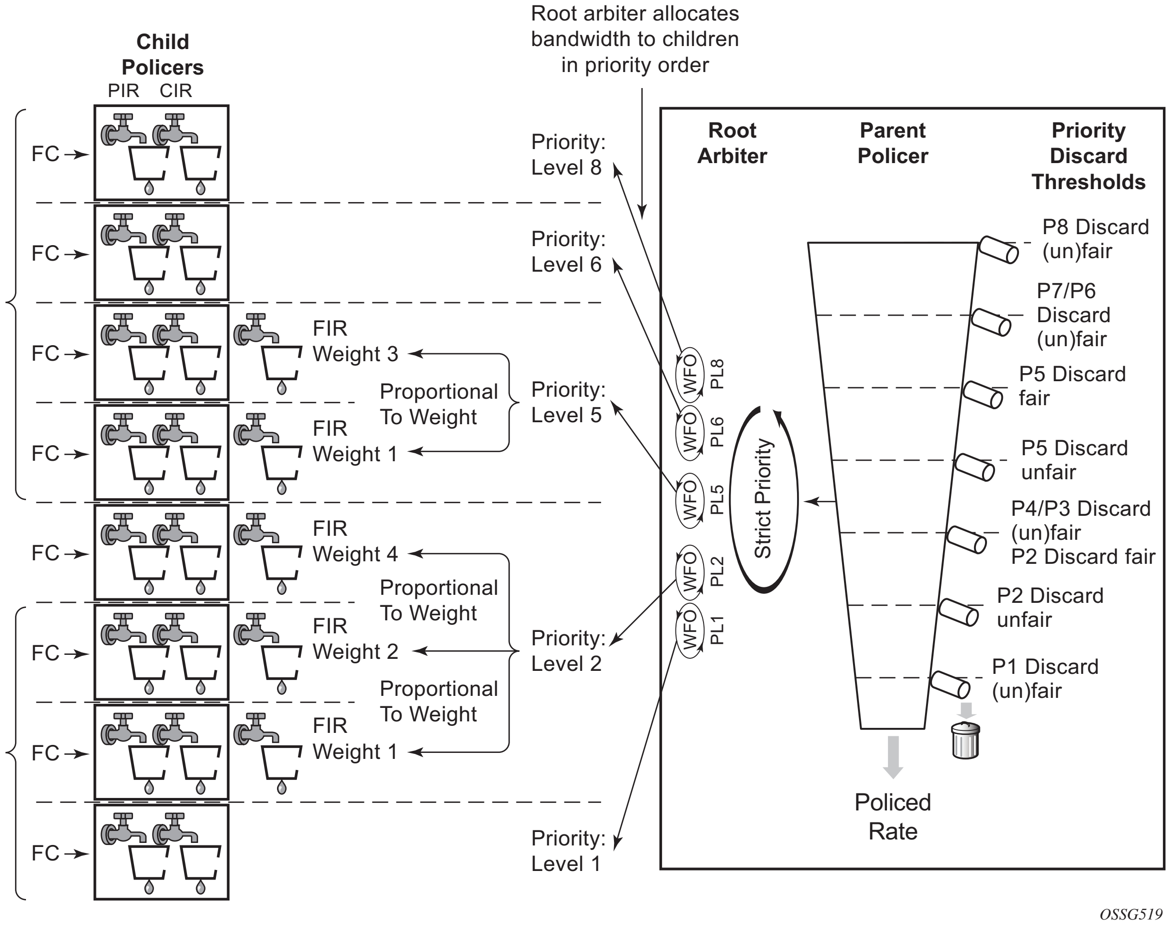
The final configuration aspect to consider is the parent policer, specifically its multiple thresholds and how they relate to the child policers. See Figure 215.

There are 8 priority levels at the parent policer, each having an associated discard-fair and discard-unfair threshold.

The discard-fair threshold is the upper burst limit for all tokens (consequently, all packets) at the given priority, all traffic at a given priority level is discarded when its tokens arrive with this threshold being exceeded. The discard-fair thresholds enable prioritization at the parent policer by having the burst capacity for each priority threshold be larger (or equal) to those of lower priorities. For example, referring to Figure 215, the priority 6 (P6) discard-fair threshold is larger than the priority 5 (P5) discard-fair threshold with the result that even if the priority 5 and below traffic is overloading the parent policer, the priority 6 traffic has burst capacity available in order to allow some of its packets to conform and get forwarded through the parent policer.

Note that if a packet is discarded at the parent policer, the discard needs to be reflected in the associated child policer, this is achieved by also logically removing the related tokens from the child policer buckets.

Figure 215 **Parent Policer and Root Arbiter**



Each priority also has a discard-unfair threshold which discards only unfair traffic of that priority, remembering that fair and unfair are determined by the FIR bucket based on the relative weights of the children.

By default, if there are no children configured at a given priority level then both its discard-fair and discard-unfair thresholds are set to zero bytes above the previous priority's discard-fair threshold.

If there is only a single child at a priority level, the discard-fair will be greater than the previous priority’s discard-fair value (by an amount equal to the maximum of the min-thresh-separation and the mbs-contribution, see below) but the discard-unfair will be the same as the previous priority’s discard-fair threshold (there is no need for a fairness function when there is only a single child at that priority).

If there is more than one child at a priority level, the discard-unfair threshold will be greater than the previous priority’s discard-fair threshold by min-thresh-separation (see below) and the discard-fair threshold will be adjusted upwards by an amount equal to mbs-contribution minus min-thresh-separation.

The result can be summarized as follows:

- With no children at a priority level, the discard fair and unfair thresholds match the values of the previous priority.
- If there are at least two children at a priority level, the discard-unfair burst capacity equals min-thresh-separation.
- The burst capacity for a given priority level with at least one child equals the mbs-contribution, unless this is less than min-thresh-separation in which case the min-thresh-separation is used.

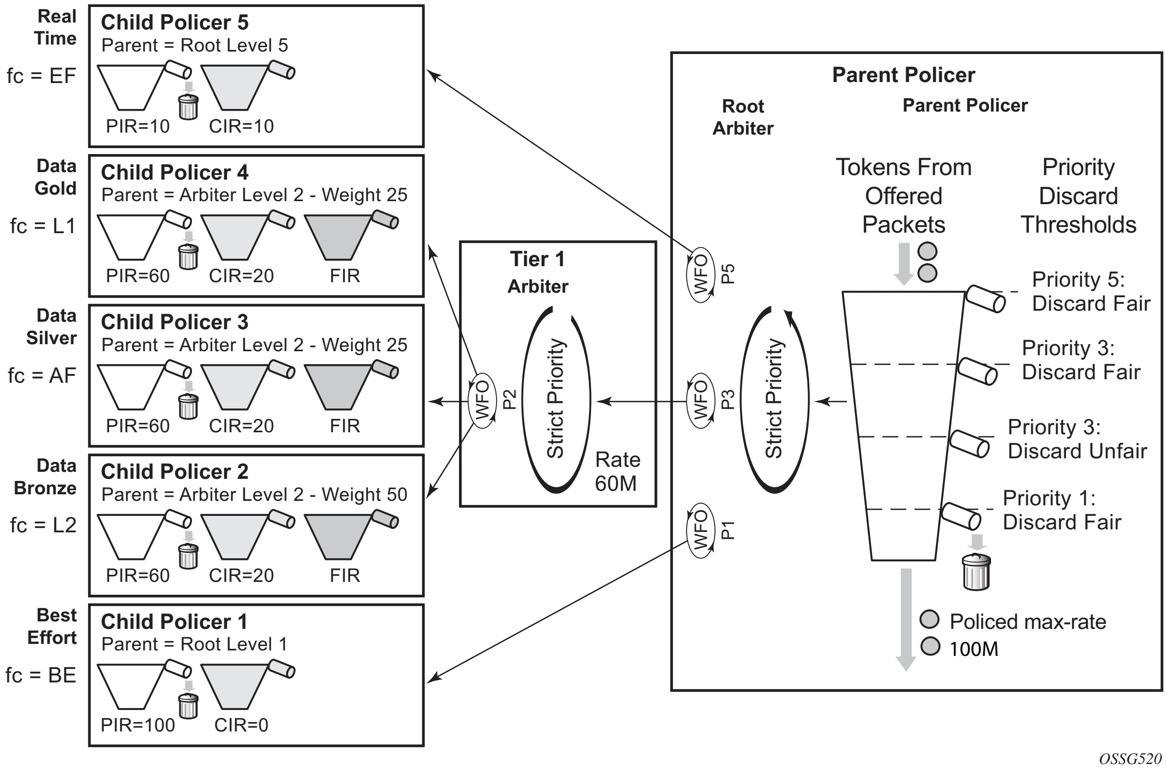
The burst tolerance for each threshold is its own burst capacity plus the sum of the burst capacities of all lower thresholds. Referring to Figure 215, the total burst capacity for priority 6 is the sum of the burst capacities for priorities 1 to 6. Note that the burst for a given FC is normally controlled by the burst allowed at the child PIR threshold, not by the parent policer.

As the burst capacity at the parent policer for a given priority level can change when adding or removing children at lower priority levels, a parameter (fixed) is available per priority threshold which causes the discard-fair and discard-unfair thresholds to be non-zero and so greater than the previous priority’s thresholds, calculated as above, even when there are no children at that priority level. An exception to this is when the mbs-contribution is set to zero with the fixed parameter configured, in which case both the discard unfair and fair for that priority level are set to zero bytes above the previous level’s thresholds (which results in the corresponding traffic being dropped).

A specific configuration and associated show output is included below to highlight the different threshold options described above.

The QoS example shown in Figure 216 is used to describe the configuration of CFHP.

Figure 216 **Configuration Example**



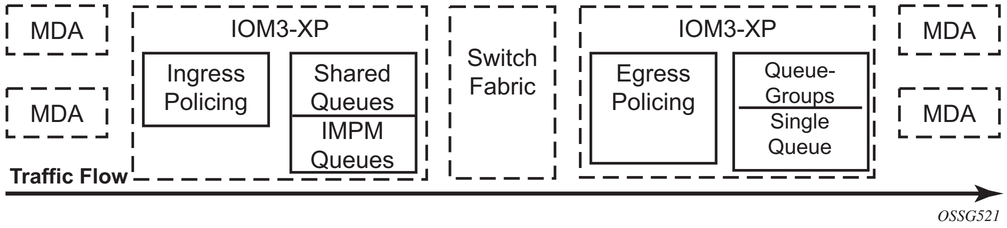
Five classes of services are accepted, each with a specific CIR and PIR. The data classes, bronze, silver and gold (L2/AF/L1), have a relative weighting of 50/25/25 at priority Level 2 of an intermediate arbiter which is constrained to 60Mbps. At the parent policer, the real time traffic (EF) is defined at level 5, with the data classes at Level 3 and a best effort class (BE) at Level 1. The overall traffic is constrained to 100Mbps at the parent policer. Only unicast traffic is policed in this example.

This example focuses on ingress policing, however, the configuration of policers, arbiters and the parent policer at the egress is almost identical to that at the ingress, the only difference being the particular statistics that can be collected.

There is a difference between ingress and egress policing in terms of how the ingress traffic accesses the switch fabric and the egress traffic access the port after it has been policed. In both cases, unicast access is enabled through a set of policer-output-queues, which are shared-queues at the ingress and queue-groups at the egress (at the egress, user defined queue-groups can be used). It is also possible to use a single service queue to access the egress port. Ingress multipoint traffic accesses the switch fabric using the Ingress Multicast Path Management (IMPM) queues.

This is shown in Figure 217 on an IOM3-XP (other line cards have the same logic).

Figure 217 **Post Policing Queues**



The differences between the ingress and egress policing configuration will be highlighted in the associated sections.

Configuration

To achieve the QoS shown in Figure 216, configure a SAP-ingress QoS policy to define the child policers and a policer-control-policy to define the intermediate arbiter and the root arbiter/parent policer. As this example is for ingress, the unicast traffic will pass through a set of shared queues called policer-output-queues, which could be modified if required.

Policers

Policers control the CIR and PIR rates for each of the traffic classes and are defined in a SAP-ingress QoS policy. The focus here are parameters related to policing.

The configuration of a child (or standalone) policer is similar to that of a queue.

```
config>qos>sap-ingress# policer policer-id [create]
    description "description-string"
    adaptation-rule [pir {max | min | closest}] [cir {max | min | closest}]
    stat-mode {no-stats|minimal|offered-profile-no-cir|
           offered-priority-no-cir|offered-profile-cir|offered-priority-cir|
           offered-total-cir|offered-limited-profile-cir}
    rate {max | kilobits-per-second} [cir {max | kilobits-per-second}]
    percent-rate pir-percent [cir cir-percent]
    mbs size [bytes | kilobytes]
    cbs size [bytes | kilobytes]
    high-prio-only [default | percent-of-mbs]
    parent {root | arbiter-name} [level level] [weight weight-within-level]
    packet-byte-offset {add bytes | subtract bytes}
```

Parameters:

- description — This configures a text string, up to 80 characters, which can be used to describe the use of the policy.
- adaptation-rule — The hardware supports distinct values for the rates. This parameter tells the system how the rate configured should be mapped onto the possible hardware values. min results in the next higher hardware value being used, max results in the next lower hardware value being used and closest results in the closest available hardware value being used. As can be seen, it is possible to set the adaptation-rule independently for the CIR and PIR.
Default: closest
- stat-mode — This defines the traffic statistics collected by the policer, summarized in Table 19.

Table 19 **Policer stat-mode**

stat-mode	Ingress		Egress	
no-stats	0	Neither policer nor parent arbiter account are required.	0	Neither policer nor parent arbiter accounting are required.
minimal (default)	1	Basic policer accounting (default).	1	Basic policer accounting (default).
offered-profile-no-cir	2	All ingress packets are either in-profile or out-of-profile.	2	Accounting for egress offered profile is required. No visibility for CIR profile state output.
offered-priority-no-cir	2	Ingress packet burst priority accounting is the primary requirement.		N/A
offered-limited-profile-cir	3	Ingress color-aware profiling is in use but packets are not being classified as in-profile.		N/A
offered-profile-cir	4	Ingress color-aware profiling is in use and packets are undefined or classified as out-of-profile or in-profile.	4	Egress profile reclassification is performed.
offered-priority-cir	4	Ingress policer is used in color-blind mode and ingress packet priority and CIR state output accounting is needed.		N/A
offered-total-cir	2	Ingress priority and ingress profile accounting is not needed. CIR profiling is in use.	2	Offered profile visibility is not required (such as, all offered packets have the same profile) and CIR profiling is in use.
	Counter resources needed for this stat mode			

- rate and cir — The rate defines the PIR and the cir defines the CIR, both are in Kbps. The parameters rate and percent-rate are mutually exclusive and will overwrite each other when configured in the same policy.
  Range: PIR=1 to 20,000,000 Kbps or max ; CIR=0 to 20,000,000 Kbps or max
  Default: rate(PIR)=max ; cir=0
- percent-rate and cir — The percent-rate defines the PIR and the cir defines the CIR with their values being a percentage of the maximum policer rate of 20Gbps. The parameters rate and percent-rate are mutually exclusive and will overwrite each other when configured in the same policy.
  Range: pir-percent = [0.01..100.00]; cir-percent = [0.00..100.00]
  Default: pir-percent = 100; cir-percent = 0.00
- mbs and cbs — The mbs defines the MBS for the PIR bucket and the cbs defines the CBS for the CIR bucket, both can be configured in bytes or kilobytes.
  Note that the PIR MBS applies to high burst priority packets (these are packets whose classification match criteria is configured with priority high at the ingress and are in-profile packets at the egress).
  Range: mbs=0 to 4194304 bytes; cbs=0 to 4194304 bytes
  Note: mbs=0 prevents any traffic from being forwarded.
  Default: mbs=10ms of traffic or 64KB if PIR=max; cbs=10ms of traffic or 64KB if CIR=max
- high-prio-only — This defines a second burst threshold within the PIR bucket to give a maximum burst size for low burst priority packets (these are packets whose classification match criteria is configured with priority low at the ingress and are out-of-profile packets at the egress). It is configured as a percentage of the MBS.
  Default: 10%
- parent — This parameter is used when hierarchical policing is being performed and points to the parent arbiter (which could be the root arbiter or an intermediate arbiter), giving the level to which this policer connects to its parent arbiter and its relative weight compared to other children at the same level. Note that for a child policer to be associated with a parent, its stat-mode cannot be no-stats.
  Range: level=1 to 8; weight=1 to 100
  Default: level=1; weight=1
- packet-byte-offset — This changes the packet size used for accounting purposes, both in terms of the CIR and PIR rates and what is reported in the statistics. The change can either add or subtract a number of bytes. For example:
  - To have the policer work on Layer 2 frame size including inter-frame gap and preamble, add 20 bytes.
  - To have the policer work on IP packet size instead of the default layer 2 frame size, subtract the encapsulation overhead:

    14 bytes L2 + 4bytes VLAN ID + 4 bytes FCS = 22 bytes

Range: add-bytes=0 to 31; sub-bytes=1 to 32
Default: add-bytes=0; sub-bytes=0

A FC must be assigned to the policer in order for the policer to be instantiated (allocating a hardware policer).

By default, any unicast traffic assigned to the FC at the ingress will be processed by the policer, non-unicast traffic would continue to use the multipoint queue. At the egress all traffic assigned to the FC is processed by the policer (as there is no distinction between unicast and non-unicast traffic at the egress).

If required, non-unicast traffic can be policed in IES/VPRN and VPLS services at the ingress (note: all Epipe traffic is treated as unicast). Within an IES/VPRN service, multicast traffic can be assigned to a specific ingress policer on a PIM enabled IP interface. When the service is VPLS, broadcast, unknown unicast and multicast traffic can be individually assigned to ingress policers. In each of these cases, the policers used could be separate from the unicast policer, resulting in the instantiation of additional hardware policers, or a single policer could be used for multiple traffic types (this differs from the queuing implementation where separate queue types are used for unicast and non-unicast traffic).

```
config>qos>sap-ingress>fc#
    broadcast-policer <policer-id>
    unknown-policer <policer-id>
    multicast-policer <policer-id>
```

As mentioned above, the ingress policed unicast traffic passes through a set of shared-queues (policer-output-queues) to access the switch fabric with the multipoint traffic using the IMPM queues.

When policers are required at the egress, a SAP-egress policy is used. The configuration of the policers is almost identical to that used in the SAP-ingress policy, the only difference being the available stat-modes (as shown above).

At the egress, the policed traffic can also be directed to a specific queue-group (instead of the default policer-output-queues) and to a specific queue within that queue-group, as follows:

```
config>qos>sap-egress>fc# policer <policer-id> [group <queue-group-name> [queue
<queueid>]]
```

It is also possible to direct the egress policed traffic to a single service queue if specific egress queuing is required, as follows:

```
config>qos>sap-egress>fc# policer <policer-id> queue <queue-id>
```

Multiple egress policers in a SAP-egress policy can use the same local queue and other forwarding classes can directly use the same local queue that is being used by policers.

Parent Policer and Arbiters

The parent policer and its associated root arbiter, together with the tier 1 and 2 arbiters, are configured within a policer-control-policy.

```
config>qos# policer-control-policy policy-name [create]
    description description-string
    root
        max-rate {kilobits-per-second | max}
        priority-mbs-thresholds
            min-thresh-separation size [bytes|kilobytes]
            priority level
                mbs-contribution size [bytes|kilobytes] [fixed]
    tier 1
        arbiter arbiter-name [create]
            description escription-string
            rate {kilobits-per-second|max}
            parent root [level priority-level] [weight weight-within-level]
    tier 2
        arbiter arbiter-name [create]
            description description-string
            rate {kilobits-per-second | max}
            parent {root|arbiter-name} [level priority-level] [weight weight-within-
level]
```

Parameters:

- description — This configures a text string, up to 80 characters, which can be used to describe the use of the policy.
- root — This section defines the configuration of the parent policer and the root arbiter.
  - max-rate — This defines the policed rate of the parent policer, the rate at which the bucket is drained. It is defined in Kbps with an option to use max, in which case the maximum possible rate is used.
    Range: 1 to 20,000,000Kbps or max
    Default: max
  - priority-mbs-thresholds
    This section defines the thresholds used for the 8 priorities available in the parent policer.

- min-thresh-separation — This defines the minimum separation between any two active thresholds in the parent policer in units of bytes or kilobytes.

  It should be set to a value greater than the maximum packet size used for traffic passing through the policer. This ensures that a single packet arriving in the parent policer will not cause the depth of tokens to cross two burst thresholds, if this did happen it would result in the prioritization failing as a given priority level could be starved of burst capacity by a lower priority traffic.

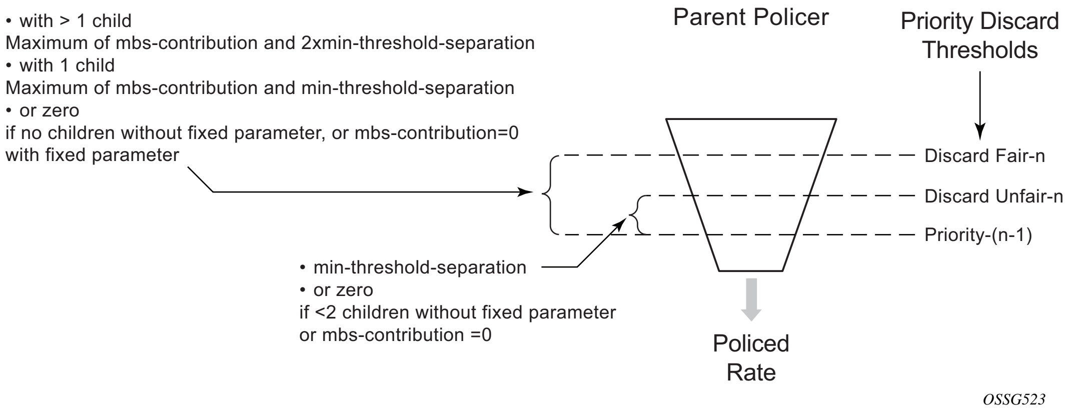
  This parameter is also used as the burst capacity for each priority level's unfair packets.
  Range: 0 to 4194304 bytes
  Default: 1536 bytes

- mbs-contribution — This is normally used to define the amount of packet burst capacity required at the parent policer for a particular priority level with at least one child, keeping in mind that the total capacity is the sum of this plus that of all lower thresholds. The actual burst capacity used depends also on the setting of min-thresh-separation, as described earlier. This permits the tuning of the burst capacity at the parent for any children at a given priority level. A conservative setting would ensure that the burst at the parent policer for a given priority is the sum of the bursts of all children at that priority. Less conservative settings could use a lower value and assume some level of oversubscription.

  The use of the fixed parameter causes both the fair and unfair discard thresholds to be non-zero even when there are no children assigned to this priority level (unless the mbs-contribution is set to zero).
  Range: 0 to 4194304
  Default: 8192 bytes

  The relationship between these two parameters is shown in Figure 218.

Figure 218 **Parent Policer Thresholds**


• with > 1 child
Maximum of mbs-contribution and 2xmin-threshold-separation
• with 1 child
Maximum of mbs-contribution and min-threshold-separation
• or zero
if no children without fixed parameter, or mbs-contribution=0
with fixed parameter
Parent Policer
Priority Discard
Thresholds
Discard Fair-n
Discard Unfair-n
Priority-(n-1)
• min-threshold-separation
• or zero
if <2 children without fixed parameter
or mbs-contribution =0
Policed
Rate
OSSG523

- tier

  This section defines the configuration of any intermediate tier 1 or 2 arbiters.

  – arbiter

    This specifies the name of the arbiter.

    - description — This configures a text string, up to 80 characters, which can be used to describe the use of the policy.
    - rate — This defines the rate of the arbiter, it is the maximum rate at which the arbiter will distribute burst capacity to its children. It is defined in Kbps with an option to use max, in which case the maximum possible rate is used.
      Range: 1 to 20,000,000Kbps or max
      Default: max
    - parent — This parameter is used when hierarchical policing is being performed and points to the parent arbiter (which could be the root arbiter or a tier 1 arbiter), giving the level to which this arbiter connects to its parent arbiter and its relative weight compared to other children at the same level.
      Range: level=1 to 8; weight=1 to 100
      Default: level=1; weight=1

Access to Switch Fabric and Egress Port

After the traffic has been processed by the policers it must pass through a set of queues in order to access the switch fabric at the ingress or the port at the egress.

For the ingress unicast traffic, there is a set of shared-queues (one queue per FC for each possible switch fabric destination) called policer output queues. Note that only their queue characteristics can be configured, the FC to queue mapping is fixed. Also, the PIR/CIR rates only affect the packet scheduling, they do not alter the packet profile state. The details of shared-queues are beyond the scope of this note.

```
config>qos# shared-queue "policer-output-queues"
    description description-string
    fc <fc-name> [create]
        broadcast-queue <queue-id>
        multicast-queue <queue-id>
        queue <queue-id>
        unknown-queue <queue-id>
    queue queue-id [queue-type] [multipoint] [create]
        cbs percent
        mbs percent
        high-prio-only percent
        pool pool-name
        rate percent [cir percent]
```

Multipoint traffic uses the IMPM queues to access the switch fabric. For the egress to the port, either a queue-group or a single service queue is used. There is a default queue-group called policer-output-queues or a user configured queue-group can also be used.

As mentioned above, when a policer is assigned to a specific queue-group (default or user defined) it is optionally possible to configure explicitly the queue to be used. Within the queue-group it is also possible to redirect a FC for policed traffic to a specific queue, using the FC parameter. The preference of the FC to queue mapping is (in order, highest to lowest):

1. Explicitly configured in SAP-egress FC definition
2. Mapped using FC parameter within queue-group definition
3. Default is to use queue 1

```
config>qos>qgrps>egr# queue-group queue-group-name [create]
    description description-string
    queue queue-id [queue-type] [create]
        adaptation-rule [pir adaptation-rule] [cir adaptation-rule]
        burst-limit size [bytes|kilobytes]
        cbs size-in-kbytes
        high-prio-only percent
        mbs size [bytes|kilobytes]
        parent scheduler-name [weight weight] [level level] [cir-weight cir-weight]
                              [cir-level cir-level]
        percent-rate pir-percent [cir cir-percent]
        pool pool-name
        port-parent [weight weight] [level level] [cir-weight cir-weight]
                    [cir-level cir-level]
        rate pir-rate [cir cir-rate]
        xp-specific
```

```
            wred-queue [policy slope-policy-name]
    fc fc-name [create]
        queue queue-id
```

The default policer-output-queues queue-group consists of two queues; queue 1 being best-effort and queue 2 being expedite. The lowest four FCs (BE, L2, AF, L1) are assigned to queue 1 and the highest four queues (H2, EF, H1, NC) are assigned to queue 2. It may be important to change the queue 2 definition in the queue-group to have CIR=PIR when there are other best-effort queues using a non-zero CIR on the same egress port. This ensures that the policed traffic using queue 2 will be scheduled before any other best-effort within CIR traffic. It also results in the queue CBS being non-zero, allowing the queue 2 traffic access to reserved buffer space.

```
A:PE-1>config>qos# queue-group-templates egress queue-group "policer-output-queues"
A:PE-1>cfg>qos>qgrps>egr>qgrp# info detail
----------------------------------------------
                    description "Default egress policer output queues."
                    queue 1 best-effort create
                        no parent
                        no port-parent
                        adaptation-rule pir closest cir closest
                        rate max cir 0
                        cbs default
                        mbs default
                        high-prio-only default
                        no pool
                        xp-specific
                            no wred-queue
                        exit
                        no burst-limit
                    exit
                    queue 2 expedite create
                        no parent
                        no port-parent
                        adaptation-rule pir closest cir closest
                        rate max cir 0
                        cbs default
                        mbs default
                        high-prio-only default
                        no pool
                        xp-specific
                            no wred-queue
                        exit
                        no burst-limit
                    exit
                    fc af create
                        queue 1
                    exit
                    fc be create
                        queue 1
                    exit
                    fc ef create
                        queue 2
                    exit
```

```
fc h1 create
    queue 2
exit
fc h2 create
    queue 2
exit
fc l1 create
    queue 1
exit
fc l2 create
    queue 1
exit
fc nc create
    queue 2
exit
```

The remaining details of queue-groups are beyond the scope of this section.

Applying the SAP Ingress and Policer Control Policy

The SAP ingress policy and policer control policy are both applied under the associated SAP. After applying these, it is possible to override the configuration of specific policers and/or the policer control policy. This is shown below. The parameter values are the same as detailed for the policies, as above.

```
config>service><service>#
    sap sap-id [create]
        [ingress|egress]
            qos policy-id
            policer-control-policy policy-name
            policer-override
                policer policer-id [create]
cbs size [bytes|kilobytes]
     mbs size [bytes|kilobytes]
packet-byte-offset {add add-bytes | subtract sub-bytes}
rate {rate | max} [cir {max | rate}
percent-rate <pir-percent> [cir <cir-percent>]
stat-mode stat-mode
            policer-control-override [create]
                max-rate {rate | max}
                priority-mbs-thresholds
min-thresh-separation size [bytes | kilobytes]
priority level
                        mbs-contribution size [bytes | kilobytes]
```

The SAP ingress policy and policer control policy required for the configuration example in Figure 216 is shown below.

```
#--------------------------------------------------
echo "QoS Policy Configuration"
```

```
#--------------------------------------------------
    qos
        policer-control-policy "cfhp-1" create
            root
                max-rate 100000
            exit
            tier 1
                arbiter "a3" create
                    parent "root" level 3
                    rate 60000
                exit
            exit
        exit
        sap-ingress 10 create
            queue 1 create
            exit
            queue 11 multipoint create
            exit
            policer 1 create
                stat-mode offered-total-cir
                parent "root"
                rate 100000
                high-prio-only 0
            exit
            policer 2 create
                stat-mode offered-total-cir
                parent "a3" level 2 weight 50
                rate 60000 cir 20000
                high-prio-only 0
            exit
            policer 3 create
                stat-mode offered-total-cir
                parent "a3" level 2 weight 25
                rate 60000 cir 20000
                high-prio-only 0
            exit
            policer 4 create
                stat-mode offered-total-cir
                parent "a3" level 2 weight 25
                rate 60000 cir 20000
                high-prio-only 0
            exit
            policer 5 create
                stat-mode offered-total-cir
                parent "root" level 5
                rate 10000 cir 10000
                high-prio-only 0
            exit
            fc "af" create
                policer 3
            exit
            fc "be" create
                policer 1
            exit
            fc "ef" create
                policer 5
            exit
            fc "l1" create
                policer 4
            exit
            fc "l2" create
                policer 2
            exit
            dot1p 1 fc "be"
            dot1p 2 fc "l2"
            dot1p 3 fc "af"
            dot1p 4 fc "l1"
            dot1p 5 fc "ef"
        exit
```

Traffic is classified based on dot1p values, each of which is assigned to an individual FC which in turn is assigned to a policer. The policer rates are configured as required for the example with an appropriate stat-mode. Default values are used for the policer burst thresholds. As all FCs are low burst priority by default, the high-prio-only has been set to zero in order to allow the traffic to use all of the MBS available at the PIR bucket.

Policers 2, 3 and 4 are parented to the arbiter “a3” with the required weights and at a single level (Level 2). In this example it does not matter which level of “a3” is used to parent these policers, the important aspect is the level at which “a3” is parented to the root. Consequently, these policers use the Level 3 parent policer thresholds (not the level they are parented on a“a3” not Level 2). Arbiter “a3” has a rate of 60Mbps so that its children cannot exceed this rate (except up to the burst tolerances).

Policers 1 and 5 are directly parented to the root arbiter, together with tier 1 arbiter “a3”.

The total capacity for the 5 traffic streams is constrained to 100Mbps by the parent policer, again with the default burst tolerances at the root arbiter.

The SAP-ingress and policer-control-policies are applied to a SAP within an Epipe.

```
#--------------------------------------------------
echo "Service Configuration"
#--------------------------------------------------
    service
        epipe 1 customer 1 create
            sap 1/1/3:1 create
                ingress
                    policer-control-policy "cfhp-1"
                    qos 10
                exit
            exit
            sap 1/1/4:1 create
            exit
            no shutdown
        exit
    exit
```

The following configuration is used to highlight the relative thresholds in the parent policer when a priority level has 0, 1 or 2 associated children, both with and without using the fixed parameter.

```
--------------------------------------------------
echo "QoS Policy Configuration"
#--------------------------------------------------
    qos
       policer-control-policy "cfhp-2" create
            root
                max-rate 100000
                priority-mbs-thresholds
                    min-thresh-separation 256 bytes
                    priority 1
                        mbs-contribution 1 kilobytes
                    exit
                    priority 2
                        mbs-contribution 1 kilobytes
                    exit
                    priority 3
                        mbs-contribution 1 kilobytes
                    exit
                    priority 4
                        mbs-contribution 1 kilobytes fixed
                    exit
                    priority 5
                        mbs-contribution 1 kilobytes fixed
                    exit
                    priority 6
                        mbs-contribution 1 kilobytes fixed
                    exit
                exit
            exit
        exit
        sap-ingress 20 create
            queue 1 create
            exit
            queue 11 multipoint create
            exit
            policer 1 create
                parent "root" level 2
            exit
            policer 2 create
                parent "root" level 3
            exit
            policer 3 create
                parent "root" level 3
            exit
            policer 4 create
                parent "root" level 5
            exit
            policer 5 create
                parent "root" level 6
            exit
            policer 6 create
                parent "root" level 6
            exit
            fc "af" create
```

```
                policer 3
            exit
            fc "be" create
                policer 1
            exit
            fc "ef" create
                policer 6
            exit
            fc "h2" create
                policer 5
            exit
            fc "l1" create
                policer 4
            exit
            fc "l2" create
                policer 2
            exit
        exit
#--------------------------------------------------
echo "Service Configuration"
#--------------------------------------------------
    service
        epipe 2 customer 1 create
            sap 1/1/3:2 create
                ingress
                    policer-control-policy "cfhp-2"
                    qos 20
                exit
            exit
            sap 1/1/4:2 create
            exit
            no shutdown
        exit
```

A policer-control-policy can also be applied under a multi-service site (MSS) so that the hierarchical policing applies to traffic on multiple SAPs, potentially from different services. The MSS can only be assigned to a port, which could be a LAG, but it is not possible to assign an MSS to a card. When MSS are used, policer overrides are not supported.

```
config>service><service>#
    service
        customer customer-id [create]
            multi-service-site customer-site-name [create]
                assignment port port-id
                egress
                    policer-control-policy name
                ingress
                    policer-control-policy name
        service-type
            sap sap-id
                multi-service-site customer-site-name
                ingress
                    qos policy-id
                egress
                    qos policy-id
```

Show Output

After configuring the example as described in the previous section, steady state traffic was sent through the Epipe to overload each of the policers and the show output below was collected. This output focuses on the policer and arbiter details.

The following shows the policers on the SAP and their current state.

```
A:PE-1# show qos policer sap 1/1/3:1
===============================================================================
Policer Information (Summary), Slot 1
===============================================================================
-------------------------------------------------------------------------------
Name                FC-Maps       MBS       HP-Only A.PIR    A.CIR
Direction                         CBS       Depth   O.PIR    O.CIR    O.FIR
-------------------------------------------------------------------------------
1->1/1/3:1->1
Ingress             be            124 KB    0 KB    100000   0
                                  0 KB      82      30000    0        30000
1->1/1/3:1->2
Ingress             l2            76 KB     0 KB    60000    20000
                                  25 KB     77846   30000    20000    30000
1->1/1/3:1->3
Ingress             af            76 KB     0 KB    60000    20000
                                  25 KB     77824   15000    15000    15000
1->1/1/3:1->4
Ingress             l1            76 KB     0 KB    60000    20000
                                  25 KB     77868   15000    15000    15000
1->1/1/3:1->5
Ingress             ef            12800 B   0 KB    10000    10000
                                  12800 B   12834   10000    10000    10000
===============================================================================
A:PE-1#
```

The output above shows the configured values for the policers, e.g. PIR and CIR, together with their operational (current) state, such as PIR, CIR and FIR. The depth of each of the PIR buckets is also shown.

The detailed state of each policer can be seen by adding the parameter detail. The following is the output for policer 3.

```
A:PE-1# show qos policer sap 1/1/3:1 ingress detail
...
===============================================================================
Policer Info (1->1/1/3:1->3), Slot 1
===============================================================================
Policer Name       : 1->1/1/3:1->3
Direction          : Ingress            Fwding Plane       : 1
FC-Map             : af
Depth PIR          : 77842 Bytes        Depth CIR          : 25618 Bytes
Depth FIR          : 77842 Bytes
MBS                : 76 KB              CBS                : 25 KB
Hi Prio Only       : 0 KB               Pkt Byte Offset    : 0
```

```
Admin PIR          : 60000 Kbps         Admin CIR          : 20000 Kbps
Oper PIR           : 15000 Kbps         Oper CIR           : 15000 Kbps
Oper FIR           : 15000 Kbps
Stat Mode          : offered-total-cir
PIR Adaption       : closest            CIR Adaption       : closest
Parent Arbiter Name: a3
-------------------------------------------------------------------------------
Arbiter Member Information
-------------------------------------------------------------------------------
Offered Rate       : 45800 Kbps
Level              : 2                  Weight             : 25
Parent PIR         : 15000 Kbps         Parent FIR         : 15000 Kbps
Consumed           : 15000 Kbps
-------------------------------------------------------------------------------
===============================================================================...
A:PE-1#
```

Notice that the above output shows the depth of the PIR, CIR and FIR buckets together with their operational rates. This can be used to explain the operation of the policers in this example and is discussed later in this section.

The stat-mode of offered-total-cir configured on policer 3 results in these statistics being collected.

```
A:PE-1# show service id 1 sap 1/1/3:1 stats
===============================================================================
...
-------------------------------------------------------------------------------
Sap per Policer stats
-------------------------------------------------------------------------------
                        Packets                 Octets

Ingress Policer 1 (Stats mode: offered-total-cir)
Off. All              : 2690893                 172217152
Dro. InProf           : 0                       0
Dro. OutProf          : 967465                  61917760
For. InProf           : 0                       0
For. OutProf          : 1723428                 110299392

Ingress Policer 2 (Stats mode: offered-total-cir)
Off. All              : 2690988                 172223232
Dro. InProf           : 0                       0
Dro. OutProf          : 909492                  58207488
For. InProf           : 1178507                 75424448
For. OutProf          : 602989                  38591296
...
```

The following output is included for reference and shows the statistics which are collected for each of the ingress and egress stat-modes.

```
PE-1# show service id 2 sap 1/1/1:2 stats
...
-------------------------------------------------------------------------------
```

```
Sap per Policer stats
-------------------------------------------------------------------------------
                        Packets                 Octets

Ingress Policer 1 (Stats mode: no-stats)

Ingress Policer 2 (Stats mode: minimal)
Off. All              : 0                       0
For. All              : 0                       0
Dro. All              : 0                       0

Ingress Policer 3 (Stats mode: offered-profile-no-cir)
Off. InProf           : 0                       0
Off. OutProf          : 0                       0
For. InProf           : 0                       0
For. OutProf          : 0                       0
Dro. InProf           : 0                       0
Dro. OutProf          : 0                       0

Ingress Policer 4 (Stats mode: offered-priority-no-cir)
Off. HiPrio           : 0                       0
Off. LowPrio          : 0                       0
For. HiPrio           : 0                       0
For. LoPrio           : 0                       0
Dro. HiPrio           : 0                       0
Dro. LowPrio          : 0                       0

Ingress Policer 5 (Stats mode: offered-profile-cir)
Off. InProf           : 0                       0
Off. OutProf          : 0                       0
Off. Uncolor          : 0                       0
For. InProf           : 0                       0
For. OutProf          : 0                       0
Dro. InProf           : 0                       0
Dro. OutProf          : 0                       0

Ingress Policer 6 (Stats mode: offered-priority-cir)
Off. HiPrio           : 0                       0
Off. LowPrio          : 0                       0
For. InProf           : 0                       0
For. OutProf          : 0                       0
Dro. InProf           : 0                       0
Dro. OutProf          : 0                       0

Ingress Policer 7 (Stats mode: offered-total-cir)
Off. All              : 0                       0
For. InProf           : 0                       0
For. OutProf          : 0                       0
Dro. InProf           : 0                       0
Dro. OutProf          : 0                       0

Ingress Policer 8 (Stats mode: offered-limited-profile-cir)
Off. OutProf          : 0                       0
Off. Uncolor          : 0                       0
For. InProf           : 0                       0
For. OutProf          : 0                       0
Dro. InProf           : 0                       0
Dro. OutProf          : 0                       0
```

```
Egress Policer 1 (Stats mode: no-stats)

Egress Policer 2 (Stats mode: minimal)
Off. All              : 0                       0
For. All              : 0                       0
Dro. All              : 0                       0

Egress Policer 3 (Stats mode: offered-profile-no-cir)
Off. InProf           : 0                       0
Off. OutProf          : 0                       0
For. InProf           : 0                       0
For. OutProf          : 0                       0
Dro. InProf           : 0                       0
Dro. OutProf          : 0                       0

Egress Policer 4 (Stats mode: offered-profile-cir)
Off. InProf           : 0                       0
Off. OutProf          : 0                       0
Off. Uncolor          : 0                       0
For. InProf           : 0                       0
For. OutProf          : 0                       0
Dro. InProf           : 0                       0
Dro. OutProf          : 0                       0

Egress Policer 5 (Stats mode: offered-total-cir)
Off. All              : 0                       0
For. InProf           : 0                       0
For. OutProf          : 0                       0
Dro. InProf           : 0                       0
Dro. OutProf          : 0                       0
==============================================================================
```

It is possible to show the policer-control-policy details and the SAPs with which it is associated, as shown here.

```
A:PE-1# show qos policer-control-policy cfhp-1
==============================================================================
QoS Policer Control Policy
==============================================================================
Policy-Name        : cfhp-1
Description        : (Not Specified)
Min Threshold Sep  : Def

------------------------------------------------------------------------------
Priority MBS Thresholds
------------------------------------------------------------------------------
Priority            MBS Contribution
------------------------------------------------------------------------------
1                   none
2                   none
3                   none
4                   none
5                   none
6                   none
7                   none
8                   none
```

```
-------------------------------------------------------------------------------
Tier/Arbiter                      Lvl/Wt      Rate      Parent
-------------------------------------------------------------------------------
  root                            N/A         100000    None
1 a3                              3/1         60000     root

===============================================================================
A:PE-1# show qos policer-control-policy "cfhp-1" association

===============================================================================
QoS Policer Control Policy
===============================================================================
Policy-Name        : cfhp-1
Description        : (Not Specified)

-------------------------------------------------------------------------------
Associations
-------------------------------------------------------------------------------
Service-Id         : 1 (Epipe)          Customer-Id        : 1
 - SAP : 1/1/3:1 (Ing)

===============================================================================
A:PE-1
```

The following command shows the policer hierarchy, including the child policers and their relationship to the intermediate arbiter (a3) and the root arbiter. It can be used to monitor the status of the child policers in the hierarchy. The output shows the assigned, offered and consumed capacity for each policer.

```
A:PE-1# show qos policer-hierarchy sap 1/1/3:1
===============================================================================
Policer Hierarchy - Sap 1/1/3:1
===============================================================================
Ingress Policer Control Policy : cfhp-1
Egress Policer Control Policy  :
-------------------------------------------------------------------------------
root (Ing)
|
| slot(1)
|
|--(A) : a3 (Sap 1/1/3:1)
|   |
|   |--(P) : Policer 1->1/1/3:1->4
|   |   |
|   |   |    [Level 2 Weight 25]
|   |   |    Assigned PIR:15000      Offered:45800
|   |   |    Consumed:15000
|   |   |
|   |   |    Assigned FIR:15000
|   |
|   |--(P) : Policer 1->1/1/3:1->3
|   |   |
|   |   |    [Level 2 Weight 25]
|   |   |    Assigned PIR:15000      Offered:45800
|   |   |    Consumed:15000
|   |   |
|   |   |    Assigned FIR:15000
|   |
|   |--(P) : Policer 1->1/1/3:1->2
|   |   |
|   |   |    [Level 2 Weight 50]
|   |   |    Assigned PIR:30000      Offered:45800
|   |   |    Consumed:30000
|   |   |
|   |   |    Assigned FIR:30000
|
|--(P) : Policer 1->1/1/3:1->5
|   |
|   |    [Level 5 Weight 1]
|   |    Assigned PIR:10000      Offered:10000
|   |    Consumed:10000
|   |
|   |    Assigned FIR:10000
|
|--(P) : Policer 1->1/1/3:1->1
|   |
|   |    [Level 1 Weight 1]
|   |    Assigned PIR:30000      Offered:45800
|   |    Consumed:30000
|   |
|   |    Assigned FIR:30000
root (Egr)
|
No Active Members Found on slot 1
===============================================================================
A:PE-1#
```

The complete information about the policer hierarchy can be seen by adding the detail parameter, as shown below, with alternative parameters to select more specific information.

- root-detail — Rates, depth and thresholds for the root arbiter.
- thresholds — CBS, MBS and high-prio-only thresholds with associated rates of child policers.
- priority-info — Discard-fair and discard-unfair thresholds, with number of associated children, for each of the root priority levels.
- depth — Parent policer and child PIR buckets depth, with PIR and FIR rate information.
- arbiter — Specific information of a given arbiter.
- port — For use with LAGs in different line cards or using adapt-qos link.

The output adds a good representation of the root arbiter thresholds, indicating the priority levels, discard-unfair and discard-fair thresholds, and how many child policers are associated with each level. It also includes the current depth of the child policer PIR buckets and the parent policer bucket.

```
A:PE-1# show qos policer-hierarchy sap 1/1/3:1 detail
===============================================================================
Policer Hierarchy - Sap 1/1/3:1
===============================================================================
Ingress Policer Control Policy : cfhp-1
Egress Policer Control Policy  :
-------------------------------------------------------------------------------
Legend :
(*) real-time dynamic value
(w) Wire rates
-------------------------------------------------------------------------------
root (Ing)
|
| slot(1)
|    MaxPIR:100000
|    ConsumedByChildren:100000
|    OperPIR:100000      OperFIR:100000
|
|    DepthPIR:8111 bytes
|  Priority 8
|    Oper Thresh Unfair:17408       Oper Thresh Fair:25600
|    Association count:0
|  Priority 7
|    Oper Thresh Unfair:17408       Oper Thresh Fair:25600
|    Association count:0
|  Priority 6
|    Oper Thresh Unfair:17408       Oper Thresh Fair:25600
|    Association count:0
|  Priority 5
|    Oper Thresh Unfair:17408       Oper Thresh Fair:25600
|    Association count:1
|  Priority 4
|    Oper Thresh Unfair:9728        Oper Thresh Fair:17408
|    Association count:0
|  Priority 3
|    Oper Thresh Unfair:9728        Oper Thresh Fair:17408
|    Association count:3
|  Priority 2
|    Oper Thresh Unfair:0           Oper Thresh Fair:8192
|    Association count:0
|  Priority 1
|    Oper Thresh Unfair:0           Oper Thresh Fair:8192
|    Association count:1
|
|--(A) : a3 (Sap 1/1/3:1)
|   |    MaxPIR:60000
|   |    ConsumedByChildren:60000
|   |    OperPIR:60000       OperFIR:60000
|   |
|   |    [Level 3 Weight 1]
|   |    Assigned PIR:60000      Offered:60000
|   |    Consumed:60000
|   |
|   |    Assigned FIR:60000
|   |
|   |--(P) : Policer 1->1/1/3:1->4
|   |   |    MaxPIR:60000        MaxCIR:20000
|   |   |    CBS:25600           MBS:77824
|   |   |    HiPrio:0
|   |   |    Depth:77876
|   |   |
|   |   |    OperPIR:15000        OperCIR:15000
|   |   |    OperFIR:15000
|   |   |    PacketByteOffset:0
|   |   |    StatMode: offered-total-cir
|   |   |
|   |   |    [Level 2 Weight 25]
|   |   |    Assigned PIR:15000      Offered:45800
|   |   |    Consumed:15000
|   |   |
|   |   |    Assigned FIR:15000
|   |
|   |--(P) : Policer 1->1/1/3:1->3
|   |   |    MaxPIR:60000       MaxCIR:20000
|   |   |    CBS:25600          MBS:77824
|   |   |    HiPrio:0
|   |   |    Depth:77834
|   |   |
|   |   |    OperPIR:15000        OperCIR:15000
|   |   |    OperFIR:15000
|   |   |    PacketByteOffset:0
|   |   |    StatMode: offered-total-cir
|   |   |
|   |   |    [Level 2 Weight 25]
|   |   |    Assigned PIR:15000      Offered:45800
|   |   |    Consumed:15000
|   |   |
|   |   |    Assigned FIR:15000
|   |
|   |--(P) : Policer 1->1/1/3:1->2
|   |   |    MaxPIR:60000       MaxCIR:20000
|   |   |    CBS:25600          MBS:77824
|   |   |    HiPrio:0
|   |   |    Depth:77848
|   |   |
|   |   |    OperPIR:30000        OperCIR:20000
|   |   |    OperFIR:30000
|   |   |    PacketByteOffset:0
|   |   |    StatMode: offered-total-cir
|   |   |
|   |   |    [Level 2 Weight 50]
|   |   |    Assigned PIR:30000      Offered:45800
|   |   |    Consumed:30000
|   |   |
|   |   |    Assigned FIR:30000
|
|--(P) : Policer 1->1/1/3:1->5
|   |    MaxPIR:10000       MaxCIR:10000
|   |    CBS:12800          MBS:12800
|   |    HiPrio:0
|   |    Depth:12854
|   |
|   |    OperPIR:10000        OperCIR:10000
|   |    OperFIR:10000
|   |    PacketByteOffset:0
|   |    StatMode: offered-total-cir
|   |
|   |    [Level 5 Weight 1]
|   |    Assigned PIR:10000      Offered:10000
|   |    Consumed:10000
|   |
|   |    Assigned FIR:10000
|
|--(P) : Policer 1->1/1/3:1->1
|   |    MaxPIR:100000       MaxCIR:0
|   |    CBS:0               MBS:126976
|   |    HiPrio:0
|   |    Depth:135
|   |
|   |    OperPIR:30000        OperCIR:0
|   |    OperFIR:30000
|   |    PacketByteOffset:0
|   |    StatMode: offered-total-cir
|   |
|   |    [Level 1 Weight 1]
|   |    Assigned PIR:30000      Offered:45800
|   |    Consumed:30000
|   |
|   |    Assigned FIR:30000


root (Egr)
|
No Active Members Found on slot 1

===============================================================================
A:PE-1#
```

The output above gives the depth of the parent policer, which can be used with the output below to explain the operation of the policing in this example.

```
A:PE-1# show qos policer sap 1/1/3:1 detail | match expression "Slot | Bytes | Kbps"
Policer Info (1->1/1/3:1->1), Slot 1
Depth PIR          : 153 Bytes          Depth CIR          : 0 Bytes
Depth FIR          : 153 Bytes
Admin PIR          : 100000 Kbps        Admin CIR          : 0 Kbps
Oper PIR           : 30000 Kbps         Oper CIR           : 0 Kbps
Oper FIR           : 30000 Kbps
Offered Rate       : 45800 Kbps
Parent PIR         : 30000 Kbps         Parent FIR         : 30000 Kbps
Consumed           : 30000 Kbps
Policer Info (1->1/1/3:1->2), Slot 1
Depth PIR          : 77828 Bytes        Depth CIR          : 25624 Bytes
Depth FIR          : 77828 Bytes
Admin PIR          : 60000 Kbps         Admin CIR          : 20000 Kbps
Oper PIR           : 30000 Kbps         Oper CIR           : 20000 Kbps
Oper FIR           : 30000 Kbps
Offered Rate       : 45800 Kbps
Parent PIR         : 30000 Kbps         Parent FIR         : 30000 Kbps
Consumed           : 30000 Kbps
Policer Info (1->1/1/3:1->3), Slot 1
Depth PIR          : 77858 Bytes        Depth CIR          : 25634 Bytes
Depth FIR          : 77858 Bytes
Admin PIR          : 60000 Kbps         Admin CIR          : 20000 Kbps
Oper PIR           : 15000 Kbps         Oper CIR           : 15000 Kbps
Oper FIR           : 15000 Kbps
Offered Rate       : 45800 Kbps
Parent PIR         : 15000 Kbps         Parent FIR         : 15000 Kbps
Consumed           : 15000 Kbps
Policer Info (1->1/1/3:1->4), Slot 1
Depth PIR          : 77838 Bytes        Depth CIR          : 25614 Bytes
Depth FIR          : 77838 Bytes
Admin PIR          : 60000 Kbps         Admin CIR          : 20000 Kbps
Oper PIR           : 15000 Kbps         Oper CIR           : 15000 Kbps
Oper FIR           : 15000 Kbps
Offered Rate       : 45800 Kbps
Parent PIR         : 15000 Kbps         Parent FIR         : 15000 Kbps
Consumed           : 15000 Kbps
Policer Info (1->1/1/3:1->5), Slot 1
Depth PIR          : 12814 Bytes        Depth CIR          : 12814 Bytes
Depth FIR          : 12814 Bytes
Admin PIR          : 10000 Kbps         Admin CIR          : 10000 Kbps
Oper PIR           : 10000 Kbps         Oper CIR           : 10000 Kbps
Oper FIR           : 10000 Kbps
Offered Rate       : 10000 Kbps
Parent PIR         : 10000 Kbps         Parent FIR         : 10000 Kbps
Consumed           : 10000 Kbps
A:PE-1#
```

From the output above, it can be seen that the offered rate for policers 1-4 is 45800Kbps, in fact it is the same for policer 5 but this is capped at the admin PIR rate, 10000Kbps.

The depth of the parent policer is only 8111 bytes, so this is not causing any discarding of priority 2-5 traffic at the parent policer as their discard thresholds are all above this value. Therefore the drops in policers 2-5 are all occurring in the child policers.

Policer 5 is consuming all of its operational capacity (PIR, CIR and FIR), and it can be seen that the level of the PIR bucket is 12814 bytes, which is slightly above its MBS of 12800 bytes. The level of the PIR bucket will oscillate around the MBS value as tokens are added to exceed the threshold (causing discards) then the draining reduces the level to just below the threshold (allowing forwarding).

Policers 2-4 are functioning in the same way as policer 5, as can be seen from their PIR bucket levels (levels are 77828 bytes with MBS of 77824), resulting in the PIR buckets constraining the rates of the traffic through these policers. This is happening because the arbiter “a3” is distributing its 60000Kbps in the configured ratio to these policers, which changes the operational PIR to 30000Kbps for policer 2 and 15000Kbps for policers 3 and 4, all being below the offered traffic rate. A similar effect can be seen with the CIR rates and bucket depths, as the operational CIR rate of policer 2 has reached its administrative value with those of policer 3 and 4 being constrained by the operational PIR. The CIR bucket depths are just above the CBS, again this will oscillate causing traffic to both in-profile and out-of-profile. As this is steady state traffic, the operational FIR rates for these policers have settled to match their operational PIR rates.

Policer 1 is also discarding traffic at the PIR bucket but it is also discarding traffic at the parent policer. This can be seen by the fact that policer 1 PIR depth is nowhere near its MBS whereas the parent policer level is just below the priority 1 discard-fair threshold. The level of the parent policer bucket will oscillate around this threshold causing policer 1 traffic to be discarded, which in turn is reflected back into the level of tokens in the policer 1 PIR bucket.

As this example is based on ingress unicast policing, the traffic exits the policers and then accesses the switch fabric using a set of shared-queue (policer-output-queues). The parameters for these queues can be seen using the following **show** command.

```
A:PE-1# show qos shared-queue "policer-output-queues" detail
===============================================================================
QoS Shared Queue Policy
===============================================================================
-------------------------------------------------------------------------------
Shared Queue Policy (policer-output-queues)
-------------------------------------------------------------------------------
Policy         : policer-output-queues
Description    : Default Policer Output Shared Queue Policy

-------------------------------------------------------------------------------
Queue CIR       PIR       CBS     MBS     HiPrio  Multipoint Pool-Name
-------------------------------------------------------------------------------
1     0         100       1       50      10      FALSE
2     25        100       3       50      10      FALSE
3     25        100       10      50      10      FALSE
4     25        100       3       25      10      FALSE
5     100       100       10      50      10      FALSE
6     100       100       10      50      10      FALSE
7     10        100       3       25      10      FALSE
8     10        100       3       25      10      FALSE
9     0         100       1       50      10      TRUE
10    25        100       3       50      10      TRUE
11    25        100       10      50      10      TRUE
12    25        100       3       25      10      TRUE
13    100       100       10      50      10      TRUE
14    100       100       10      50      10      TRUE
15    10        100       3       25      10      TRUE
16    10        100       3       25      10      TRUE

-------------------------------------------------------------------------------
FC    UCastQ    MCastQ    BCastQ  UnknownQ
-------------------------------------------------------------------------------
be    1         9         9       9
l2    2         10        10      10
af    3         11        11      11
l1    4         12        12      12
h2    5         13        13      13
```

```
ef    6         14        14      14
h1    7         15        15      15
nc    8         16        16      16

-------------------------------------------------------------------------------
Associations
-------------------------------------------------------------------------------
Service : 1             SAP : 1/1/3:1
===============================================================================
A:PE-1#
```

For egress policing, policed traffic can access the exit port by a queue-group, the default being called policer-output-queues. The following shows the parameters for these queues.

```
A:PE-1# show qos queue-group "policer-output-queues" detail
===============================================================================
QoS Queue-Group Ingress
===============================================================================
===============================================================================
QoS Queue-Group Egress
===============================================================================
-------------------------------------------------------------------------------
QoS Queue Group
-------------------------------------------------------------------------------
Group-Name     : policer-output-queues
Description    : Default egress policer output queues.

-------------------------------------------------------------------------------
Q  CIR Admin PIR Admin CBS         HiPrio PIR Lvl/Wt    Parent    BurstLimit(B)
   CIR Rule  PIR Rule  MBS                CIR Lvl/Wt    Wred-Queue     Slope
   Named-Buffer Pool
-------------------------------------------------------------------------------
1  0         max       def         def    1/1           None          default
   closest   closest   def                0/1           disabled      default
   (not-assigned)
2  0         max       def         def    1/1           None          default
   closest   closest   def                0/1           disabled      default
   (not-assigned)

===============================================================================
Queue Group Ports (access)
===============================================================================
Port                Sched Pol           Acctg Pol Stats     Description
-------------------------------------------------------------------------------
1/1/3                                   0         No
1/1/4                                   0         No
-------------------------------------------------------------------------------

===============================================================================
Queue Group Ports (network)
===============================================================================
Port                Sched Pol           Acctg Pol Stats     Description
-------------------------------------------------------------------------------
No Matching Entries

===============================================================================
Queue Group Sap FC Maps
===============================================================================
Sap Policy     FC Name             Queue Id
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================
A:PE-1#
```

The following output shows the relative thresholds in the parent policer when a priority level has 0, 1 or 2 associated children, both with and without using the fixed parameter.

```
A:PE-1# show qos policer-hierarchy sap 1/1/3:2 ingress priority-info
===============================================================================
Policer Hierarchy - Sap 1/1/3:2
===============================================================================
Ingress Policer Control Policy : cfhp-2
-------------------------------------------------------------------------------
root (Ing)
|
| slot(1)
|  Priority 8
|    Oper Thresh Unfair:4352        Oper Thresh Fair:5120
|    Association count:0
|  Priority 7
|    Oper Thresh Unfair:4352        Oper Thresh Fair:5120
|    Association count:0
|  Priority 6
|    Oper Thresh Unfair:4352        Oper Thresh Fair:5120
|    Association count:2 fixed
|  Priority 5
|    Oper Thresh Unfair:3328        Oper Thresh Fair:4096
|    Association count:1 fixed
|  Priority 4
|    Oper Thresh Unfair:2304        Oper Thresh Fair:3072
|    Association count:0 fixed
|  Priority 3
|    Oper Thresh Unfair:1280        Oper Thresh Fair:2048
|    Association count:2
|  Priority 2
|    Oper Thresh Unfair:0           Oper Thresh Fair:1024
|    Association count:1
|  Priority 1
|    Oper Thresh Unfair:0           Oper Thresh Fair:0
|    Association count:0


===============================================================================
A:PE-1#
```

Where

- Priority Level 1 has no children so both its fair and unfair thresholds are 0.

- Priority Level 2 has one child so its unfair threshold is 0 and its fair threshold is at the configured mbs-contribution [1024 bytes] (given that this is larger than the min-thresh-separation).
- Priority Level 3 has two children so its unfair threshold is equal to the min-thresh-separation plus the fair threshold of priority 2 [256+1024=1280 bytes]. Its fair threshold is effectively the mbs-contribution plus the fair threshold of priority 2 [1024+1024=2048 bytes] (given that the mbs-contribution is larger than 2x min-thresh-separation).
- Priorities 4, 5 and 6 have the fixed parameter configured. Even though priority 4 has no children, priority 5 has only one child and priority 6 has two children, all three priorities have the same incremental values for their unfair and fair discard threshold. This result in
  - Priority 4’s unfair threshold being equal to the min-thresh-separation plus the fair threshold of priority 3 [256+2048=2304 bytes]. Its fair threshold is effectively the mbs-contribution plus the fair threshold of priority 3 [1024+2048=3072 bytes] (given that the mbs-contribution is larger than 2x min-thresh-separation).
  - Priority 5’s unfair threshold being equal to the min-thresh-separation plus the fair threshold of priority 4 [256+3072=3328 bytes]. Its fair threshold is effectively the mbs-contribution plus the fair threshold of priority 4 [1024+3072=4096 bytes] (given that the mbs-contribution is larger than 2x min-thresh-separation).
  - Priority 6’s unfair threshold being equal to the min-thresh-separation plus the fair threshold of priority 5 [256+4096=4352 bytes]. Its fair threshold is effectively the mbs-contribution plus the fair threshold of priority 5 [1024+4096=5120 bytes] (given that the mbs-contribution is larger than 2x min-thresh-separation).

Note that the above parameter values were chosen to exactly match available hardware values to simplify the output.

Conclusion

This note has described the configuration of Class Fair Hierarchical Policing for SAPs. This hardware policing provides low latency ingress and egress prioritized traffic control with the ability to provide fairness between child policers at the same parent policer priority level.

FP and Port Queue Groups

This chapter provides information about FP and port queue groups.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter is applicable to the 7950 XRS-16c/20, 7750 SR-7/12, 7750 SR-a4/8, 7750 SR-c4/12, and 7450 ESS-6/6v/7/12 platforms and assumes only FP2- and higher-based line cards are used.

Port queue groups can be configured on FP1 line cards but not all functions described in this chapter are supported on FP1 line cards. FP queue groups are not supported on FP1 line cards because FP queue groups rely on hardware policing.

The configuration was tested on release 13.0.R7. There are no other specific prerequisites for this configuration.

Overview

Queue groups provide flexible QoS control beyond that available by default for SAPs and network interfaces.

Many applications require detailed QoS control for SAPs, with aggregated QoS control across the core on network interfaces. Queue groups allow the reverse, specifically aggregated QoS control for multiple SAPs and per-network interface QoS control. This is summarized in Table 20.

Table 20 **Default QoS and Queue Group Comparison**

	Default	Queue group
SAPs	• Per-SAP ingress queues and policers • Per-SAP egress queues and policers	• Policers for a set of SAPs at ingress • Queues for a set of SAPs at egress
Network interfaces	• Per-MDA ingress network queues • Per-port egress network queues	• Policers per network interface at ingress • Policers and queues per network interface at egress

Queue groups were introduced in SR OS release 7 as a mechanism of grouping a set of queues and enhanced in subsequent releases. Release 10 added the ability to configure policers within queue groups and introduced a more flexible configuration where the queue group template could be applied multiple times with different instances. The term queue groups was retained even though they can now contain queues, policers, or both. This chapter describes this new configuration.

Queue groups can be used for other applications than those listed in Table 20; for example:

- Pseudowire (PW) QoS

  Providing QoS control for spoke SDPs in the various pipe services, different types of VPLS services, and associated IES/VPRN interfaces (see chapter Pseudowire QoS).
- Carrier Supporting Carrier (CSC) services

  Providing QoS control for CSC network interfaces in VPRN services.
- Ingress QoS control on VPRN network interfaces

  Providing for ingress QoS control of unicast traffic into a VPRN over automatically created or manually created bindings in a VPRN service.
- Network egress QoS

  Providing queue rates in kbps instead of the percentage of the port rate, and queue CBS/MBS in kbytes instead of a fractional percentage of the pool size.

When egress queue groups are used for SAPs, the groups provide a similar functionality to multi-service sites in that both can provide an overall rate for a set of SAPs. However, multi-service sites also provide per-SAP QoS control, whereas the queue groups do not.

Queue groups are not applicable to subscriber management except when egress policing is used.

Note: The queue groups described in this chapter are different from those that are configured when using a high-scale Ethernet MDA (HS-MDAv2).

Configuration

Following the description of the overall configuration of queue groups, their configuration is illustrated using examples on a SAP, a network interface, and with egress subscriber policing.

The steps required to configure a queue group are summarized as follows:

- Create a queue group template:
  - Ingress or egress
- Apply an instance of the queue group template to:
  - FP ingress
    - Access or network
  - Ethernet egress port
    - Access or network
- Redirect traffic to a policer or queue in the FP or port queue group instance per forwarding class (FC) within a:
  - SAP ingress or egress QoS policy
  - Network QoS policy for both ingress and egress
- Consider post-egress policer queuing

Creating Queue Group Templates

Queue group templates are configured separately for ingress and egress; an ingress queue group template with the same name as an egress queue group template is a different object.

Ingress queue group templates can contain policers or queues, but not both.

Egress queue group templates can contain policers or queues, or both; queue 1 is created by default and cannot be deleted. A forwarding class (FC) can also be configured to determine the mapping of the FC of the forwarded traffic to a policer or queue.

To summarize the use of policers with respect to queue groups:

- Ingress FP queue groups can only use policers.
- Port egress access queue groups cannot use policers.
- Network ingress and egress applications can only use policers in a queue group.

The configuration of an ingress and egress queue group template is as follows:

```
configure
    qos
        queue-group-templates
            ingress
                queue-group <queue-group-name> [create]
                    description <description-string>
                    policer <policer-id> [create]
                    queue <queue-id> [multipoint] [<queue-type>]
                                     [<queue-mode>] [create]
            egress
                queue-group <queue-group-name> [create]
                    description <description-string>
                    fc <fc-name> [create]
                        queue <queue-id>
                    policer <policer-id> [create]
                    queue <queue-id> [<queue-type>] [create]
```

The configuration of the policer and queue is similar to the configuration in a SAP ingress and egress QoS policy. However, the SAP ingress QoS policy allows the configuration of a percent-rate and the SAP egress QoS policy allows the configuration of an avg-frame-overhead, neither of which is available in a queue group template. An ingress queue group template supports the creation of up to 32 queues and 32 policers, while an egress queue group template supports the creation of up to 8 queues and 8 policers.

Because an egress queue group template uses kbps for the CIR/PIR rates and bytes/kbytes for the CBS/MBS, unlike a network-queue QoS policy, a network egress interface can use them when an egress queue group template is applied.

The system instantiates the following queue group templates by default. Table 22 shows the ingress queue group templates:

Table 21 **Queue Group Templates - Ingress**

Group name	Description
_tmnx_nat_ing_q_grp **_tmnx_nat_ing_q_grp_v2** **_tmnx_lns_esm_ing_q_grp**	NAT/LNS Ingress Queue Group Template NAT/LNS Ingress Queue Group Template for ISAv2 LNS ESM Ingress Queue Group Template

Table 22 shows the egress queue group templates:

Table 22 **Queue Group Templates - Egress**

Group name	Description
_tmnx_nat_egr_q_grp **policer-output-queues** **_tmnx_nat_egr_q_grp_v2** **_tmnx_lns_esm_egr_q_grp**	NAT/LNS Egress Queue Group Template Default egress policer output queues NAT/LNS Egress Queue Group Template for ISAv2 LNS ESM Egress Queue Group Template

This chapter will only discuss the policer-output-queues queue group from the preceding table, an instance of which is created on all egress access and hybrid ports, to be used for egress policed traffic.

Applying Queue Group Templates

Ingress

An ingress queue group template containing only policers can be applied to an FP ingress. When applied, an instance identifier must be specified, which represents the instantiated instance of the related queue group template.

If an attempt is made to configure a queue group template containing queues for an FP ingress, the following error message appears:

```
*A:PE-1>config#  card 5 fp 1 ingress access queue-group "qg2" instance 1 create
MINOR: CHMGR #1164 Cannot attach a Queue Group containing queues
```

Because 7750 SR-a4/8 platforms do not support SAP or network hardware policers, FP ingress queue groups are not supported on those platforms.

Queue group templates containing only queues can be applied to port ingress. However, only one such queue group can be applied per ingress port.

An ingress template can be applied to an FP ingress as either an ingress access queue group or an ingress network queue group, or both. In each case:

- An accounting policy can be applied.
- Statistics collection can be enabled.
- A description can be configured.
- A policer control policy can be applied, containing parent arbiters for the queue group policers, and its parameters can be overridden.
- Policer parameters configured within the queue group template can also be overridden.

Note: When the queue group applies to a tunnel object that can move between different network interfaces, and consequently different network ingress FPs (for example, PW QoS), a network ingress queue group instance must be applied to each FP ingress that could be used.

The configuration of ingress queue group templates for ingress FPs is as follows:

```
configure
    card <slot-number>
        fp [<fp-number>]
            ingress
                access
                    queue-group <queue-group-name>
                                instance <[1..65535]> [create]
                        accounting-policy <acct-policy-id>
                        collect-stats
                        description <description-string>
                        policer-control-override [create]
                            max-rate {<rate> | max}
                            priority-mbs-thresholds
                                min-thresh-separation <size> [bytes | kilobytes]
                                priority <level>
                                   mbs-contribution <size> [bytes | kilobytes]
                        policer-control-policy
                                <policer-control-policy-name>
                        policer-override
                            policer <policer-id> [create]
               network
                    queue-group <queue-group-name>
                                instance <[1..65535]> [create]
                        accounting-policy <acct-policy-id>
                        collect-stats
                        description <description-string>
                        policer-control-override [create]
                            max-rate {<rate> | max}
                            priority-mbs-thresholds
```

```
                                    min-thresh-separation <size> [bytes | kilobytes]
                                    priority <level>
                                       mbs-contribution <size> [bytes | kilobytes]
                            policer-control-policy
                                    <policer-control-policy-name>
                            policer-override
                                policer <policer-id> [create]
```

Note: Configure FP ingress access and network queue group instances consistently across FPs relating to a LAG.

A single ingress template containing only queues can also be applied to an Ethernet port ingress, which can be used only for access ingress:

- An accounting policy can be applied.
- Statistics collection can be enabled.
- A description can be configured.
- Queue parameters configured within the queue group template can be overridden.
- A scheduler policy can be applied containing parent schedulers for the queue group queues.

The configuration of an ingress queue group template for an ingress port is as follows:

```
configure
    port <port-id>
        ethernet
            access
                ingress
                    queue-group <queue-group-name> [create]
                        accounting-policy <acct-policy-id>
                        collect-stats
                        description <description-string>
                        queue-overrides
                            queue <queue-id> [create]
                        scheduler-policy <scheduler-policy-name>
```

Egress

An egress queue group template containing policers or queues, or both, can be applied to an Ethernet port network egress. Only queue group templates containing queues (not policers) can be applied to an Ethernet port access egress. When applied, an instance identifier can be specified, which represents the instantiated instance of the related queue group template; the default being 1.

If an attempt is made to configure a queue group template containing policers for an Ethernet port access egress, the following error message appears:

```
*A:PE-1# configure port 5/1/5 ethernet access egress queue-
group "qg1" instance 1 create
MINOR: CLI Could not create/change "qg1".
MINOR: PMGR #1324 Cannot attach a Queue Group containing policers
```

An egress template can be applied as either an egress access queue group or an egress network queue group, or both. In each case:

- An accounting policy can be applied.
- Statistics collection can be enabled.
- A description can be configured.
- Queue parameters configured within the queue group template can be overridden.
- A scheduler policy can be applied containing parent schedulers for the queue group queues.

When applied to the Ethernet access egress, a host-match can be configured, which is used to select a queue group for subscriber egress policed traffic; see later in this chapter for details.

When applied to the Ethernet network egress, a policer control policy can be applied, which contains parent arbiters for the queue group policers.

Note: When the queue group applies to a tunnel object that can move between different network interfaces, and consequently different network egress ports (for example, PW QoS), a network egress queue group instance must be applied to each network egress port that could be used.

The configuration of egress queue group templates for egress ports is as follows:

```
configure
    port <port-id>
        ethernet
            access
                egress
                    queue-group <queue-group-name> [create]
                                [instance <instance-id>]
                        accounting-policy <acct-policy-id>
                        agg-rate
                            limit-unused-bandwidth
                            queue-frame-based-accounting
                            rate <kilobits-per-second>
                        collect-stats
                        description <description-string>
                        host-match dest <destination-string> [create]
```

```
         queue-overrides
             queue <queue-id> [create]
         scheduler-policy <scheduler-policy-name>
network
     queue-group <queue-group-name> [create]
                 [instance <instance-id>]
         accounting-policy <acct-policy-id>
         agg-rate
             limit-unused-bandwidth
             queue-frame-based-accounting
             rate <kilobits-per-second>
         collect-stats
         description <description-string>
         policer-control-policy
                 <policer-control-policy-name>
         queue-overrides
             queue <queue-id> [create]
         scheduler-policy <scheduler-policy-name>
```

Note: When port egress queue groups are used with a LAG, the system enforces a consistent configuration between the ports (based on only the configuration of the primary LAG port).

Redirecting Traffic to a Queue Group Queue or Policer

There are multiple ways to redirect traffic to a queue group queue or policer. The redirection is always on a per-FC basis so that different FCs within the same policy can use a mix of a local SAP policer, a local SAP or network queue, or a queue group policer or queue (where applicable).

Redirection to a queue group queue or policer is not supported for subscribers; attempting to do so will display the following errors:

```
*A:PE-1>config>qos>sap-ingress# fc "af" policer 1 fp-redirect-group
MINOR: QOS #1489 Cannot assign a queue-group because SLA profile references to this
policy exist

*A:PE-1>config>qos>sap-egress# fc af policer 1 port-redirect-group-queue
MINOR: QOS #1628 Cannot assign a queue-group because SLA profile references to this
policy exist

*A:PE-1>config>qos>sap-egress# fc af queue 1 port-redirect-group-queue
MINOR: QOS #1628 Cannot assign a queue-group because SLA profile references to this
policy exist

*A:PE-1>config>subscr-mgmt>sla-prof>ingress# qos 40
MINOR: SUBMGR #1110 Cannot assign Qos Policy, QOS Plcy contains queue-
group references

*A:PE-1>config>subscr-mgmt>sla-prof>egress# qos 40
MINOR: SUBMGR #1110 Cannot assign Qos Policy, QOS Plcy contains queue-
```

```
group references
```

SAP Ingress Redirection

Traffic can be redirected per FC to a policer in an FP ingress access queue group, using the fp-redirect-group parameter in the SAP ingress QoS policy, as follows:

```
configure
    qos
        sap-ingress <policy-id> [create]
            fc <fc-name> [create]
                policer <policer-id> fp-redirect-group
```

The policer-id is referencing a policer in the FP ingress queue group.

The queue group name and instance to be used is not yet configured, allowing this QoS policy to be applied to objects using different FP ingress queue group instances, which results in greater flexibility.

Compare this preferred configuration to the original configuration used for ingress port queue groups, as follows, where the queue group name is specified within the SAP ingress QoS policy and the instance is not available:

```
configure
    qos
        sap-ingress <policy-id> [create]
            fc <fc-name> [create]
                queue <queue-id> group <queue-group-name>
```

In the preferred configuration, redirection is completed by applying the SAP ingress QoS policy to the SAP with the queue group name and instance to be used, as follows:

```
configure
    service
        {apipe|cpipe|epipe|fpipe|ipipe} <service-id>
            sap <sap-id>
                ingress
                    qos <policy-id> fp-redirect-group <queue-group-name>
                                             instance <instance-id>
        {ies|vprn} <service-id>
            interface <ip-int-name>
                sap <sap-id>
                    ingress
                        qos <policy-id>
                            fp-redirect-group <queue-group-name>
                                     instance <instance-id>
```

SAP Egress Redirection

Traffic can be redirected per FC to a queue in a port egress access queue group, using the port-redirect-group parameter, as follows:

```
configure
    qos
        sap-egress <policy-id> [create]
            fc <fc-name> [create]
                queue <queue-id> port-redirect-group-queue
```

The queue-id is referencing a queue in the port egress queue group.

The queue group name and instance to be used is not yet configured, allowing this QoS policy to be applied to objects using different port egress queue groups.

Compare this preferred configuration to the original configuration, which allows the queue group name with an instance to be specified:

```
queue <id> {group <grp-name> [instance instance-id]}
```

In the preferred configuration, redirection is completed by applying the SAP egress QoS policy to the SAP with the queue group name and instance to be used, as follows:

```
configure
    service
        {apipe|cpipe|epipe|fpipe|ipipe} <service-id>
            sap <sap-id>
                egress
                    qos <policy-id>
                        port-redirect-group <queue-group-name>
                                        instance <instance-id>
        {ies|vprn} <service-id>
            interface <ip-int-name>
                sap <sap-id>
                    egress
                        qos <policy-id>
                            port-redirect-group <queue-group-name>
                                            instance <instance-id>
```

Network Ingress Redirection

Traffic can be redirected per FC to a policer in an FP ingress network queue group, using the fp-redirect-group parameter, as follows:

```
configure
    qos
        network <network-policy-id> [create]
```

```
            ingress
                fc <fc-name>
                    fp-redirect-group <policer-type> <policer-id>
```

The policer-id is referencing a policer in the FP ingress queue group.

The traffic usage for each policer type for the supported services is shown in Table 4.

Table 23 **Network Ingress FP Queue Group Policer Usage**

Policer Type	Usage
broadcast-policer	Broadcast traffic for PW QoS in a VPLS service
mcast-policer	Multipoint traffic (except for ingress QoS control on VPRN network interfaces and CSC services where the IP multicast traffic uses the ingress network queues or queue group related to the network interface) Multicast traffic for PW QoS in a VPLS service
unknown-policer	Unknown traffic for PW QoS in a VPLS service
policer	Unicast traffic

The queue group name and instance to be used is not yet configured, allowing this QoS policy to be applied to objects using different FP ingress queue group instances.

The redirection is completed by applying the network QoS policy to the required object with the queue group name and instance to be used, as follows (only the network interface redirection is shown):

```
configure
    router
        interface <interface-name>
            qos <network-policy-id>
                        ingress-fp-redirect-group <queue-group-name>
                        ingress-instance <instance-id>
```

Network Egress Redirection

Traffic can be redirected per FC to a policer or a queue in a port egress network queue group, using the port-redirect-group parameter, as follows:

```
configure
    qos
        network <network-policy-id> [create]
            egress
                fc <fc-name>
```

```
                    port-redirect-group {queue <queue-id> |
                                  policer <plcr-id> [queue <queue-id>]}
```

The queue-id and policer-id are referencing a queue and a policer in the port egress queue group.

The queue group name and instance to be used are not yet configured, allowing this QoS policy to be applied to objects using different port egress queue groups.

The redirection is completed by applying the network QoS policy to the required object with the queue group name and instance to be used, as follows (only the network interface redirection is shown):

```
configure
    router
        interface <interface-name>
            qos <network-policy-id>
                        egress-fp-redirect-group <queue-group-name>
                        egress-instance <instance-id>
```

Note: Non-IPv4/non-IPv6/non-MPLS packets are not subject to the redirection to an egress queue group instance and will remain on the regular port network egress queues. When using an egress port scheduler, parent the related regular network port queues to appropriate port scheduler priority levels, to ensure the required operation under port congestion. This is important for protocol traffic, such as LACP, EFMOAM, ETH-CFM, ARP, and IS-IS, which by default use the FC nc regular network port queue.

Post-egress Policer Queuing

The queuing of traffic exiting an egress policer is described as follows for SAP egress and network egress.

SAP egress policed traffic exits a port, using either a local queue or an egress queue group instance queue. If an egress queue group instance queue is to be used, the policed traffic can be mapped to the queue in the following ways, listed in reverse preference order (first is lowest preference):

1. By default, SAP egress policed traffic exits using a queue in the policer-output-queues queue group instance when only a policer is configured:

```
configure
    qos
        sap-egress <policy-id> [create]
            fc <fc-name> [create]
                policer <policer-id>
```

The system always creates this egress queue group template and applies it to all access and hybrid ports (for use by the access part of the hybrid port) as instance 1. The queue group template contains the mapping for all FCs to one of the two created queues to determine which queue is to be used.

The default configuration of the policer-output-queues group template is as follows (without the queue details):

```
configure
    qos
        queue-group-templates
            egress
                queue-group "policer-output-queues" create
                    description "Default egress policer
                                 output queues."
                    queue 1 best-effort create
                    exit
                    queue 2 expedite create
                    exit
                    fc af create
                        queue 1
                    exit
                    fc be create
                        queue 1
                        exit
                    fc ef create
                        queue 2
                    exit
                    fc h1 create
                        queue 2
                    exit
                    fc h2 create
                        queue 2
                    exit
                    fc l1 create
                        queue 1
                    exit
                    fc l2 create
                        queue 1
                    exit
                    fc nc create
                        queue 2
                    exit
                exit
```

This queue group template cannot be deleted, but its configuration can be modified. The configuration commands under a port access egress queue group instance are also available under the policer-output-queues queue group instance.

2. The FC in the SAP egress QoS policy can be mapped to a policer with its traffic redirected to a port access egress queue group, as follows:

```
configure
    qos
        sap-egress <policy-id> [create]
            fc <fc-name> [create]
```

```
                    policer <policer-id> port-redirect-group-queue
```

The QoS policy must be applied to the SAP with a redirection to a queue group instance. By default, queue 1 in the queue group will be used, but other queues can be created and used by creating an FC-to-queue mapping to them.

3. Similar to step 2, the FC in the SAP egress QoS policy can be mapped to a policer with its traffic redirected to a port access egress queue group instance and the queue within that queue group instance to be used, as follows:

```
configure
    qos
        sap-egress <policy-id> [create]
            fc <fc-name> [create]
                policer <policer-id> port-redirect-group-queue queue <id>
```

The QoS policy must again be applied to the SAP with a redirection to a queue group instance. In this case, the queue specified with the port-redirect-group-queue will be used.

Network egress can only use policers that are created within an egress queue group instance, and this requires the egress FC to have the port-redirect-group parameter configured.

With the following configuration, traffic exits the port on the network egress queue to which its FC is mapped (not on a queue group queue):

```
configure
    qos
        network <network-policy-id> [create]
            egress
                fc <fc-name>
                    port-redirect-group policer <plcr-id>
```

If a queue is specified on the port-redirect-group statement, as follows, the traffic exits on the referenced queue within the port network egress queue group instance:

```
configure
    qos
        network <network-policy-id> [create]
            egress
                fc <fc-name>
                    port-redirect-group policer <plcr-id> queue <queue-id>
```

Configuration Examples

The following configuration examples are for the use of queue groups with a SAP, network interface, and egress policed traffic subscriber traffic. Different queue group templates and instances have been used to highlight the flexibility of the configuration.

SAP Configuration Example

In this example, a SAP is created in an IES service using ingress queue group qg1 instance 1 and egress queue group qg1 instance 1.

First, the queue group templates are configured, as follows:

```
configure
    qos
        queue-group-templates
            ingress
                queue-group "qg1" create
                    policer 1 create
                    exit
                    policer 2 create
                    exit
                exit
            exit
            egress
                queue-group "qg1" create
                    queue 1 best-effort create
                    exit
                    queue 2 expedite create
                    exit
                exit
                queue-group "policer-output-queues" create
                    queue 3 best-effort create
                    exit
                    fc l2 create
                        queue 3
                    exit
                exit
```

The policer-output-queues queue group template has been modified to add an extra queue and map FC l2 to it. This is used by the traffic mapped to policer 1 in the SAP egress QoS policy.

The ingress template is applied to card 5 fp 1 to create an FP access queue group instance, as follows:

```
configure
    card 5
        fp 1
```

```
            ingress
                access
                    queue-group "qg1" instance 1 create
                    exit
                exit
```

The egress template is applied to port 5/1/5 to create a port access queue group instance, as follows:

```
configure
    port 5/1/5
        ethernet
            mode hybrid
            encap-type dot1q
            access
                egress
                    queue-group "qg1" instance 1 create
                    exit
                exit
```

The SAP ingress QoS policy is created to redirect FC af to policer 1 and FC ef to policer 2 in the FP ingress queue group. To emphasize that the redirection is per FC and can coexist with locally mapped queues or policers, FC be is mapped to the local queue 1 and FC l2 to local queue 2, as follows:

```
configure
    qos
        sap-ingress 10 create
            queue 1 create
            exit
            queue 2 create
            exit
            queue 11 multipoint create
            exit
            fc "af" create
                policer 1 fp-redirect-group
            exit
            fc "be" create
                queue 1
            exit
            fc "ef" create
                policer 2 fp-redirect-group
            exit
            fc "l2" create
                queue 2
            exit
            dot1p 0 fc "be"
            dot1p 1 fc "l2"
            dot1p 2 fc "af"
            dot1p 3 fc "ef"
        exit
```

The SAP egress QoS policy is created to redirect FC af to queue 1 in the port egress queue group and FC ef to local policer 2, with its traffic exiting through queue 2 in the port egress queue group. FC be is mapped to the local queue 1 and FC l2 to local policer 1. The policer 1 traffic will exit using queue 3 of the policer-output-queues queue group, because the mapping for FC l2 has been modified in its template, as follows:

```
configure
    qos
        sap-egress 10 create
            queue 1 create
            exit
            policer 1 create
            exit
            policer 2 create
            exit
            fc af create
                queue 1 port-redirect-group-queue
            exit
            fc be create
                queue 1
            exit
            fc ef create
                policer 2 port-redirect-group-queue queue 2
            exit
            fc l2 create
                policer 1
            exit
        exit
```

The IES service is created with an interface on SAP 5/1/5:1, using the preceding ingress and egress QoS policies and queue group instances. A second interface (not shown) is used to forward the traffic to and from this interface, as follows:

```
configure
    service
        ies 1 customer 1 create
           interface "PE-1-int1-2" create
                address 10.1.2.1/24
                sap 5/1/5:1 create
                    ingress
                        qos 10 fp-redirect-group "qg1" instance 1
                    exit
                    egress
                        qos 10 port-redirect-group "qg1" instance 1
                    exit
                exit
            exit
            no shutdown
        exit
```

The following output shows the configuration and the QoS state after sending traffic through the service.

The traffic statistics are either counted in the SAP queues and policers, or in the queue group instance queues and policers, but not in both. However, summary statistics per SAP are available when using FP ingress queue groups, as shown.

The queue group templates can be shown. The output shows the details related to the ingress queue group template:

```
*A:PE-1# show qos queue-group "qg1" ingress detail
===============================================================================
QoS Queue-Group Ingress
===============================================================================
-------------------------------------------------------------------------------
QoS Queue Group
-------------------------------------------------------------------------------
Group-Name     : qg1
Description    : (Not Specified)
-------------------------------------------------------------------------------
Q  Mode    CIR Admin  PIR Admin  CBS          HiPrio  PIR Lvl/Wt Parent
           CIR Rule   PIR Rule   MBS                  CIR Lvl/Wt BurstLimit(B)
           Named-Buffer Pool     Pkt Bt Ofst          Adv Config Policy Name
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================
Queue Group Ports
===============================================================================
Port                Sched Pol           Acctg Pol Stats     Description
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================
Queue Group Sap FC Maps
===============================================================================
Sap Policy     FC Name             Queue (id type)
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================
Queue Group FP Maps
===============================================================================
Card Num       Fp Num              Instance            Type
-------------------------------------------------------------------------------
5              1                   1                   Access
-------------------------------------------------------------------------------
Entries found: 1
-------------------------------------------------------------------------------
===============================================================================
Queue Group Policer
===============================================================================
Policer Id     : 1
Description    : (Not Specified)
PIR Adptn      : closest                     CIR Adptn    : closest
Parent         : none                        Level        : 1
Weight         : 1                           Adv. Cfg Plcy: none
Admin PIR      : max                         Admin CIR    : 0
CBS            : def                         MBS          : def
Hi Prio Only   : def                         Pkt Offset   : 0
Profile Capped : Disabled
StatMode       : minimal
===============================================================================
Policer Id     : 2
Description    : (Not Specified)
PIR Adptn      : closest                       CIR Adptn    : closest
Parent         : none                          Level        : 1
Weight         : 1                             Adv. Cfg Plcy: none
Admin PIR      : max                           Admin CIR    : 0
CBS            : def                           MBS          : def
Hi Prio Only   : def                           Pkt Offset   : 0
Profile Capped : Disabled
StatMode       : minimal
===============================================================================
*A:PE-1#
```

The following output shows the details related to the egress queue group template:

```
*A:PE-1# show qos queue-group "qg1" egress detail
===============================================================================
QoS Queue-Group Egress
===============================================================================
-------------------------------------------------------------------------------
QoS Queue Group
-------------------------------------------------------------------------------
Group-Name     : qg1
Description    : (Not Specified)
-------------------------------------------------------------------------------
Q CIR Admin  PIR Admin  CBS          HiPrio PIR Lvl/Wt Parent     BurstLimit(B)
  CIR Rule   PIR Rule   MBS                 CIR Lvl/Wt Wred-Queue Slope
  Named-Buffer Pool     Pkt Bt Ofst         Adv Config Policy Name
-------------------------------------------------------------------------------
1 0          max        def          def    1/1        None       default
  closest    closest    def                 0/1        disabled   default
  (not-assigned)        add 0               (not-assigned)
2 0          max        def          def    1/1        None       default
  closest    closest    def                 0/1        disabled   default
  (not-assigned)        add 0               (not-assigned)
===============================================================================
Queue Group FC Mapping
===============================================================================
FC Name                                Queue-Id
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================
===============================================================================
Queue Group Ports (access)
===============================================================================
Port      Sched Pol          Acctg Pol Stats Description          QGrp-Instance
-------------------------------------------------------------------------------
5/1/5                        0         No                         1
-------------------------------------------------------------------------------
===============================================================================
Queue Group Ports (network)
===============================================================================
Port   Sched Pol    Policer-Ctrl-Pol  Acctg Pol Stats Description QGrp-Instance
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================
Qos Sap-Egress FC Group-Queue References
===============================================================================
Sap Policy     FC Name             Queue Id
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================
Qos Sap-Egress FC Port-Redirect-Group-Queue References
===============================================================================
Sap Policy     FC Name             Queue Id
-------------------------------------------------------------------------------
10             af                  1
10             ef                  2
-------------------------------------------------------------------------------
Entries found: 2
-------------------------------------------------------------------------------
===============================================================================
Queue Group Policer
===============================================================================
No Matching Entries
-------------------------------------------------------------------------------
HSMDA PIR Admin  Packet  WRR     MBS       Slope Plcy        WRR Plcy
Queue PIR Rule   Offset  Weight            Max Class         Burst Lmt
-------------------------------------------------------------------------------
1     max        add 0   1       default   default           n/a
      closest                              8                 default
2     max        add 0   1       default   default           n/a
      closest                              8                 default
3     max        add 0   1       default   default           n/a
      closest                              8                 default
4     max        add 0   1       default   default           n/a
      closest                              8                 default
5     max        add 0   1       default   default           n/a
      closest                              8                 default
6     max        add 0   1       default   default           n/a
      closest                              8                 default
7     max        add 0   1       default   default           n/a
      closest                              8                 default
8     max        add 0   1       default   default           n/a
      closest                              8                 default
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

The preceding output shows that the ingress template has an instance 1, which is applied to card 5 fp 1, and the egress template has an instance 1, which is applied to port 5/1/5.

The remapping of FC l2 to queue 3 in the policer-output-queues queue group template is as follows:

```
*A:PE-1# show qos queue-group "policer-output-queues" egress detail | match post-
lines 12 "Queue Group FC Mapping"
Queue Group FC Mapping
===============================================================================
FC Name                                 Queue-Id
-------------------------------------------------------------------------------
af                                      1
be                                      1
ef                                      2
```

```
h1                                      2
h2                                      2
l1                                      1
l2                                      3
nc                                      2
===============================================================================
*A:PE-1#
```

For SAP ingress QoS policy 10, the redirection is true for FC af using policer 1 and FC ef using policer 2 to an FP ingress queue group, as follows:

```
*A:PE-1# show qos sap-ingress 10 detail | match post-
lines 5 expression "af Unicast|ef Unicast"
FC                 : af Unicast
-------------------------------------------------------------------------------
Policer            : 1                  Queue              : def
Queue-Group        : None
FP Queue-Group     : True
FC                 : ef Unicast
-------------------------------------------------------------------------------
Policer            : 2                  Queue              : def
Queue-Group        : None
FP Queue-Group     : True
```

In the following output for SAP egress QoS policy 10:

- The redirection is true for FC af using queue 1 in the queue group instance, to be specified when the policy is applied to a SAP.
- The redirection is true for FC ef using local policer 2 with its traffic exiting queue 2 in the queue group instance, to be specified when the policy is applied to a SAP.
- FC l2 is using local policer 1 with its traffic exiting using queue 3 in the policer-output-queues port access egress queue group.

```
*A:PE-1# show qos sap-egress 10 detail | match post-lines 6 expression "Queue-Group"
FC  Queue  Queue-Group                      InstanceId   SapBReDir  Plcr
-------------------------------------------------------------------------------
be  1      n/a                              n/a          False      None
l2  3      policer-output-queues            1            False      1
af  1      n/a                              n/a          True       None
ef  2      n/a                              n/a          True       2
```

The QoS information, including the queue group instances being used for the IES service SAP, is shown as follows:

```
*A:PE-1# show service id 1 sap 5/1/5:1 detail | match post-lines 4 QOS
QOS
-------------------------------------------------------------------------------
Ingress qos-policy : 10                       Egress qos-policy : 10
Ingress FP QGrp    : qg1                      Egress Port QGrp  : qg1
Ing FP QGrp Inst   : 1                        Egr Port QGrp Inst: 1
```

In the following output for the egress SAP, FC ef is using policer 2 and its traffic exits using queue 2 in queue group qg1 instance 1, while FC l2 is using policer 1 and its traffic exits using queue 3 of the policer-output-queues queue group instance 1:

```
*A:PE-1# show qos policer sap 5/1/5:1 egress detail | match post-
lines 4 "Policer Info"
Policer Info (1->5/1/5:1->1), Slot 5
===============================================================================
Policer Name       : 1->5/1/5:1->1
Direction          : Egress             Fwding Plane       : 1
FC->[QGrp:Inst->]Q : l2->policer-output-queues:1->3
Policer Info (1->5/1/5:1->2), Slot 5
===============================================================================
Policer Name       : 1->5/1/5:1->2
Direction          : Egress             Fwding Plane       : 1
FC->[QGrp:Inst->]Q : ef->qg1:1->2
```

The associations of the port access egress queue group are shown as follows:

```
*A:PE-1# show port 5/1/5 queue-group "qg1" instance 1 egress access associations
===============================================================================
Ethernet port 5/1/5 Access Egress queue-group
===============================================================================
Queue-Group Name   : qg1
Instance-Id        : 1
-------------------------------------------------------------------------------
Subscriber-Host Queue-Group Associations
-------------------------------------------------------------------------------
No associations
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
SAP Based Queue-Group Association
-------------------------------------------------------------------------------
Service-Id               SAP                      Qos Policy-Id
-------------------------------------------------------------------------------
1 (IES)                  5/1/5:1                  10
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Qos Policy Based Queue-Group Association
-------------------------------------------------------------------------------
FC Name   Qos Policy-Id       Local Policer-Id    Local Queue-Id      QGrp-Q
-------------------------------------------------------------------------------
No associations
-------------------------------------------------------------------------------
===============================================================================
===============================================================================
*A:PE-1#
```

After traffic is sent through the service, the FP ingress access queue group policer statistics are as follows:

```
*A:PE-1# show card 5 fp 1 ingress queue-
group "qg1" instance 1 mode access statistics
===============================================================================
Card:5  Acc.QGrp: qg1  Instance: 1
===============================================================================
Group Name    : qg1
Description   : (Not Specified)
Pol Ctl Pol   : None                    Acct Pol      : None
Collect Stats : disabled
-------------------------------------------------------------------------------
Statistics
-------------------------------------------------------------------------------
                        Packets                 Octets
Ing. Policer:  1  Grp: qg1 (Stats mode: minimal)
Off. All              : 1000                    128000
Dro. All              : 0                       0
For. All              : 1000                    128000
Ing. Policer:  2  Grp: qg1 (Stats mode: minimal)
Off. All              : 1000                    128000
Dro. All              : 0                       0
For. All              : 1000                    128000
===============================================================================
*A:PE-1#
```

The traffic sent through the port egress access queue group queues is as follows:

```
*A:PE-1# show port 5/1/5 queue-group "qg1" instance 1 egress access statistics
-------------------------------------------------------------------------------
Ethernet port 5/1/5 Access Egress queue-group
-------------------------------------------------------------------------------
                        Packets                 Octets
Egress Queue: 1 Group: qg1 Instance: 1
In Profile forwarded  : 0                       0
In Profile dropped    : 0                       0
Out Profile forwarded : 1000                    128000
Out Profile dropped   : 0                       0
Egress Queue: 2 Group: qg1 Instance: 1
In Profile forwarded  : 0                       0
In Profile dropped    : 0                       0
Out Profile forwarded : 1000                    128000
Out Profile dropped   : 0                       0
-------------------------------------------------------------------------------
*A:PE-1#
```

Finally, the FC l2 traffic using the egress policer 1 is in queue 3, in the policer-output-queues queue group instance statistics, as follows:

```
*A:PE-1# show port 5/1/5 queue-group "policer-output-
queues" instance 1 egress access statistics
-------------------------------------------------------------------------------
Ethernet port 5/1/5 Access Egress queue-group
-------------------------------------------------------------------------------
                        Packets                 Octets
Egress Queue: 1 Group: policer-output-queues Instance: 1
In Profile forwarded  : 0                       0
In Profile dropped    : 0                       0
Out Profile forwarded : 0                       0
Out Profile dropped   : 0                       0
Egress Queue: 2 Group: policer-output-queues Instance: 1
In Profile forwarded  : 0                       0
In Profile dropped    : 0                       0
Out Profile forwarded : 0                       0
Out Profile dropped   : 0                       0
Egress Queue: 3 Group: policer-output-queues Instance: 1
In Profile forwarded  : 0                       0
In Profile dropped    : 0                       0
Out Profile forwarded : 1000                    128000
Out Profile dropped   : 0                       0
-------------------------------------------------------------------------------
*A:PE-1#
```

The number of valid ingress packets received on a SAP, or subscribers on that SAP, in the sap-stats output, are shown as follows. The received valid counter includes both the local SAP counters and the counters from the related FP ingress queue group instance. This is useful to display SAP-level traffic statistics when forwarding classes in a SAP ingress policy have been redirected to an ingress queue group.

```
*A:PE-1# show service id 1 sap 5/1/5:1 sap-stats | match post-
lines 6 "Forwarding Engine Stats"
Forwarding Engine Stats
Dropped               : 0                       0
Received Valid        : 4000                    512000
Off. HiPrio           : 0                       0
Off. LowPrio          : 2000                    256000
Off. Uncolor          : 0                       0
Off. Managed          : 0                       0
```

Traffic forwarded through FP ingress access, port ingress access, and port egress access queue groups can be monitored, as follows:

```
monitor card <slot-number> fp <fp-number>  ingress {access|network} queue-group
<queue-group-name> instance <instance-id> [interval <seconds>][repeat<repeat>]
policer <policer-id> [absolute | percent-rate [<reference-rate>]]

monitor port queue-group <queue-group-name> ingress <access> ingress-queue <ingress-
queue-id> [interval <seconds>] [repeat <repeat>] [absolute|rate]

monitor port queue-group <queue-group-name> egress <access> [instance <instance-
id>] [egress-queue <egress-queue-id>] [interval <seconds>] [repeat <repeat>]
[absolute|rate]
```

The summary of the queue groups applied to a port is shown as follows:

```
*A:PE-1# show port 5/1/5 queue-group summary
===============================================================================
Port queue-group summary
===============================================================================
Access-egress queue groups:
----------------------------
qg1
policer-output-queues
Total number of access-egress queue groups  : 2
Network-egress queue groups:
----------------------------
Total number of network-egress queue groups : 0
Access-ingress queue groups:
----------------------------
```

```
Total number of access-ingress queue groups : 0
===============================================================================
*A:PE-1#
```

The total usage of queue groups is shown as follows:

```
*A:PE-1# show qos queue-group summary
===============================================================================
Queue-group instances per card
===============================================================================
card      port-acc-ing  port-acc-egr  port-nw-egr   fp-acc-ing    fp-nw-ing
-------------------------------------------------------------------------------
1         0             0             0             0             0
2         0             0             0             0             0
3         0             0             0             0             0
4         0             0             0             0             0
5         0             3             0             1             0
-------------------------------------------------------------------------------
Total ingress QG templates per system :  4
Total egress  QG templates per system :  5
===============================================================================
*A:PE-1#
```

The preceding output includes the created ingress template plus the three system-created ingress templates (making four in total), and the created egress template plus the four system-created egress templates (making five in total). There is one applied FP access ingress queue group instance on card 5. There are three port access egress queue group instances (the applied queue group instance and two instances of the policer-output-queues queue group), one on each access port used for IES service interfaces.

Network Interface Configuration Example

A network interface is created using ingress queue group qg2 instance 2 and egress queue group qg2 instance 2. If the goal is to provide per-network interface QoS on a single port, each network interface would be configured on a separate VLAN.

First, the queue group templates are configured, as follows:

```
configure
    qos
        queue-group-templates
            ingress
                queue-group "qg2" create
                    policer 1 create
                    exit
                    policer 2 create
                    exit
                exit
            exit
            egress
```

```
                    queue-group "qg2" create
                        queue 1 best-effort create
                        exit
                        queue 2 expedite create
                        exit
                        policer 1 create
                        exit
                        policer 2 create
                        exit
                    exit
                exit
```

The ingress template is applied to card 5 fp 1 to create an FP network queue group instance, as follows:

```
configure
    card 5
        fp 1
            ingress
                network
                    queue-group "qg2" instance 2 create
                    exit
                exit
```

The egress template is applied to port 5/1/5 to create a port network queue group instance, as follows:

```
configure
    port 5/1/5
        ethernet
            mode hybrid
            encap-type dot1q
            network
                egress
                    queue-group "qg2" instance 2 create
                    exit
                exit
```

The network QoS policy is created to:

- Ingress

  Redirect FC af to policer 1 and FC ef to policer 2 in the FP ingress queue group. FC be and FC l2 continue to use their default mapping to the local network ingress queues 1 and 2, respectively.
- Egress

  Redirect FC af to queue 1 in the port egress queue group and FC ef to policer 2 in the port egress queue group, with its traffic exiting through queue 2 of the port egress queue group. FC l2 is redirected to policer 1 in the port egress queue group, with its traffic exiting through the default network egress queue mapped by FC l2; that is, queue 2. FC be continues to use the default network egress queue 1.

```
configure
    qos
        network 10 create
            ingress
                dot1p 0 fc be profile out
                dot1p 1 fc l2 profile out
                dot1p 2 fc af profile out
                dot1p 3 fc ef profile in
                fc af
                    fp-redirect-group policer 1
                exit
                fc ef
                    fp-redirect-group policer 2
                exit
            exit
            egress
                fc af
                    port-redirect-group queue 1
                exit
                fc ef
                    port-redirect-group policer 2 queue 2
                exit
                fc l2
                    port-redirect-group policer 1
                exit
            exit
        exit
```

The network interface is created on port 5/1/5:2 using the preceding network QoS policy with the ingress and egress being redirected to created queue group instances. A second interface (not shown) is used to forward the traffic to and from this network interface.

```
configure
    router Base
        interface "PE-1-int2-2"
            address 10.2.2.1/24
            port 5/1/5:2
            qos 10 egress-port-redirect-group "qg2" egress-instance 2
                   ingress-fp-redirect-group "qg2" ingress-instance 2
            no shutdown
        exit
```

The following output shows the details of the configuration and the QoS state after sending traffic through the network interface.

The traffic statistics are either counted in the network interface queues or in the queue group instance queues and policers, but not in both.

The queue group templates can be shown. The output shows the details related to the ingress queue group template:

```
*A:PE-1# show qos queue-group "qg2" ingress detail
===============================================================================
QoS Queue-Group Ingress
```

```
===============================================================================
-------------------------------------------------------------------------------
QoS Queue Group
-------------------------------------------------------------------------------
Group-Name     : qg2
Description    : (Not Specified)
-------------------------------------------------------------------------------
Q  Mode    CIR Admin  PIR Admin  CBS          HiPrio  PIR Lvl/Wt Parent
           CIR Rule   PIR Rule   MBS                  CIR Lvl/Wt BurstLimit(B)
           Named-Buffer Pool     Pkt Bt Ofst          Adv Config Policy Name
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================
Queue Group Ports
===============================================================================
Port                Sched Pol           Acctg Pol Stats     Description
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================
Queue Group Sap FC Maps
===============================================================================
Sap Policy     FC Name             Queue (id type)
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================
Queue Group FP Maps
===============================================================================
Card Num       Fp Num              Instance            Type
-------------------------------------------------------------------------------
5              1                   2                   Network
-------------------------------------------------------------------------------
Entries found: 1
-------------------------------------------------------------------------------
===============================================================================
Queue Group Policer
===============================================================================
Policer Id     : 1
Description    : (Not Specified)
PIR Adptn      : closest                      CIR Adptn    : closest
Parent         : none                         Level        : 1
Weight         : 1                            Adv. Cfg Plcy: none
Admin PIR      : max                          Admin CIR    : 0
CBS            : def                          MBS          : def
Hi Prio Only   : def                          Pkt Offset   : 0
Profile Capped : Disabled
StatMode       : minimal
===============================================================================
Policer Id     : 2
Description    : (Not Specified)
PIR Adptn      : closest                      CIR Adptn    : closest
Parent         : none                         Level        : 1
Weight         : 1                            Adv. Cfg Plcy: none
Admin PIR      : max                          Admin CIR    : 0
CBS            : def                          MBS          : def
Hi Prio Only   : def                          Pkt Offset   : 0
Profile Capped : Disabled
StatMode       : minimal
===============================================================================
*A:PE-1#
```

The following output shows the details related to the egress queue group template:

```
*A:PE-1# show qos queue-group "qg2" egress detail
===============================================================================
QoS Queue-Group Egress
===============================================================================
-------------------------------------------------------------------------------
QoS Queue Group
-------------------------------------------------------------------------------
Group-Name     : qg2
Description    : (Not Specified)
-------------------------------------------------------------------------------
Q CIR Admin  PIR Admin  CBS          HiPrio PIR Lvl/Wt Parent     BurstLimit(B)
  CIR Rule   PIR Rule   MBS                 CIR Lvl/Wt Wred-Queue Slope
  Named-Buffer Pool     Pkt Bt Ofst         Adv Config Policy Name
-------------------------------------------------------------------------------
1 0          max        def          def    1/1        None       default
  closest    closest    def                 0/1        disabled   default
  (not-assigned)        add 0               (not-assigned)
2 0          max        def          def    1/1        None       default
  closest    closest    def                 0/1        disabled   default
  (not-assigned)        add 0               (not-assigned)
===============================================================================
Queue Group FC Mapping
===============================================================================
FC Name                                 Queue-Id
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================
===============================================================================
Queue Group Ports (access)
===============================================================================
Port      Sched Pol          Acctg Pol Stats Description          QGrp-Instance
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================
Queue Group Ports (network)
===============================================================================
Port   Sched Pol    Policer-Ctrl-Pol  Acctg Pol Stats Description QGrp-Instance
-------------------------------------------------------------------------------
5/1/5                                           No                2
-------------------------------------------------------------------------------
===============================================================================
Qos Sap-Egress FC Group-Queue References
===============================================================================
Sap Policy     FC Name             Queue Id
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================
Qos Sap-Egress FC Port-Redirect-Group-Queue References
===============================================================================
Sap Policy     FC Name             Queue Id
-------------------------------------------------------------------------------
No Matching Entries
===============================================================================
Queue Group Policer
===============================================================================
Policer Id     : 1
Description    : (Not Specified)
```

```
PIR Adptn      : closest                    CIR Adptn    : closest
Parent         : none                       Level        : 1
Weight         : 1                          Adv. Cfg Plcy: none
Admin PIR      : max                        Admin CIR    : 0
CBS            : def                        MBS          : def
Hi Prio Only   : def                        Pkt Offset   : 0
Profile Capped : Disabled
StatMode       : minimal
===============================================================================
Policer Id     : 2
Description    : (Not Specified)
PIR Adptn      : closest                    CIR Adptn    : closest
Parent         : none                       Level        : 1
Weight         : 1                          Adv. Cfg Plcy: none
Admin PIR      : max                        Admin CIR    : 0
CBS            : def                        MBS          : def
Hi Prio Only   : def                        Pkt Offset   : 0
Profile Capped : Disabled
StatMode       : minimal
-------------------------------------------------------------------------------
HSMDA PIR Admin  Packet  WRR     MBS       Slope Plcy        WRR Plcy
Queue PIR Rule   Offset  Weight            Max Class         Burst Lmt
-------------------------------------------------------------------------------
1     max        add 0   1       default   default           n/a
      closest                              8                 default
2     max        add 0   1       default   default           n/a
      closest                              8                 default
3     max        add 0   1       default   default           n/a
      closest                              8                 default
4     max        add 0   1       default   default           n/a
      closest                              8                 default
5     max        add 0   1       default   default           n/a
      closest                              8                 default
6     max        add 0   1       default   default           n/a
      closest                              8                 default
7     max        add 0   1       default   default           n/a
      closest                              8                 default
8     max        add 0   1       default   default           n/a
      closest                              8                 default
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

The preceding output shows that the ingress template has an instance 2, which is applied to card 5 fp 1, and the egress template has an instance 2, which is applied to port 5/1/5.

The details of network QoS policy 10 shows the redirection to the ingress and egress queue groups, as follows:

```
*A:PE-1# show qos network 10 detail | match expression "Egress Forwarding Class
Mapping|Ingress Forwarding Class Mapping|FC Value|Redirect"
Egress Forwarding Class Mapping
FC Value          : 0                   FC Name           : be
Redirect Grp Q    :  None               Redirect Grp Plcr:  None
FC Value          : 1                   FC Name           : l2
Redirect Grp Q    :  None               Redirect Grp Plcr:  1
```

```
FC Value         : 2                      FC Name          : af
Redirect Grp Q   :  1                     Redirect Grp Plcr:  None
FC Value         : 3                      FC Name          : l1
Redirect Grp Q   :  None                  Redirect Grp Plcr:  None
FC Value         : 4                      FC Name          : h2
Redirect Grp Q   :  None                  Redirect Grp Plcr:  None
FC Value         : 5                      FC Name          : ef
Redirect Grp Q   :  2                     Redirect Grp Plcr:  2
FC Value         : 6                      FC Name          : h1
Redirect Grp Q   :  None                  Redirect Grp Plcr:  None
FC Value         : 7                      FC Name          : nc
Redirect Grp Q   :  None                  Redirect Grp Plcr:  None
Ingress Forwarding Class Mapping
FC Value                : 0             FC Name                 : be
Redirect UniCast Plcr   : None          Redirect MultiCast Plcr : None
Redirect BroadCast Plcr : None          Redirect Unknown Plcr   : None
FC Value                : 1             FC Name                 : l2
Redirect UniCast Plcr   : None          Redirect MultiCast Plcr : None
Redirect BroadCast Plcr : None          Redirect Unknown Plcr   : None
FC Value                : 2             FC Name                 : af
Redirect UniCast Plcr   : 1             Redirect MultiCast Plcr : None
Redirect BroadCast Plcr : None          Redirect Unknown Plcr   : None
FC Value                : 3             FC Name                 : l1
Redirect UniCast Plcr   : None          Redirect MultiCast Plcr : None
Redirect BroadCast Plcr : None          Redirect Unknown Plcr   : None
FC Value                : 4             FC Name                 : h2
Redirect UniCast Plcr   : None          Redirect MultiCast Plcr : None
Redirect BroadCast Plcr : None          Redirect Unknown Plcr   : None
FC Value                : 5             FC Name                 : ef
Redirect UniCast Plcr   : 2             Redirect MultiCast Plcr : None
Redirect BroadCast Plcr : None          Redirect Unknown Plcr   : None
FC Value                : 6             FC Name                 : h1
Redirect UniCast Plcr   : None          Redirect MultiCast Plcr : None
Redirect BroadCast Plcr : None          Redirect Unknown Plcr   : None
FC Value                : 7             FC Name                 : nc
Redirect UniCast Plcr   : None          Redirect MultiCast Plcr : None
Redirect BroadCast Plcr : None          Redirect Unknown Plcr   : None
*A:PE-1#
```

The preceding output shows that:

- Ingress
  - FC af is redirected to unicast policer 1 in the queue group instance, to be specified when the policy is applied to a network interface.
  - FC ef is redirected to unicast policer 2 in the queue group instance, to be specified when the policy is applied to a network interface.
- Egress
  - FC af is redirected to queue 1 in the queue group instance, to be specified when the policy is applied to a network interface.
  - FC ef is redirected to policer 2 and queue 2 in the queue group instance, to be specified when the policy is applied to a network interface, so that the policer 2 traffic exits using queue 2 in the port network egress queue group instance.

The queue group instances used by the network interface are shown as follows:

```
*A:PE-1# show router interface "PE-1-int2-2" detail | match post-lines 3 "QoS Queue-
Group Redirection Details"
QoS Queue-Group Redirection Details
-------------------------------------------------------------------------------
Ingress FP QGrp  : qg2                  Egress Port QGrp  : qg2
Ing FP QGrp Inst : 2                    Egr Port QGrp Inst: 2
```

After traffic is sent through the network, it can be shown in the FP ingress network queue group policers, as follows:

```
*A:PE-1# show card 5 fp 1 ingress queue-group "qg2" instance 2 mode network
statistics
===============================================================================
Card:5  Net.QGrp: qg2  Instance: 2
===============================================================================
Group Name    : qg2
Description   : (Not Specified)
Pol Ctl Pol   : None                    Acct Pol      : None
Collect Stats : disabled
-------------------------------------------------------------------------------
Statistics
-------------------------------------------------------------------------------
                        Packets                 Octets
Ing. Policer:  1  Grp: qg2 (Stats mode: minimal)
Off. All             : 1000                    128000
Dro. All             : 0                       0
For. All             : 1000                    128000
Ing. Policer:  2  Grp: qg2 (Stats mode: minimal)
Off. All             : 1000                    128000
Dro. All             : 0                       0
For. All             : 1000                    128000
===============================================================================
*A:PE-1#
```

The traffic sent through the port egress network queue group queues can also be shown, as follows:

```
*A:PE-1# show port 5/1/5 queue-group "qg2" instance 2 egress network statistics
-------------------------------------------------------------------------------
Ethernet port 5/1/5 Network Egress queue-group
-------------------------------------------------------------------------------
                        Packets                 Octets
Egress Queue:  1    Group: qg2    Instance-Id:  2
In Profile forwarded  : 0                       0
In Profile dropped    : 0                       0
Out Profile forwarded : 1000                    128000
Out Profile dropped   : 0                       0
Egress Queue:  2    Group: qg2    Instance-Id:  2
In Profile forwarded  : 0                       0
In Profile dropped    : 0                       0
Out Profile forwarded : 1000                    128000
Out Profile dropped   : 0                       0
Egress Policer:  1  Group: qg2  Instance-Id: 2
Stats mode: minimal
```

```
Off. All              : 1000                    128000
Dro. All              : 0                       0
For. All              : 1000                    128000
Egress Policer:  2  Group: qg2  Instance-Id: 2
Stats mode: minimal
Off. All              : 1000                    128000
Dro. All              : 0                       0
For. All              : 1000                    128000
-------------------------------------------------------------------------------
*A:PE-1#
```

In the preceding output, the traffic on queue 1 is FC af, on policer 1 is FC l2, and on policer 2 is FC ef, with the post-policed traffic on queue 2.

Finally, the FC l2 traffic using the network egress queue group policer 1 can be shown in queue 2 of the default network egress queues, as follows:

```
*A:PE-1# show port 5/1/5 detail | match post-lines 4 "Egress Queue  2"
Egress Queue  2              Packets                 Octets
     In Profile forwarded  :    0                       0
     In Profile dropped    :    0                       0
     Out Profile forwarded :    1000                    128000
     Out Profile dropped   :    0                       0
```

Traffic forwarded through both FP ingress network and port egress network queue groups can be monitored, as follows:

```
monitor card <slot-number> fp <fp-number>  ingress {access|network} queue-group
<queue-group-name> instance <instance-id> [interval <seconds>][repeat<repeat>]
policer <policer-id> [absolute | percent-rate [<reference-rate>]]

monitor port queue-group <queue-group-name> egress <access> [instance <instance-id>]
[egress-queue <egress-queue-id>] [interval <seconds>] [repeat <repeat>]
[absolute|rate]
```

The summary of the queue groups applied to a port is shown as follows:

```
*A:PE-1# show port 5/1/5 queue-group summary

===============================================================================
Port queue-group summary
===============================================================================
Access-egress queue groups:
----------------------------
policer-output-queues
Total number of access-egress queue groups  : 1

Network-egress queue groups:
----------------------------
qg2
Total number of network-egress queue groups : 1

Access-ingress queue groups:
----------------------------
Total number of access-ingress queue groups : 0
```

```
===============================================================================
*A:PE-1#
```

The total usage of queue groups is shown as follows:

```
*A:PE-1# show qos queue-group summary

===============================================================================
Queue-group instances per card
===============================================================================
card      port-acc-ing  port-acc-egr  port-nw-egr   fp-acc-ing    fp-nw-ing
-------------------------------------------------------------------------------
1         0             0             0             0             0
2         0             0             0             0             0
3         0             0             0             0             0
4         0             0             0             0             0
5         0             2             1             0             1
-------------------------------------------------------------------------------
Total ingress QG templates per system :  4
Total egress  QG templates per system :  5
===============================================================================
*A:PE-1#
```

The preceding output includes the created ingress template plus the three system-created ingress templates (making four in total), and the created egress template plus the four system-created egress templates (making five in total). There is the one applied FP network ingress queue group instance on card 5. There is the one created port network egress queue group instance. There are also two port access egress queue group instances, which are the policer-output-queues queue group instances associated with the access port used for IES service interface (not discussed) and the access side of the hybrid port 5/1/5.

Egress Policed Subscriber Configuration Example

Queue groups are only applicable to subscribers for egress policed traffic. By default, subscriber egress policed traffic exits the port using a queue in the egress access policer-output-queues queue group instance. The queue used is determined by the FC mapping in the policer-output-queues queue group template. This is the same default operation as in the SAP example.

The subscriber policed traffic can be sent to a different queue group instance using the inter-dest-id and a host-match (described when applying an egress access queue group template to a port), which represent an intermediate destination, such as a downstream DSLAM or GPON OLT. The inter-dest-id can be associated with a subscriber host when it is created; this would usually be received from the DHCP, or RADIUS server, or Diameter Gx (Policy and Rule Charging Function), or the local user database, or configured under a static host.

An alternative to host matching on an inter-dest-id is to match on the top VLAN tag when a QinQ SAP is configured using a default inter-dest-id.

A default inter-dest-id can be configured in IES, VPRN, and VPLS services, and under an MSAP policy, as follows:

```
configure
    service
        {ies|vprn} <service-id>
            subscriber-interface <ip-int-name>
                group-interface <ip-int-name> [create]
                    sap <sap-id>
                        sub-sla-mgmt
                              def-inter-dest-id {<inter-dest-string>|use-top-q}


configure
    service
        vpls <service-id>
            sap <sap-id>
                sub-sla-mgmt
                    def-inter-dest-id {<inter-dest-string>|use-top-q}


configure
    subscriber-mgmt
        msap-policy <msap-policy-name> [create]
                sub-sla-mgmt
                    def-inter-dest-id {<inter-dest-string>|use-top-q}
```

The egress queue group template is configured as follows:

```
configure
    qos
        queue-group-templates
            egress
                queue-group "qg3" create
                    queue 1 best-effort create
                    exit
                    queue 2 expedite create
                    exit
                    fc af create
                        queue 1
                    exit
                    fc ef create
                        queue 2
                    exit
                exit
            exit
```

The egress template is applied to port 5/1/5 to create a port access queue group instance. A host match is configured under the created queue group instance on an egress access port. In the following, the host match dslam-1 is used:

```
configure
    port 5/1/5
```

```
        ethernet
            mode hybrid
            encap-type dot1q
            access
                egress
                    queue-group "qg3" instance 1 create
                        host-match dest "dslam-1" create
                    exit
                exit
```

A host-match can only be configured under instance 1 of a port access egress queue group; if its configuration is attempted on a different instance, the following error is displayed:

```
*A:PE-1# configure port 5/1/5 ethernet access egress queue-group "qg3" instance 3
create host-match dest another-dslam create
MINOR: PMGR #1337 Host match entries only supported on port access egress queue
groups with system default instance 1
*A:PE-1#
```

The subscriber host uses a SAP egress QoS policy in an egress SLA profile to map FCs to egress queue and policers. The SAP egress QoS policy is created with FC af using policer 1 and FC ef using policer 2, as follows:

```
configure
    qos
        sap-egress 30 create
            queue 1 create
            exit
            queue 2 create
            exit
            policer 1 create
            exit
            policer 2 create
            exit
            fc af create
                policer 1
            exit
            fc be create
                queue 1
            exit
            fc ef create
                policer 2
            exit
            fc l2 create
                queue 2
            exit
        exit
```

To redirect the subscriber egress policed traffic to the access egress queue group qg3 instance 1 on port 5/1/5, which is configured with the host-match, an inter-dest-id is configured for the created subscriber static host, as follows:

```
configure
    service
```

```
        vprn 3 customer 1 create
            route-distinguisher 65536:200
            subscriber-interface "sub-int-1" create
                address 10.3.2.1/24
                group-interface "group-int-1" create
                    arp-populate
                    sap 5/1/5:3 create
                        sub-sla-mgmt
                            def-sub-profile "basic-sub"
                            def-sla-profile "basic-sla"
                            multi-sub-sap 200
                            single-sub-parameters
                                profiled-traffic-only
                            exit
                            no shutdown
                        exit
                        static-host ip 10.3.2.2 mac 00:00:10:03:02:02 create
                            inter-dest-id "dslam-1"
                            sla-profile "basic-sla"
                            sub-profile "basic-sub"
                            subscriber "sub1"
                            no shutdown
                        exit
                    exit
```

The host match configured under the egress queue group instance is shown as follows:

```
A:PE-1# show port 5/1/5 queue-group "qg3" instance 1 access | match post-
lines 3 Host-Matches
Host-Matches
-------------------------------------------------------------------------------
Dest: dslam-1
-------------------------------------------------------------------------------
```

When the subscriber host is created, the inter-dest-id for the subscriber host is shown as follows:

```
A:PE-1# show service active-subscribers subscriber "sub1" detail | match expression
"Subscriber sub1 |Sub. Int Dest Id"
Subscriber sub1 (basic-sub)
Sub. Int Dest Id : "dslam-1"
```

The inter-dest-id dslam-1 is matched against the host-match destination configured on the access egress group instances on the port on which the host is being created. If a match is found, the subscriber egress policed traffic will use that egress queue group instance, with the actual queue used being selected by the FC-to-queue mapping in the related queue group template. Otherwise, the default policer-output-queues queue group instance will be used.

The egress queue group instance subscriber host associations are shown as follows:

```
A:PE-1# show port 5/1/5 queue-group egress "qg3" associations | match post-
lines 6 Subscriber-Host
```

```
Subscriber-Host Queue-Group Associations
-------------------------------------------------------------------------------
svc-id : 3 (VPRN)
 sap   : 5/1/5:3
 subscr: sub1
 ip    : 10.3.2.2
 mac   : 00:00:10:03:02:02  pppoe-sid: N/A
```

The following output shows that FC ef is using policer 2 and its traffic exits using queue 2 in queue group qg3, while FC af is using policer 1 and its traffic exits using queue 1 in queue group qg3.

```
A:PE-1# show qos policer subscriber "sub1" egress detail | match post-
lines 4 "Policer Info"
Policer Info (Sub=1:1 3->5/1/5:3->2), Slot 5
===============================================================================
Policer Name        : Sub=1:1 3->5/1/5:3->2
Direction           : Egress             Fwding Plane       : 1
FC->[QGrp:Inst->]Q : ef->qg3->2
Policer Info (Sub=1:1 3->5/1/5:3->1), Slot 5
===============================================================================
Policer Name        : Sub=1:1 3->5/1/5:3->1
Direction           : Egress             Fwding Plane       : 1
FC->[QGrp:Inst->]Q : af->qg3->1
```

After traffic is sent through the subscriber, the egress policed traffic can be shown in the port egress access queue group instance, as follows:

```
A:PE-1# show port 5/1/5 queue-group "qg3" instance 1 access egress statistics

-------------------------------------------------------------------------------
Ethernet port 5/1/5 Access Egress queue-group
-------------------------------------------------------------------------------
                        Packets                 Octets

Egress Queue: 1 Group: qg3 Instance: 1
In Profile forwarded  : 0                       0
In Profile dropped    : 0                       0
Out Profile forwarded : 1000                    128000
Out Profile dropped   : 0                       0
Egress Queue: 2 Group: qg3 Instance: 1
In Profile forwarded  : 0                       0
In Profile dropped    : 0                       0
Out Profile forwarded : 1000                    128000
Out Profile dropped   : 0                       0
-------------------------------------------------------------------------------
A:PE-1#
```

The traffic statistics are either counted in the subscriber queues or policers, or in the queue group instance queues, but not in both. However, summary statistics per SAP are available when using FP ingress queue groups.

The number of valid ingress packets received on a SAP, or subscribers on that SAP, can be shown in the sap-stats output, as follows. The received valid counter includes both the local SAP counters and the counters from the related FP ingress queue group instance. This is useful to display SAP-level traffic statistics when forwarding classes in a SAP ingress policy have been redirected to an ingress queue group.

```
*A:PE-1# show service id 3 sap 5/1/5:3 sap-stats | match post-
lines 6 "Forwarding Engine Stats"
Forwarding Engine Stats
Dropped               : 0                      0
Received Valid        : 4000                   512000
Off. HiPrio           : 0                      0
Off. LowPrio          : 0                      0
Off. Uncolor          : 0                      0
Off. Managed          : 0                      0
```

Traffic forwarded through port egress access queue groups can be monitored, as follows:

```
monitor port queue-group <queue-group-name> egress <access> [instance <instance-id>]
[egress-queue <egress-queue-id>] [interval <seconds>] [repeat <repeat>]
[absolute|rate]
```

The summary of the queue groups applied to a port is shown as follows:

```
A:PE-1# show port 5/1/5 queue-group summary

===============================================================================
Port queue-group summary
===============================================================================
Access-egress queue groups:
----------------------------
qg3
policer-output-queues
Total number of access-egress queue groups  : 2

Network-egress queue groups:
----------------------------
Total number of network-egress queue groups : 0

Access-ingress queue groups:
----------------------------
Total number of access-ingress queue groups : 0
===============================================================================
A:PE-1#
```

The total usage of queue groups is shown as follows:

```
A:PE-1# show qos queue-group summary

===============================================================================
Queue-group instances per card
===============================================================================
card      port-acc-ing  port-acc-egr  port-nw-egr   fp-acc-ing    fp-nw-ing
-------------------------------------------------------------------------------
```

```
1         0             0             0             0             0
2         0             0             0             0             0
3         0             0             0             0             0
4         0             0             0             0             0
5         0             3             0             0             0
-------------------------------------------------------------------------------
Total ingress QG templates per system :  3
Total egress  QG templates per system :  5
===============================================================================
A:PE-1#
```

The preceding output includes the three system-created ingress templates, and the created egress template plus the four system-created egress templates (making five in total). There are three port access egress queue group instances (the applied queue group instance and two instances of the policer-output-queues queue group), one on each access port used for VPRN service interfaces.

Conclusion

This chapter described the use of queue groups as a mechanism to provide an aggregate QoS control for multiple SAPs and per-network interface QoS control. The configuration steps and commands are described, followed by example configurations on a SAP, network interface, and for egress policed traffic subscriber traffic.

High Scale QoS IOM: QoS, Service, and Network Configuration

This chapter provides information about High Scale QoS IOM: QoS, Service, and Network Configuration.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter is applicable to the 7750 SR-7/12/12e platforms and describes the High Scale QoS (HSQ) IOM. The configuration was tested on Release 15.0.R5.

Overview

This chapter describes the QoS operation and configuration of the HSQ IOM, with a focus on services and network interfaces. For the subscriber management configuration, see chapters *High Scale QoS IOM in ESM Context - Single SLA Mode* and *High Scale QoS IOM in ESM Context - Expanded SLA Mode*.

The HSQ IOM is an FP3-based IOM that has a multicore CPU and accepts up to two MDA-e cards. The HSQ IOM supports an enhanced egress QoS architecture to provide scalable network, service, and subscriber QoS. At ingress, the HSQ IOM supports regular FP3 QoS with a high ingress policer scaling. This chapter focuses on the HSQ IOM egress QoS.

The HSQ IOM supports six scheduling classes across multiple hierarchical levels of hardware egress shaping with very stringent egress burst control. The scheduling allows a mix of strict priority and weighted round-robin (WRR). A flexible buffer pool structure permits both buffer isolation and buffer oversubscription for the queue buffer allocation.

The HSQ IOM supports 768k queues, which are grouped into 96k queue groups; each comprises eight queues (referred to as HSQ queue groups). HSQ queue groups are used for SAP egress queues, network egress queues, and both access and network egress queue group instance queues.

The SAP egress, network egress, and access and network egress queue group related commands that are not supported with an HSQ IOM are provided in the associated configuration chapters following. In addition, the following are not applicable to the HSQ:

- QoS related
  - Egress access and network MDA and port pools
  - All HSMDA commands
  - All VPORT related commands
  - PBB egress B-SAP per ISID shaping
  - Port **hybrid-buffer-allocation egr-weight**
- MPLS related
  - Generalized Multiprotocol Label Switching (GMPLS) UNI
- Service related
  - G.8031 protected Ethernet tunnels
- System related
  - Port cross-connects (PXC)
  - Ethernet satellite host ports
  - Soft reset

The operation of the HSQ IOM is described in the following sections:

- Shaping
- Scheduling
- Buffer Management
- LAGs

Shaping

The HSQ egress shaping uses the following objects:

- HSQ queue groups

  An HSQ queue group comprises eight egress queues with two WRR groups. One HS queue group is allocated to each of the following:

  - An egress SAP
  - An egress network port
  - An egress access queue group instance
  - An egress network queue group instance
  - A subscriber egress (single SLA profile instance in single HS SLA mode). Enhanced Subscriber Management (ESM) is beyond the scope of this chapter.

- Primary shapers

  In the context of this chapter, a primary shaper is allocated for each secondary shaper because it is required in the hierarchy, but it does not perform any QoS control. Primary shapers are also allocated for each subscriber egress configured with multiple SLA profile instances in extended HS SLA mode, however, ESM is beyond the scope of this chapter.

- Secondary shapers

  Secondary shapers provide an abstraction to be used for QoS control of traffic to a downstream device such as an access node. Shaping can be performed on the entire traffic or on each scheduling class within the secondary shaper.

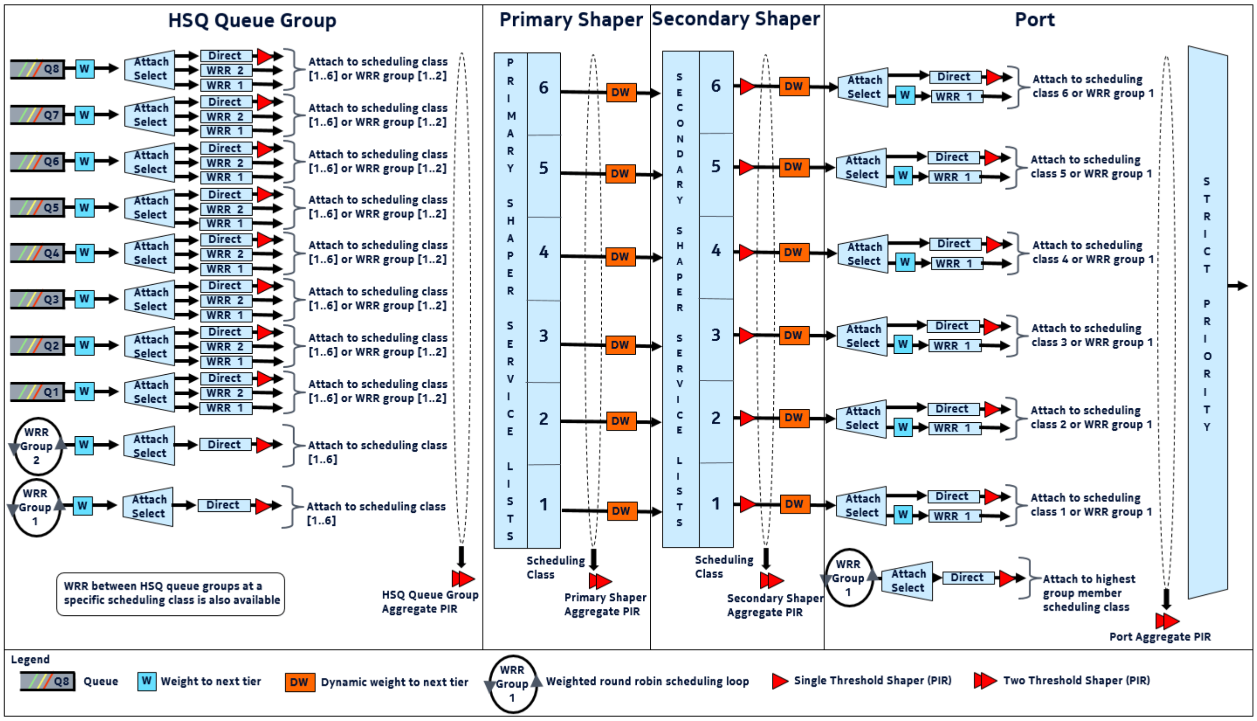
- Ports

  The traffic forwarded to each port can be shaped. In addition, traffic in each scheduling class within a port can be shaped individually or within a single WRR group.

Six scheduling classes are supported across all the preceding objects.

The egress QoS scheduling hierarchy is shown in Figure 219.

Figure 219 **Egress HSQ IOM Scheduling Hierarchy**


HSQ Queue Group
Primary Shaper
Secondary Shaper
Port
Q8
Q7
Q6
Q5
Q4
Q3
Q2
Q1
W
Attach Select
Direct
WRR 2
WRR 1
Attach to scheduling class [1..6] or WRR group [1..2]
WRR Group 2
WRR Group 1
Attach to scheduling class [1..6]
PRIMARY SHAPER SERVICE LISTS
6
5
4
3
2
1
DW
Scheduling Class
HSQ Queue Group Aggregate PIR
Primary Shaper Aggregate PIR
SECONDARY SHAPER SERVICE LISTS
Secondary Shaper Aggregate PIR
WRR 1
Attach to scheduling class 6 or WRR group 1
Attach to scheduling class 5 or WRR group 1
Attach to scheduling class 4 or WRR group 1
Attach to scheduling class 3 or WRR group 1
Attach to scheduling class 2 or WRR group 1
Attach to scheduling class 1 or WRR group 1
Attach to highest group member scheduling class
STRICT PRIORITY
Port Aggregate PIR
WRR between HSQ queue groups at a specific scheduling class is also available
Legend
Queue
Weight to next tier
Dynamic weight to next tier
Weighted round robin scheduling loop
Single Threshold Shaper (PIR)
Two Threshold Shaper (PIR)

The available egress shaping is described in detail, as follows:

- Per-queue or per-WRR group of queues
- Per-HSQ queue group aggregate
- Per-primary shaper aggregate
- Per-secondary shaper aggregate
- Per-secondary shaper per scheduling class
- Per-port aggregate
- Per-port per scheduling class

Per-Queue or Per-WRR Group of Queues

Each queue can be independently shaped by configuring its PIR and attaching it to a primary shaper scheduling class. Alternatively, it can be shaped together with other queues in the same HSQ queue group as part of a WRR group. The WRR group can have a configured rate, and also needs to be attached to a primary shaper scheduling class.

There are eight queues and two WRR groups available within an HSQ queue group, which attach to the six primary shaper scheduling classes. Only one object (queue or WRR group) per HSQ queue group can attach to a scheduling class at any time, so to make use of all queues in an HSQ queue group, at least three queues must be attached to a WRR group. Queues and WRR groups can remain unattached from a scheduling class, in which case the related queues discard all received packets.

The queue PIR is configured under the queue within a SAP egress QoS policy for services, in a network queue policy for network interfaces, or in an egress queue group template for both access and network egress queue group instances. The queue CIR is ignored when the policy is applied to an HSQ IOM. The per-WRR group PIR is configured within the same policies under the **hs-wrr-group** context. Queue and WRR group PIR use packet-based accounting (L2 rate), which can be adjusted using the queue **packet-byte-offset** parameter for SAP egress and egress queue group instances.

The attachment of a queue or WRR group to a scheduling class is configured within an **hs-attachment-policy**. A default **hs-attachment-policy** (which is not configurable) is created by the system and is applied to all SAP egress QoS policies, network queue policies, and egress queue group templates. The default policy has queues 1 to 3 attached to WRR group 1, which is attached to scheduling class 1, and queues 4 to 8 attached directly to scheduling classes 2 to 6.

When creating a new **hs-attachment-policy**, the following rules apply to the queue and WRR attachment:

- A queue must be attached to a scheduling class, or a WRR group, which is also attached to a scheduling class, so as to forward packets.
- Only one queue or WRR group can be attached to a scheduling class per HSQ queue group.
- Queues can only be attached to scheduling classes in an ascending order; for example, if queue 2 is attached to scheduling class 2, then queue 1 cannot attach to scheduling classes 3 to 6.
- The queue identifiers must be contiguous when attaching queues to a WRR group.
- Queues attached to WRR group 1 must have lower queue identifiers than those attached to WRR group 2.
- The maximum number of queues attached to a WRR group is six: six to group 1 or six to group 2, or six to a combination of groups 1 and 2.
- WRR group 2 can only be attached to a scheduling class after WRR group 1 has at least one attached queue and has been attached to a scheduling class.
- WRR group 2 must be attached to a higher scheduling class than WRR group 1.

Per-HSQ Queue Group Aggregate

A per-HSQ queue group aggregate shapes traffic forwarded by all the queues in its associated HSQ queue group to an aggregate rate. This is applicable to SAP egress queues, and to both access and network egress queue group instances. It is not applicable to network egress queues.

The per-HSQ queue group aggregate PIR is configurable as an egress aggregate rate limit applied under a SAP or a port access or network egress queue group instance. The HSQ queue group PIR uses packet-based accounting (L2 rate), which can be adjusted using the queue **packet-byte-offset** parameter for SAP egress and egress queue group instances.

When using HSQ queue groups with access or network egress queue group instances on 100G ports, the **hs-turbo** parameter can be configured under the port queue group instance to allow the corresponding HSQ queue group queues to achieve a higher throughput. The **hs-turbo** parameter is not applicable to 10G ports and so is ignored when configured under a queue group instance on a 10G port.

Per-Primary Shaper Aggregate

A primary shaper aggregate shapes the traffic forwarded by all of the HSQ queue groups connected to the primary shaper to an aggregate rate.

User-configured primary shapers are not applicable to SAP egress HSQ queue groups, network egress HSQ queue groups, or both access and network egress queue group instance HSQ queue groups. However, the hierarchy shown in Figure 219 is always conformed to, so by default these HSQ queue groups always connect to a system-created per-port default primary shaper that has its aggregate PIR rate set to the maximum rate, so as not to constrain the traffic rate at this level.

The system also instantiates a primary shaper, again with its aggregate PIR set to the maximum rate, when the first egress SAP or pseudowire SAP (PW-SAP) is associated with a secondary shaper. This primary shaper is then used by all HSQ queue groups associated with that secondary shaper. User-configured primary shaper aggregates are applicable to ESM, which is beyond the scope of this chapter.

Per-Secondary Shaper Aggregate

Secondary shapers are aimed at providing QoS control for traffic forwarded to a specific downstream device, such as an access node.

A secondary shaper aggregate shapes the traffic forwarded by all of its connected primary shapers (and HSQ queue groups). Secondary shapers are applicable to SAP egress queues, but not to network egress or to both access and network egress queue group instance HSQ queue groups. The hierarchy shown in Figure 219 is always conformed to, so by default all primary shapers (and their HSQ queue groups) always connect to a system-created per-port default secondary shaper that has its aggregate PIR rate set to the maximum rate, so as not to constrain the traffic rate at this level. Secondary shaper aggregates are also applicable to ESM, which is beyond the scope of this chapter.

Multiple HS secondary shapers can be created under the *config>port>ethernet> egress* context using the **hs-secondary-shaper** statement. The **HS secondary shaper** aggregate PIR is configured under the associated secondary shaper. A default hs-secondary-shaper is applied under each HSQ egress port with an aggregate PIR rate **max**, which can be configured if required. The secondary shaper PIR uses frame-based accounting (L1 rate) and is not affected by a queue **packet-byte-offset** parameter.

SAP egress HSQ queue groups are connected to an HS secondary shaper using the **hs-secondary-shaper** parameter under a queue override, which is configured under the SAP egress context. When the first egress SAP or PW-SAP is associated with a user-configured HS secondary shaper, the system instantiates a default primary shaper for that secondary shaper.

Per-Secondary Shaper per Scheduling Class

Each of the six scheduling classes can be individually shaped within an HS secondary shaper. The HS secondary shaper scheduling class PIR is configured under the associated secondary shaper. The default HS secondary shaper scheduling class PIRs are set to **max** and can also be modified. The secondary shaper scheduling class PIR uses frame-based accounting (L1 rate) and is not affected by a queue **packet-byte-offset** parameter.

Per-Port Aggregate

A per-port aggregate shapes the traffic forwarded by its connected secondary shapers, that is, all the traffic egressing out of the physical port. It is applicable to SAP egress, network egress, and both access and network egress queue group instance traffic. A default HS scheduler policy (which is not configurable) is applied to all ports.

A user-defined HS scheduler policy can be created in which the port aggregate PIR (**max-rate**) can be configured and the policy then applied under the *config>port> ethernet>egress* context. Only a single HS scheduler policy is supported on each port. The port aggregate PIR uses frame-based accounting (L1 rate) and is not affected by a queue *packet-byte-offset* parameter.

An alternative to configuring a per-port aggregate is to configure an **egress-rate** on the port. This provides more granular control as it is configured in kb/s (whereas the per-port aggregate is in Mb/s). The HSQ **egress-rate** is based on the Ethernet size of the packet including the IFG (Inter-Frame-Gap) and preamble.

Per-Port per Scheduling Class

Each of the six scheduling classes can also be individually shaped per port by configuring a scheduling class PIR within an HS scheduler policy. The scheduling classes can also be grouped in a single WRR group at each egress port with each class being assigned a weight within the group.

The scheduling class identifiers must be contiguous within the WRR group and the group is scheduled at the scheduling class of its highest member scheduling class. Both the scheduling class PIR and the WRR group PIR are set to **max** in the default HS scheduler policy, with the WRR group being unused. The port scheduling class PIR uses frame-based accounting (L1 rate) and is not affected by a queue **packet-byte-offset** parameter.

Scheduling

The scheduling allows a mix of strict priority and WRR. There are six scheduling classes, which are implemented from the HSQ queue group queues through the primary shaper, secondary shaper, and port. The scheduling classes are serviced in a strict priority order (scheduling class 6 having the highest priority and scheduling class 1 having the lowest priority), with WRR groups at the HSQ queue group and port levels, and a dynamic weight at the primary and secondary shaper levels.

Packet forwarding is achieved using service lists; the objects at each level are on a service list at that level if they are in a state ready to send packets, or are off the service list if they have exceeded their configured PIR together with its related burst. When a port has a scheduling opportunity, it selects the secondary shaper to be serviced next, which selects the primary shaper to be serviced next, which selects the HSQ queue group to be serviced next, which selects a queue to be serviced next, resulting in a packet from that queue being forwarded.

At the HSQ queue group level, queues can be attached to one of two WRR groups, each of which is scheduled at a single scheduling class with packets being taken from the constituent queues based on a configured queue weighting. The weight is configured using the **hs-wrr-weight** under the **queue** statement within a SAP egress QoS policy, a network queue policy, or in an egress queue group template.

Weighting is also supported between queues and WRR groups in different HSQ queue groups per-primary shaper scheduling class. This allows the capacity available at the primary shaper scheduling class to be shared in a WRR manner between the HSQ queue group queues and WRR groups attached to that scheduling class. This is configured within a SAP egress QoS policy, network queue policy, and egress queue group template, using the **hs-class-weight** parameter under the respective **queue** or **hs-wrr-group** statement.

This weighting should not be confused with the **hs-wrr-weight** parameter, which specifies the relative weights of different queues within the same HSQ queue group WRR group. This **hs-class-weight** parameter could be used to give unequal shares of the available capacity to different types of service offerings.

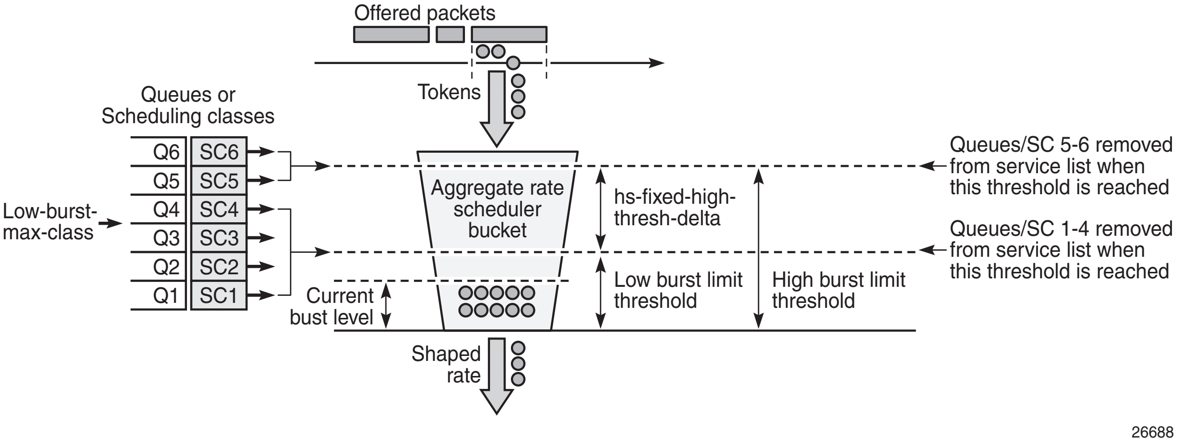
There is a single WRR group at the port level that allows multiple scheduling classes to be collapsed to a single class per port with each class in the group being assigned a weight. The weight of each scheduling class in the group is configured within the applied HS scheduler policy.

The dynamic weights at the primary and secondary shapers are managed by the system, based on the number of pending packets for each of the shapers, not on the number of attached objects in each. The more pending packets a shaper has, the higher the weight it gets. The goal is to ensure a balanced distribution of capacity between each of the primary shapers and each of the secondary shapers. For example, this allows a secondary shaper with 10 000 active HSQ queue groups to receive proportionately more scheduling opportunities than another secondary shaper with only 100 active HSQ queue groups.

The HSQ queue group and secondary shaper aggregate rates are implemented as a set of token buckets to control the aggregate rates. As packets are transmitted from each, the scheduler updates its bucket states based on the number of bytes forwarded. Two thresholds are used within each bucket to provide more granular control over this scheduling behavior: a low burst limit threshold and a high burst limit threshold.

These thresholds control when their respective queues are removed from the scheduler list, thereby allowing the queues using the high threshold to continue to forward packets even after the queues using the low threshold are no longer being serviced. This is shown in Figure 220. The **low-burst-max-class** parameter defines which queues use each of the thresholds, and is described following.

Figure 220 **HSQ Queue Group and Secondary Shaper Aggregate Scheduler Bucket**


Offered packets
Tokens
Queues or
Scheduling classes
Q6
SC6
Q5
SC5
Q4
SC4
Q3
SC3
Q2
SC2
Q1
SC1
Low-burst-max-class
Aggregate rate scheduler bucket
hs-fixed-high-thresh-delta
Current bust level
Low burst limit threshold
High burst limit threshold
Queues/SC 5-6 removed from service list when this threshold is reached
Queues/SC 1-4 removed from service list when this threshold is reached
Shaped rate
26688

Tokens representing the bytes in the packets are added to the bucket as packets are forwarded. Tokens are drained from the scheduler bucket at the configured aggregate PIR rate. If the rate at which packets are forwarded (tokens are added) exceeds the shaping rate (tokens drained), a depth of tokens builds up in the bucket. If the depth reaches the low burst limit threshold, the queues using the low threshold are removed from the scheduling list. If the depth continues to increase and reaches the high threshold, the remaining queues are removed from the scheduler list.

The low burst threshold depth is determined by the system. It is equivalent to the burst control group visitation time used by the FP egress queue scheduler. The shaping rate tokens are periodically removed from the bucket by the system by decrementing the current burst size. This period must be small enough to ensure that the resulting decrement does not cause the bucket depth to be negative, which is not permitted. Because the bucket depth cannot be negative, any potential negative decrement is lost, which equates to a loss of scheduling opportunities and the queue would underrun.

The high burst limit threshold uses a fixed increment on top of the low burst limit threshold. This fixed increment is configured under card>fp>egress using the **hs-fixed-high-thresh-delta** parameter and has a default value of 4000 bytes. It is recommended to set this parameter to a value at least two times the maximum packet size to prevent the classes using the low burst threshold from affecting those using the high burst threshold when forwarding larger packets. An insufficient burst threshold delta defeats the intended purpose of mapping classes to the high burst threshold.

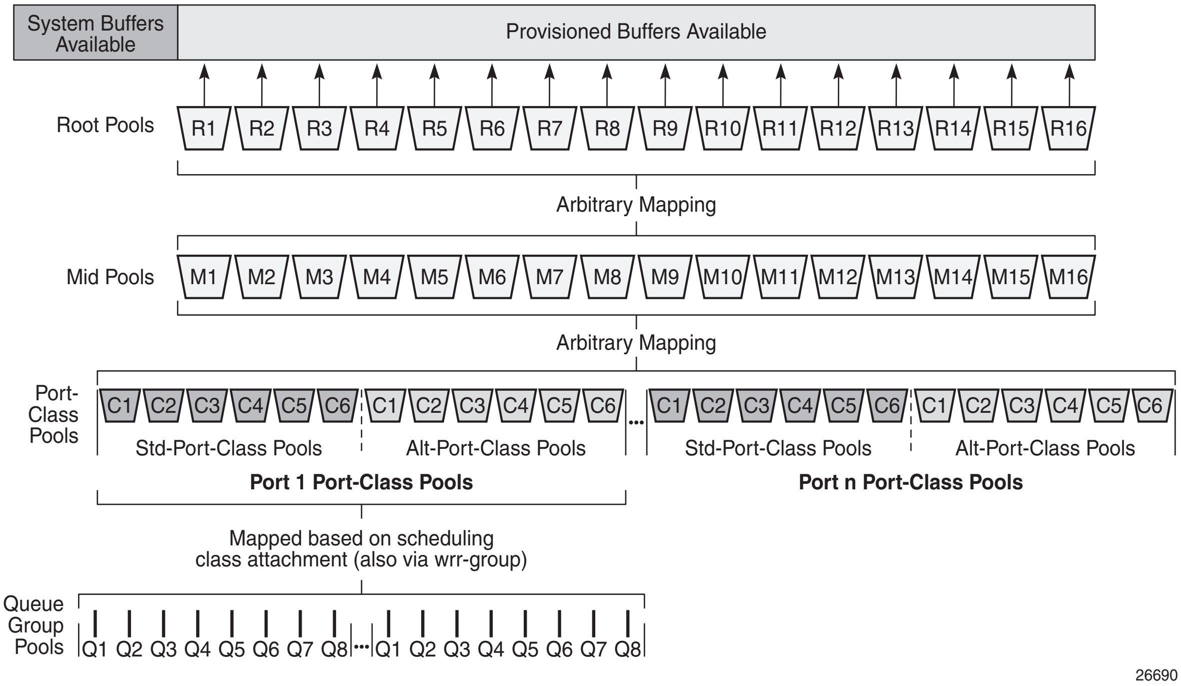
The **low-burst-max-class** parameter in the HS attachment policy (for the HSQ queue group aggregate rate) or under the secondary shaper configuration (for the secondary shaper aggregate rate) configures which queues use the low burst limit threshold and which use the high burst limit threshold. This parameter has a default max class of 6 in both contexts. As the name of the parameter implies, the specified class is the highest class that uses the low burst threshold; classes above the specified class use the high burst threshold.

Buffer Management

The HSQ supports a flexible buffer management configuration that allows both buffer isolation and buffer oversubscription for the queue buffer allocation. There are four levels to the buffer hierarchy, which are shown in Figure 221:

- Root pools
- Mid pools
- Port class pools
- Queue group queues

Figure 221 **HSQ Buffer Pool Hierarchy**


System Buffers Available
Provisioned Buffers Available
Root Pools
R1 R2 R3 R4 R5 R6 R7 R8 R9 R10 R11 R12 R13 R14 R15 R16
Arbitrary Mapping
Mid Pools
M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11 M12 M13 M14 M15 M16
Arbitrary Mapping
Port-Class Pools
C1 C2 C3 C4 C5 C6 C1 C2 C3 C4 C5 C6 ... C1 C2 C3 C4 C5 C6 C1 C2 C3 C4 C5 C6
Std-Port-Class Pools
Alt-Port-Class Pools
Std-Port-Class Pools
Alt-Port-Class Pools
Port 1 Port-Class Pools
Port n Port-Class Pools
Mapped based on scheduling class attachment (also via wrr-group)
Queue Group Pools
Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 ... Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8
26690

The total buffer allocation is divided into a system-reserved portion and a user-provisioned portion. The system buffers are allocated 5% of the total buffers in the default **hs-pool-policy**, which is applied to all HSQ IOMs under card>fp>egress. This value can be modified by creating a new **hs-pool-policy**, setting its **system-reserve** parameter, and applying the policy on an HSQ IOM. The user-provisioned portion is allocated the remainder of the available buffers, which can be configured as follows.

Root Pools

The root pools represent the total number of available buffers that can be provisioned. Up to 16 root pools can be configured, each having an allocation weight to determine its allocation of the available buffers. A root pool with an allocation weight of zero is not allocated any buffers. Root pools cannot oversubscribe the real buffers on the IOM. The use of multiple root pools provides buffer isolation between the queues using each root pool. At least one root pool (root pool 1) must be assigned buffers by having a non-zero allocation weight.

A high watermark is maintained for the buffer usage in each root pool. A slope policy is applied to each root pool to handle congestion control, the default being the *_tmnx_hs_default* slope policy. Root pools are configured per FP in an **hs-pool-policy** applied under *card>fp>egress*. Root pools 1 and 2 have an allocation weight of 75 and 25, respectively, in the default **hs-pool-policy**, with the remaining pools having a weight of 0.

Mid Pools

The mid pools are an abstract pool mapping mechanism. Each mid pool can be parented to a single **parent root pool** using its parent-root-pool parameter. Mid pools cannot be parented to a root pool without buffers and mid pools are unused if not parented to a root pool. Up to 16 mid pools are available and at least one mid pool must be parented to a root pool for its queues to buffer packets. The number of buffers in a mid pool is configured as a percentage of its parent root pool size using the **allocation-percent** parameter.

Mid pools can facilitate buffer isolation by being mapped to different root pools. Mapping multiple mid pools to the same root pool allows the buffers of that root pool to be shared by those child mid pools, and if the sum of the child mid pool allocation percent is greater than 100, then the root pool will be oversubscribed accordingly.

An oversubscription factor can also be applied to each mid pool (using the **port-bw-oversub-factor** parameter) to permit its child class pools to oversubscribe it. This does not change the size of the mid pool, but allows the mid pool size to be increased in the calculation of each of its child port class pools.

A high watermark is maintained for the buffer usage in each mid pool. A slope policy is applied to each mid pool to handle congestion control, the default being the *_tmnx_hs_default* slope policy. Mid pools are configured per FP egress in an **hs-pool-policy** applied under *card>fp>egress*. In the default **hs-pool-policy**, mid pools 1 to 4 are parented to root pool 1 with allocation percentages of 40, 35, 30, and 25; mid pools 5 and 6 are parented to root pool 2 with allocation percentages of 80 and 20; and mid pools 7 to 16 are not parented to any root pool. All mid pools have a **port-bw-oversub-factor** of 1.

Port Class Pools

Port class pools, as the name implies, are per-class pools that exist at the port level. There are two sets of port class pools per port: six standard port class pools and six alternative port class pools. The alternative set of port class pools enables additional flexibility for both buffer isolation and oversubscription by providing a simple mechanism to parent queues to different port class pools and, therefore, to different mid and root pools.

HSQ queue group queues are statically assigned to the port class pool associated with the scheduling class that they have been attached to (via a WRR group, if used): scheduling class 1 to port class pool 1, up to scheduling class 6 to port class pool 6. Port class pools are configured in an **hs-port-pool-policy**, which is applied under *config>port>ethernet>egres*s. A default **hs-port-pool-policy** in which only the standard port class pools are used is applied to all HSQ ports.

Queues can be assigned to an alternative class pool (again based on the associated scheduling class) using the **hs-alt-port-class-pool** parameter under the queue in the SAP egress QoS policy, network queue policy, or egress queue group template.

Each port class pool parents to a single mid pool using its **parent-mid-pool** parameter. Port class pools are unused if not parented to a mid pool. Each port class pool must be parented to a mid pool that is parented to a root pool for queues to buffer packets. Port class pools can facilitate buffer isolation by being parented to different mid pools that are parented to different root pools. The standard port class pools are parented to their respective mid pool (port class pool 1 to mid pool 1, up to port class pool 6 to mid pool 6) in the default **hs-port-pool-policy**, with the alternative port class pools not parented to any mid pool.

The oversubscription of port class pools in a mid pool can be achieved by configuring the **port-bw-oversub-factor** under the parent mid pool (in the **hs-pool-policy**), which is multiplied by the size of the mid pool when calculating the size of each child class pool.

A weight is configurable per port to handle the allocation of buffers to different class pools parented to the same mid pool. This is configured using the **allocation port-bw-weight** under the class pool statement, where the weight configured for a port class pool is divided by the sum of the weights of the port class pools parented to the same mid pool, to determine the proportion of the allocated buffers for that port class pool. It is also possible to configure an **explicit-percent** for a port class pool, in which case that port class pool will be allocated the configured explicit percentage of the mid pool (without any mid pool **port-bw-oversub-factor** being applied).

If there are multiple port class pools parented to the same mid pool, their buffer allocation is determined using the weighting mechanism based on the port class pool **allocation port-bw-weight** parameter. Port class pools configured with an **explicit-percent** have a weight of zero (that is, they do not participate in the weighting buffer allocation). The port class pools in the default **hs-port-pool-policy** are configured with an **allocation port-bw-weight** of 1.

The port class pools are sized dynamically to provide a fair share of a mid pool size to each of its child port class pools, based on the potential bandwidth represented by each port on which the port class pools exist. The first step is to determine the usable bandwidth of each port. The mid pool buffers are then shared between its child port class pools, based on their related port usable bandwidth. An oversubscription factor is then applied to allow the port class pools to oversubscribe their mid pool. Finally, each port mid pool buffer allocation is shared between the child port class pools on that port.

No buffers are allocated to port class pools if there are no SAPs or network interfaces configured on that port and the port is shutdown. The details of the port class pool sizing calculation are as follows (examples of each are shown in the Buffer Pools configuration section):

1. Determine each port bandwidth value.
    a. This is the minimum of the port current line rate, the port **egress-rate** limit, and the **hs-scheduler-policy max-rate** configured on the port.
    b. The port bandwidth may be further modified by the port **modify-buffer-allocation-rate egr-percentage-of-rate** command, which can increase or decrease the port bandwidth by the specified percent. This allows the port to have a higher or lower bandwidth derived weight, based on how the port is being used, instead of bandwidth alone.
2. Determine each port portion of each mid pool.

The port class pools are configured to map to the mid pools, so it is possible that not every port will have a port class pool associated with a mid pool. This requires that the system perform the relative bandwidth calculations separately per mid pool. A port without any port class pools associated with a mid pool will have a port portion of zero for that mid pool.

Per mid pool, each port portion of the mid pool size is calculated based on:

*Port_Portion = (Port_Adj_Bw / Sigma_Mid_Pool_Ports_Adj_Bw) * Mid_Pool_Size*

Where:

- *Port_Adj_Bw* is calculated in (1).
- *Sigma_Mid_Pool_Ports_Adj_Bw* is the sum of the adjusted bandwidths for all ports, with port class pools mapped to the mid pool that are not sized configured with **explicit-percent** (see (4)).
- *Mid_Pool_Size* is the mid pool parent root pool size multiplied by the mid pool allocation weight.

3. Modify the mid pool sizes by their **port-bw-oversub-factor**.

   The port bandwidth weighting mechanism allocates 100% of the mid pool size to the associated port class pools. To allow the port class pools to oversubscribe their parent mid pool, the mid pool **port-bw-oversub-factor** parameter can be used to increase the apparent size of the mid pool (this does not change the mid pool size) in the calculation in (2). This potentially provides a more efficient use of the mid pool available buffers since it is not expected that all port class pools will be using their allotted size simultaneously.

4. Determine each port class pool share of the mid pool port share.

   Multiple port class pools on the same port may be mapped to the same mid pool. This requires a mechanism to distribute the portion of the mid pool allocated to each port class pool on that port.

   Each port class pool **allocation port-bw-weight** parameter is used to determine how much of the port mid pool is given to each port class pool associated with the mid pool. A port class pool is allocated the portion of its mid pool size multiplied by its port class pool **port-bw-weight** divided by the sum of the **port-bw-weights** for all port class pools associated with that mid pool on that port.

   Alternatively, port class pools can be sized using an **explicit-percent** of the actual mid pool size (without applying the **port-bw-oversub-factor**). These class pools are assigned a **port-bw-weight** equal to zero, causing them to be excluded from the port portion distribution. It is expected (but not required) that either port bandwidth-based sizing or explicit percent-based sizing will be used, with concurrent use of both mechanisms being transitory in nature.

Whenever one of the inputs to the preceding calculations changes, the bandwidth weighted sizes for the corresponding pool class pools are recalculated.

A high watermark is maintained for the buffer usage in each port class pool. A slope policy is applied to each port class pool to handle congestion control, the default being the *_tmnx_hs_default* slope policy.

Queue Group Queues

HSQ queue group queues always operate in WRED per queue mode, supporting three WRED slopes. The total number of buffers usable by the queue is limited by the queue MBS configuration, and each packet profile type (exceed, out, in) is limited by the respective slope configuration (exceed, low, high) in the applied slope policy, if the slope is not shutdown. For a buffer to be allocated, the applicable WRED slope processing (if enabled) must accept the packet, the MBS must not be exceeded, and there must be available buffers in its parent port class pool, mid pool, and root pool.

The regular **mbs** queue parameter configuration is used within SAP egress QoS policies and egress queue group templates, using the regular defaults. In the network queue policy, the MBS is configured using the **hs-mbs** parameter, which allows a different default to be used, with its value calculated based on a percentage of one second of the queue PIR converted to bytes (the regular **mbs** parameter is ignored in the network queue policy). The queue CBS and drop tail configuration is ignored on an HSQ queue group queue.

A default slope (named *_tmnx_hs_default*) is applied to each HSQ queue, using the **policy** parameter on the **hs-wred-queue** statement within a SAP egress QoS policy, network queue policy, and egress queue group template. A user-configured regular slope policy can be applied using the same parameter and statement. The *highplus* slope and time average factor in the applied slope policy are ignored on HSQ queue group queues.

LAGs

LAGs are supported on HSQ ports. The LAG **port-type** must be set to **hs** to add an HSQ port to a LAG, at which point only HSQ ports can be added to that LAG. When an HSQ queue group is created on a LAG, an HSQ queue group is allocated on each LAG port.

LAG **access adapt-qos** modes **link** and **port-fair** are supported; **distribute** mode is not supported.

LAG **access per-fp-egr-queuing** is supported and, when configured, either **per-link-hash** or **per-service-hashing** (supported service types only) must be enabled under the LAG **access**. LAG **access per-fp-sap-instance** is supported (this requires **per-fp-egr-queuing** to be enabled).

The full configured queue MBS is applied to all the related HSQ queue group queues on the individual LAG ports.

Configuration

This section describes simple configurations using an HSQ IOM for SAP egress, network egress, and access and network egress queue group instances.

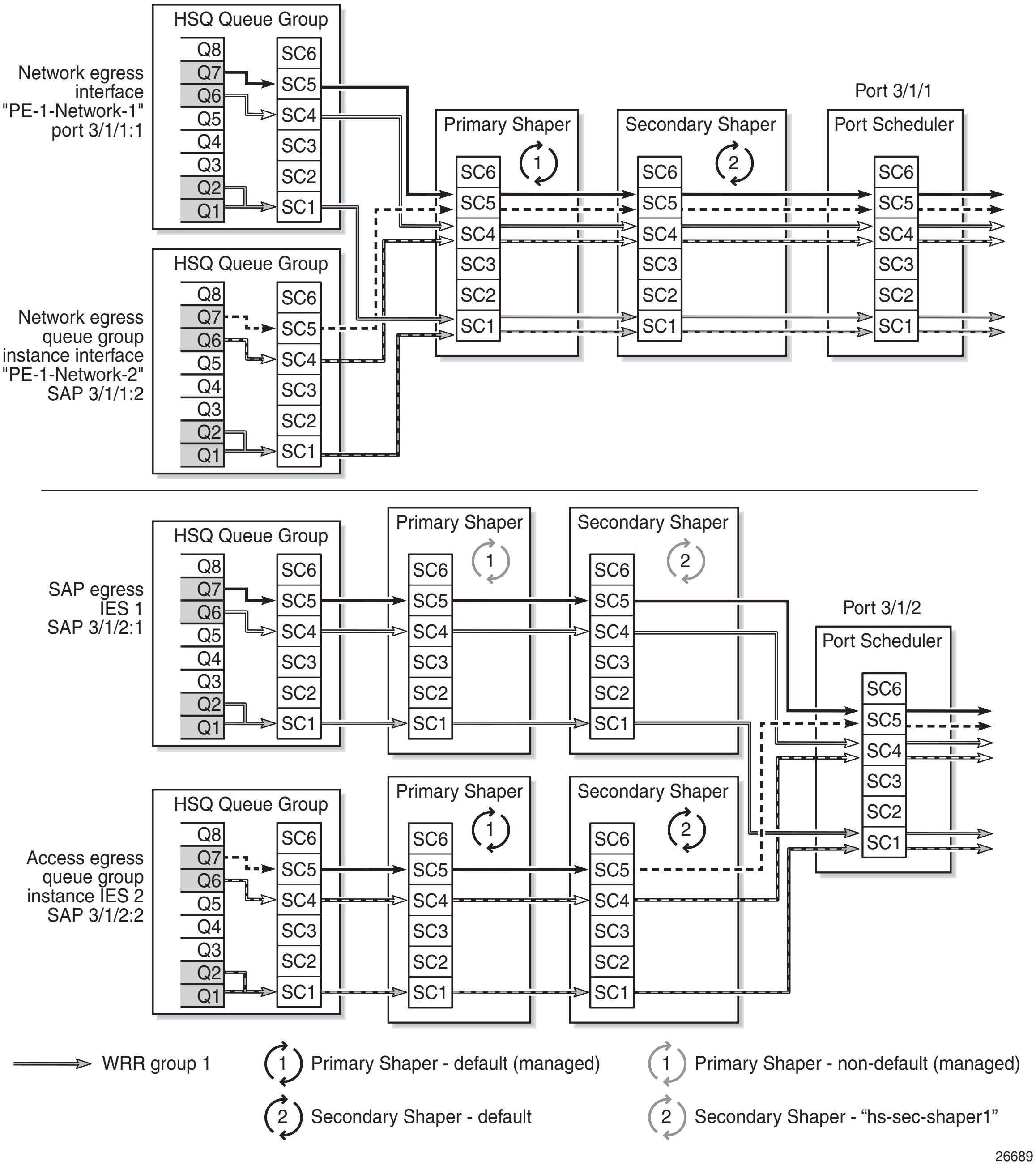
Each configuration uses the same four queues:

- Queue 7 at scheduling class 5
- Queue 6 at scheduling class 4
- Queues 1 and 2 in WRR group 1 using scheduling class 1, with queue 1 having a weight of 2 and queue 2 having a weight of 1

Scheduling class 6 has been reserved for queue 8 to be used for network protocol traffic. Queue 3 is unattached, but could at some point be added to WRR group 1. The configuration provides the future flexibility to either add queues 4 and 5 to WRR group 1, to WRR group 2, or attach them to scheduling classes 2 and 3. Although eight queues are always allocated in an HSQ queue group, only the queues to be used need to be configured.

The QoS path for the configured SAP, network interface, and access and network queue group instances with respect to their HSQ queue group, primary shaper, secondary shaper, and port scheduler, is shown in Figure 222.

Figure 222 **Configured QoS Paths**


HSQ Queue Group
Network egress interface "PE-1-Network-1" port 3/1/1:1
Network egress queue group instance interface "PE-1-Network-2" SAP 3/1/1:2
Q8
Q7
Q6
Q5
Q4
Q3
Q2
Q1
SC6
SC5
SC4
SC3
SC2
SC1
Primary Shaper
Secondary Shaper
Port 3/1/1
Port Scheduler
SAP egress IES 1 SAP 3/1/2:1
Access egress queue group instance IES 2 SAP 3/1/2:2
Port 3/1/2
1
2
WRR group 1
Primary Shaper - default (managed)
Secondary Shaper - default
Primary Shaper - non-default (managed)
Secondary Shaper - “hs-sec-shaper1”
26689

The configurations start with the generic aspects:

- Card configuration
- Buffer pools

- Shaping and scheduling
    - HSQ queue groups
    - HS secondary shapers
        - These are specific to SAP egress in the context of this chapter; however, as secondary shapers can also be used by subscribers, they are included with the generic aspects.
    - Ports

This is followed by the specific configuration related to:

- SAP egress
- Network egress
- Access and network egress queue groups

Card Configuration

An HSQ IOM is configured with the card type *iom4-e-hs* and the associated supported MDAs:

```
A:PE-1# configure card 3
A:PE-1>config>card# info
----------------------------------------------
        card-type iom4-e-hs
        mda 1
            mda-type me10-10gb-sfp+
            no shutdown
        exit
        no shutdown
----------------------------------------------
A:PE-1>config>card# exit all
A:PE-1# show card

===============================================================================
Card Summary
===============================================================================
Slot      Provisioned Type                         Admin Operational   Comments
              Equipped Type (if different)         State State
-------------------------------------------------------------------------------
2         imm-2pac-fp3                             up    up
3         iom4-e-hs                                up    up
A         cpm5                                     up    up/active
B         cpm5                                     up    up/standby
===============================================================================
A:PE-1#
```

The supported MDA types are displayed by entering a "?" after the **mda-type** parameter:

```
A:PE-1# configure card 3 mda 1 mda-type
  - mda-type <mda-type>
  - no mda-type

 <mda-type>           : me1-100gb-cfp2|me10-10gb-sfp+|me12-10/1gb-sfp+|me2-100gb-
cfp4|me2-100gb-qsfp28|me40-1gb-csfp|me6-10gb-sfp+
```

The **hs-fixed-high-thresh-delta** on card 3 fp 1 is *default*, resulting in the high burst limit threshold (which is used by queues and WRR groups attached to scheduling classes above the **low-burst-max-class**) being 4000 bytes larger than the low burst limit threshold:

```
*A:PE-1# show card 3 detail | match "HS Fixed High Threshold Delta"
    HS Fixed High Threshold Delta : default
*A:PE-1#
```

The HSQ-specific resource usage is displayed as follows:

```
*A:PE-1# tools dump resource-usage card 3 fp 1

===============================================================================
Resource Usage Information for Card Slot #3 FP #1
===============================================================================
                                                    Total  Allocated       Free
-------------------------------------------------------------------------------

                               Egress Queues |     786432        123     786309
                            Ingress Policers |     511999          1     511998
                       Ingress Policer Stats |     511967          0     511967

                 Egress HS Turbo Queue Group |         64         10         54
                       Egress HS Queue Group |      98240         36      98204
                          HS Primary Shapers +      16384         22      16362
                 HS Explicit Primary Shapers -                     0
                  HS Managed Primary Shapers -                    22
                        HS Secondary Shapers |       4096         22       4074
===============================================================================
*A:PE-1#
```

The preceding output displays the usage information of the egress queues, the ingress policers and their statistics, the HS turbo queue groups, HS queue groups, HS primary shapers (the managed primary shapers are system-created, whereas the explicit primary shapers are used for ESM), and the HS secondary shapers.

The following card commands are ignored on HSQ IOMs:

- All regular **pool** commands
- **ingress-buffer-allocation**
- **reset-on-recoverable-error**
- **virtual-scheduler-adjustment** (egress only)

The following card commands are not configurable on HSQ IOMs:

- **named-pool-mode**
- **stable-pool-sizing**
- **egress wred-queue-control**

Buffer Pools

The buffer pool configuration used in this example provides buffer isolation between traffic in queue 7, queue 6, and WRR group 1, by assigning each to a different root and mid pool. The combined queue 1 and 2 traffic share a root and mid pool.

Root and Mid Pools

The HSQ root and mid pools for an IOM are configured in an HS pool policy, which is applied under *card>fp>egress*.

An HS pool policy is configured as follows:

```
configure
    qos
        hs-pool-policy <policy-name> [create]
            description <description-string>
            mid-tier
                mid-pool <mid-pool-id>
                    allocation-percent <percent-of-parent-pool>
                    parent-root-pool <root-pool-id>
                    port-bw-oversub-factor <oversubscription-factor>
                    slope-policy <policy-name>
            root-tier
                root-pool <root-pool-id>
                    allocation-weight <pool-weight>
                    slope-policy <policy-name>
            system-reserve <percent-of-buffers>
```

Where (in order):

```
<policy-name>        : [32 chars max]
<description-string> : [80 chars max]
<mid-pool-id>        : [1..16]
<percent-of-parent*> : [0.01..100.00]
<root-pool-id>       : [1..16]
<oversubscription-*> : [1..10]
<policy-name>        : [32 chars max]
<root-pool-id>       : [1..16 | none]
<pool-weight>        : [0..100]
<policy-name>        : [32 chars max]
```

```
<percent-of-buffers> : [1.00..30.00]
```

A default HS pool policy is created by the system with the following configuration:

hs-pool-policy default

System reserve: 5%

Root pools			**Mid pools**				
Root Pool ID	**Allocation weight**	**Slope policy**	**Mid Pool ID**	**Parent mid pool**	**Allocation %**	**Port BW oversub factor**	**Slope policy**
1	75	_tmnx_hs_default	1	1	40%	1	_tmnx_hs_default
2	25	_tmnx_hs_default	2	1	35%	1	_tmnx_hs_default
3	20	_tmnx_hs_default	3	1	30%	1	_tmnx_hs_default
4-16	0		4	1	25%	1	_tmnx_hs_default
			5	2	80%	1	_tmnx_hs_default
			6	2	20%	1	_tmnx_hs_default
			7-16	None			

If a new HS pool policy is created, its initial configuration is as follows:

hs-pool-policy <new>

System reserve: 5%

Root pools			**Mid pools**				
Root Pool ID	**Allocation weight**	**Slope policy**	**Mid Pool ID**	**Parent mid pool**	**Allocation %**	**Port BW oversub factor**	**Slope policy**
1	100	_tmnx_hs_default	1-16	1	1%	1	_tmnx_hs_default
2-16							

The HS pool policy (*hs-pool-pol-1*) used for this example is shown following. Root pool 1 and mid pool 1 have been reserved for network protocol traffic and are not used. Root pool 2 and mid pool 2 are used for queue 7 traffic, root pool 3 and mid pool 3 are used for queue 6 traffic, and root pool 4 and mid pool 4 are used for queue 1 and queue 2 (WRR group 1) traffic. The buffer allocation for each pool is based on the expected traffic volumes.

Root pool 4 and mid pool 4 have a more aggressive slope policy (*hs-slope-1*) than the default HSQ slope policy (*_tmnx_hs_default*). The default HSQ slope policy is as follows (the HSQ slopes use the instantaneous queue depth so the **time-average-factor** is ignored and the *highplus* slope is also ignored):

```
*A:PE-1>config>qos# slope-policy "_tmnx_hs_default"
*A:PE-1>config>qos>slope-policy# info detail
```

```
-----------------------------------------------
            description "Default HS slope policy."
            highplus-slope
                shutdown
                start-avg 100
                max-avg 100
                max-prob 100
            exit
            high-slope
                start-avg 100
                max-avg 100
                max-prob 100
                no shutdown
            exit
            low-slope
                start-avg 90
                max-avg 90
                max-prob 100
                no shutdown
            exit
            exceed-slope
                start-avg 80
                max-avg 80
                max-prob 100
                no shutdown
            exit
            time-average-factor 7
```

The slope policy *hs-slope-1* is configured as follows:

```
A:PE-1>config>qos# slope-policy "hs-slope-1"
A:PE-1>config>qos>slope-policy# info
-----------------------------------------------
            highplus-slope
                shutdown
            exit
            high-slope
                start-avg 85
                max-avg 100
                no shutdown
            exit
            low-slope
                no shutdown
            exit
            exceed-slope
                shutdown
            exit
```

Mid pool 4 has been configured to allow a 4 times oversubscription by its child class pools.

Root pools 5 to 16 and mid pools 5 to 16 are unused.

HS pool policy *hs-pool-pol-1* is summarized as follows:

hs-pool-policy hs-pool-pol-1

System reserve: 5%

Root pools			**Mid pools**				
Root Pool ID	**Allocation weight**	**Slope policy**	**Mid Pool ID**	**Parent mid pool**	**Allocation %**	**Port BW oversub factor**	**Slope policy**
1	5	_tmnx_hs_default	1	1	100%	1	_tmnx_hs_default
2	10	_tmnx_hs_default	2	2	100%	1	_tmnx_hs_default
3	20	_tmnx_hs_default	3	3	100%	1	_tmnx_hs_default
4	65	hs-slope-1	4	4	100%	4	hs-slope-1
5-16	0	_	5-16	None			

The HS pool policy hs-pool-pol-1 is configured as follows:

```
A:PE-1>config>qos>hs-pool-policy# info
----------------------------------------------
            root-tier
                root-pool 1
                    allocation-weight 5
                exit
                root-pool 2
                    allocation-weight 10
                exit
                root-pool 3
                    allocation-weight 20
                exit
                root-pool 4
                    allocation-weight 65
                    slope-policy "hs-slope-1"
                exit
            exit
            mid-tier
                mid-pool 1
                    allocation-percent 100.00
                exit
                mid-pool 2
                    parent-root-pool 2
                    allocation-percent 100.00
                exit
                mid-pool 3
                    parent-root-pool 3
                    allocation-percent 100.00
                exit
                mid-pool 4
                    parent-root-pool 4
                    allocation-percent 100.00
                    port-bw-oversub-factor 4
                    slope-policy "hs-slope-1"
                exit
                mid-pool 5
                    parent-root-pool none
                exit
                mid-pool 6
```

```
                    parent-root-pool none
                exit
                mid-pool 7
                    parent-root-pool none
                exit
                mid-pool 8
                    parent-root-pool none
                exit
                mid-pool 9
                    parent-root-pool none
                exit
                mid-pool 10
                    parent-root-pool none
                exit
                mid-pool 11
                    parent-root-pool none
                exit
                mid-pool 12
                    parent-root-pool none
                exit
                mid-pool 13
                    parent-root-pool none
                exit
                mid-pool 14
                    parent-root-pool none
                exit
                mid-pool 15
                    parent-root-pool none
                exit
                mid-pool 16
                    parent-root-pool none
                exit
            exit
```

The HS pool policy is shown as follows:

```
*A:PE-1# show qos hs-pool-policy "hs-pool-pol-1"

===============================================================================
HS Pool Policy Information
===============================================================================
Policy Name             : hs-pool-pol-1
Description             : (Not Specified)
System Reserve          : 5.00

-------------------------------------------------------------------------------
Root Pool Information
-------------------------------------------------------------------------------
Pool Id                 : 1              Allocation Weight   : 5
Slope Policy            : _tmnx_hs_default

Pool Id                 : 2              Allocation Weight   : 10
Slope Policy            : _tmnx_hs_default

Pool Id                 : 3              Allocation Weight   : 20
Slope Policy            : _tmnx_hs_default

Pool Id                 : 4              Allocation Weight   : 65
```

```
Slope Policy            : hs-slope-1

Pool Id                 : 5             Allocation Weight   : 0
Slope Policy            : _tmnx_hs_default

Pool Id                 : 6             Allocation Weight   : 0
Slope Policy            : _tmnx_hs_default

Pool Id                 : 7             Allocation Weight   : 0
Slope Policy            : _tmnx_hs_default

Pool Id                 : 8             Allocation Weight   : 0
Slope Policy            : _tmnx_hs_default

Pool Id                 : 9             Allocation Weight   : 0
Slope Policy            : _tmnx_hs_default

Pool Id                 : 10            Allocation Weight   : 0
Slope Policy            : _tmnx_hs_default

Pool Id                 : 11            Allocation Weight   : 0
Slope Policy            : _tmnx_hs_default

Pool Id                 : 12            Allocation Weight   : 0
Slope Policy            : _tmnx_hs_default

Pool Id                 : 13            Allocation Weight   : 0
Slope Policy            : _tmnx_hs_default
Pool Id                 : 14            Alloc
ation Weight   : 0
Slope Policy            : _tmnx_hs_default
Pool Id                 : 15            Allocation Weight   : 0
Slope Policy            : _tmnx_hs_default

Pool Id                 : 16            Allocation Weight   : 0
Slope Policy            : _tmnx_hs_default

-------------------------------------------------------------------------------

-------------------------------------------------------------------------------
Mid Pool Information
-------------------------------------------------------------------------------
Pool Id                 : 1             Allocation Percent  : 100.00
Port BW Oversub Factor  : 1             Parent Root Pool    : 1
Slope Policy            : _tmnx_hs_default

Pool Id                 : 2             Allocation Percent  : 100.00
Port BW Oversub Factor  : 1             Parent Root Pool    : 2
Slope Policy            : _tmnx_hs_default

Pool Id                 : 3             Allocation Percent  : 100.00
Port BW Oversub Factor  : 1             Parent Root Pool    : 3
Slope Policy            : _tmnx_hs_default

Pool Id                 : 4             Allocation Percent  : 100.00
Port BW Oversub Factor  : 4             Parent Root Pool    : 4
Slope Policy            : hs-slope-1

Pool Id                 : 5             Allocation Percent  : 1.00
```

```
Port BW Oversub Factor  : 1               Parent Root Pool    : 0
Slope Policy            : _tmnx_hs_default

Pool Id                 : 6               Allocation Percent  : 1.00
Port BW Oversub Factor  : 1               Parent Root Pool    : 0
Slope Policy            : _tmnx_hs_default

Pool Id                 : 7               Allocation Percent  : 1.00
Port BW Oversub Factor  : 1               Parent Root Pool    : 0
Slope Policy            : _tmnx_hs_default

Pool Id                 : 8               Allocation Percent  : 1.00
Port BW Oversub Factor  : 1               Parent Root Pool    : 0
Slope Policy            : _tmnx_hs_default

Pool Id                 : 9               Allocation Percent  : 1.00
Port BW Oversub Factor  : 1               Parent Root Pool    : 0
Slope Policy            : _tmnx_hs_default

Pool Id                 : 10              Allocation Percent  : 1.00
Port BW Oversub Factor  : 1               Parent Root Pool    : 0
Slope Policy            : _tmnx_hs_default

Pool Id                 : 11              Allocation Percent  : 1.00
Port BW Oversub Factor  : 1               Parent Root Pool    : 0
Slope Policy            : _tmnx_hs_default

Pool Id                 : 12              Allocation Percent  : 1.00
Port BW Oversub Factor  : 1               Parent Root Pool    : 0
Slope Policy            : _tmnx_hs_default

Pool Id                 : 13              Allocation Percent  : 1.00
Port BW Oversub Factor  : 1               Parent Root Pool    : 0
Slope Policy            : _tmnx_hs_default

Pool Id                 : 14              Allocation Percent  : 1.00
Port BW Oversub Factor  : 1               Parent Root Pool    : 0
Slope Policy            : _tmnx_hs_default

Pool Id                 : 15              Allocation Percent  : 1.00
Port BW Oversub Factor  : 1               Parent Root Pool    : 0
Slope Policy            : _tmnx_hs_default

Pool Id                 : 16              Allocation Percent  : 1.00
Port BW Oversub Factor  : 1               Parent Root Pool    : 0
Slope Policy            : _tmnx_hs_default

-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

This HS pool policy is configured for the HSQ IOM as follows:

```
A:PE-1# configure card 3
A:PE-1>config>card# info
----------------------------------------------
        card-type iom4-e-hs
        fp 1
```

```
            egress
                hs-pool-policy "hs-pool-pol-1"
            exit
        exit
        no shutdown
----------------------------------------------
A:PE-1>config>card#
```

The association of this HS pool policy is shown as follows:

```
*A:PE-1# show qos hs-pool-policy "hs-pool-pol-1" association

===============================================================================
HS Pool Policy Information
===============================================================================
Policy Name             : hs-pool-pol-1
Description             : (Not Specified)
System Reserve          : 5.00

-------------------------------------------------------------------------------
Card Forwarding Plane (FP) Associations
-------------------------------------------------------------------------------
Card                FP
-------------------------------------------------------------------------------
3                   1
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

The resulting system and user-provisioned pool information is shown following. This output shows the total buffer allocation, number of allocated buffers, available buffer allocation, and buffer high watermarks for the system pools and user-provisioned pools. The output shows the hierarchy of the root and mid pools, with their applied slope policy and the related instantaneous slope drop probabilities (as a percentage):

```
*A:PE-1# show hs-pools 3 fp 1 egress

===============================================================================
HS Pools Card Forwarding Plane Information
===============================================================================
Card               : 3                    FP                 : 1

-------------------------------------------------------------------------------
System Pool Information
-------------------------------------------------------------------------------

Total Buffers      : 209412 KB            Allocated          : 0 KB
Available          : 209412 KB            High Water Mark    : 0 KB

-------------------------------------------------------------------------------
Buffer Pool Hierarchy Information
-------------------------------------------------------------------------------

Root Pool : 1
|  Total               : 198942 KB   Allocated         : 0 KB
|  Available           : 198942 KB   High Water Mark   : 0 KB
|  Hi-Slope Drop Prob  : 0           Lo-Slope Drop Prob: 0
|  Excd-Slope Drop Prob: 0
|  Hs Slope Policy     : _tmnx_hs_default
|
|--- Mid Pool : 1
|     | Total               : 198942 KB   Allocated         : 0 KB
|     | Available           : 198942 KB   High Water Mark   : 0 KB
|     | Hi-Slope Drop Prob  : 0           Lo-Slope Drop Prob: 0
|     | Excd-Slope Drop Prob: 0
|     | Hs Slope Policy     : _tmnx_hs_default
|     |
Root Pool : 2
|  Total               : 397886 KB   Allocated         : 0 KB
|  Available           : 397886 KB   High Water Mark   : 0 KB
|  Hi-Slope Drop Prob  : 0           Lo-Slope Drop Prob: 0
|  Excd-Slope Drop Prob: 0
|  Hs Slope Policy     : _tmnx_hs_default
|
|--- Mid Pool : 2
|     | Total               : 397886 KB   Allocated         : 0 KB
|     | Available           : 397886 KB   High Water Mark   : 0 KB
|     | Hi-Slope Drop Prob  : 0           Lo-Slope Drop Prob: 0
|     | Excd-Slope Drop Prob: 0
|     | Hs Slope Policy     : _tmnx_hs_default
|     |
Root Pool : 3
|  Total               : 795772 KB   Allocated         : 0 KB
|  Available           : 795772 KB   High Water Mark   : 0 KB
|  Hi-Slope Drop Prob  : 0           Lo-Slope Drop Prob: 0
|  Excd-Slope Drop Prob: 0
|  Hs Slope Policy     : _tmnx_hs_default
|
|--- Mid Pool : 3
|     | Total               : 795772 KB   Allocated         : 0 KB
|     | Available           : 795772 KB   High Water Mark   : 0 KB
|     | Hi-Slope Drop Prob  : 0           Lo-Slope Drop Prob: 0
|     | Excd-Slope Drop Prob: 0
|     | Hs Slope Policy     : _tmnx_hs_default
|     |
Root Pool : 4
|  Total               : 2586262 KB  Allocated         : 0 KB
|  Available           : 2586262 KB  High Water Mark   : 0 KB
|  Hi-Slope Drop Prob  : 0           Lo-Slope Drop Prob: 0
|  Excd-Slope Drop Prob: 0
|  Hs Slope Policy     : hs-slope-1
|
|--- Mid Pool : 4
|     | Total               : 2586262 KB  Allocated         : 0 KB
|     | Available           : 2586262 KB  High Water Mark   : 0 KB
|     | Hi-Slope Drop Prob  : 0           Lo-Slope Drop Prob: 0
|     | Excd-Slope Drop Prob: 0
|     | Hs Slope Policy     : hs-slope-1
|     |
```

Port Class Pools

The HSQ port class pools for a port are configured in an HS port pool policy, which is applied under *config>port>ethernet>egress*.

An HS port pool policy is configured as follows:

```
configure
    qos
        hs-port-pool-policy <policy-name> [create]
            description <description-string>
            std-port-class-pools
                class-pool <std-class-pool-id>
                    allocation explicit-percent <percent-of-parent-pool>
                    allocation port-bw-weight <pool-weight>
                    parent-mid-pool <mid-pool-id>
                    slope-policy <policy-name>
            alt-port-class-pools
                class-pool <alt-class-pool-id>
                    allocation explicit-percent <percent-of-parent-pool>
                    allocation port-bw-weight <pool-weight>
                    parent-mid-pool <mid-pool-id>
                    slope-policy <policy-name>
```

Where (in order):

```
<policy-name>         : [32 chars max]
<description-string> : [80 chars max]
<std-class-pool-id>  : [1..6]
<percent-of-parent*> : [0.01..100.00]
<pool-weight>         : [1..100]
<mid-pool-id>         : [1..16 | none]
<policy-name>         : [32 chars max]
<alt-class-pool-id>  : [1..6]
<percent-of-parent*> : [0.01..100.00]
<pool-weight>         : [1..100]
<mid-pool-id>         : [1..16 | none]
<policy-name>         : [32 chars max]
```

A default HS pool policy is created by the system with the following configuration:

hs-port-pool-policy default

Standard port class pools				**Alternative port class pools**			
Class Pool ID	**Parent mid pool**	**Allocation port bw weight**	**Slope policy**	**Class Pool ID**	**Parent mid pool**	**Allocation port bw weight**	**Slope policy**
1	1	1	_tmnx_hs_default	1-6	None		
2	2	1	_tmnx_hs_default				
3	3	1	_tmnx_hs_default				
4	4	1	_tmnx_hs_default				

hs-port-pool-policy default

Standard port class pools				Alternative port class pools			
Class Pool ID	**Parent mid pool**	**Allocation port bw weight**	**Slope policy**	**Class Pool ID**	**Parent mid pool**	**Allocation port bw weight**	**Slope policy**
5	5	1	_tmnx_hs_default				
6	6	1	_tmnx_hs_default				

Newly created HS port pool policies have the following parameters:

hs-port-pool-policy <new>

Standard port class pools				Alternative port class pools			
Class Pool ID	**Parent mid pool**	**Allocation port bw weight**	**Slope policy**	**Class Pool ID**	**Parent mid pool**	**Allocation port bw weight**	**Slope policy**
1-6	1	1	_tmnx_hs_default	1-6	None		

The HS port pool policy used for traffic in this example is shown following. Only port class pools 1, 4, 5, and 6 are used, which are parented to mid pools 4, 3, 2, and 1, respectively. As scheduling classes 2 and 3 are unused, their associated standard port class pools are not parented to a mid pool. The alternative port class pools are also unused, so are not parented to a mid pool. Standard port class pools 4 to 6 use the default HSQ slope policy with standard port class pool 1 using slope policy *hs-slope-1*.

hs-port-pool-policy hs-port-pool-pol-1

Standard port class pools				Alternative port class pools			
Class Pool ID	**Parent mid pool**	**Allocation port bw weight**	**Slope policy**	**Class Pool ID**	**Parent mid pool**	**Allocation port bw weight**	**Slope policy**
1	4	1	hs-slope-1	1-6	None		
2-3	None						
4	3	1	_tmnx_hs_default				
5	2	1	_tmnx_hs_default				
6	1	1	_tmnx_hs_default				

The HS port pool policy is configured as follows:

```
*A:PE-1>config>qos# hs-port-pool-policy "hs-port-pool-pol-1"
*A:PE-1>config>qos>hs-port-pool-policy# info
----------------------------------------------
            std-port-class-pools
                class-pool 1
                parent-mid-pool 4
                slope-policy "hs-slope-1"
            exit
            class-pool 2
                parent-mid-pool none
            exit
            class-pool 3
                parent-mid-pool none
            exit
            class-pool 4
                parent-mid-pool 3
            exit
            class-pool 5
                parent-mid-pool 2
            exit
        exit
```

The HS port pool policy is shown as follows:

```
*A:PE-1# show qos hs-port-pool-policy "hs-port-pool-pol-1"

===============================================================================
HS Port Pool Policy Information
===============================================================================
Policy Name             : hs-port-pool-pol-1
Description             : (Not Specified)

-------------------------------------------------------------------------------
Standard Port Class Pool Information
-------------------------------------------------------------------------------
Class Id                : 1             Parent Mid Pool     : 4
Alloc Port BW Weight    : 1             Alloc Explicit Prcnt: 0.00
Slope Policy            : hs-slope-1

Class Id                : 2             Parent Mid Pool     : 0
Alloc Port BW Weight    : 1             Alloc Explicit Prcnt: 0.00
Slope Policy            : _tmnx_hs_default

Class Id                : 3             Parent Mid Pool     : 0
Alloc Port BW Weight    : 1             Alloc Explicit Prcnt: 0.00
Slope Policy            : _tmnx_hs_default

Class Id                : 4             Parent Mid Pool     : 3
Alloc Port BW Weight    : 1             Alloc Explicit Prcnt: 0.00
Slope Policy            : _tmnx_hs_default

Class Id                : 5             Parent Mid Pool     : 2
Alloc Port BW Weight    : 1             Alloc Explicit Prcnt: 0.00
Slope Policy            : _tmnx_hs_default

Class Id                : 6             Parent Mid Pool     : 1
Alloc Port BW Weight    : 1             Alloc Explicit Prcnt: 0.00
Slope Policy            : _tmnx_hs_default

-------------------------------------------------------------------------------

-------------------------------------------------------------------------------
Alternate Port Class Pool Information
```

```
-------------------------------------------------------------------------------
Class Id                : 1             Parent Mid Pool     : 0
Alloc Port BW Weight    : 1             Alloc Explicit Prcnt: 0.00
Slope Policy            : _tmnx_hs_default

Class Id                : 2             Parent Mid Pool     : 0
Alloc Port BW Weight    : 1             Alloc Explicit Prcnt: 0.00
Slope Policy            : _tmnx_hs_default

Class Id                : 3             Parent Mid Pool     : 0
Alloc Port BW Weight    : 1             Alloc Explicit Prcnt: 0.00
Slope Policy            : _tmnx_hs_default

Class Id                : 4             Parent Mid Pool     : 0
Alloc Port BW Weight    : 1             Alloc Explicit Prcnt: 0.00
Slope Policy            : _tmnx_hs_default

Class Id                : 5             Parent Mid Pool     : 0
Alloc Port BW Weight    : 1             Alloc Explicit Prcnt: 0.00
Slope Policy            : _tmnx_hs_default

Class Id                : 6             Parent Mid Pool     : 0
Alloc Port BW Weight    : 1             Alloc Explicit Prcnt: 0.00
Slope Policy            : _tmnx_hs_default

-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

The preceding HS port pool policy applied to ports 3/1/1 and 3/1/2 is shown as follows:

```
*A:PE-1# show qos hs-port-pool-policy "hs-port-pool-pol-1" association

===============================================================================
HS Port Pool Policy Information
===============================================================================
Policy Name             : hs-port-pool-pol-1
Description             : (Not Specified)

-------------------------------------------------------------------------------
Port Ethernet Egress Associations
-------------------------------------------------------------------------------
3/1/1
3/1/2
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

The remaining ports (3/1/[3..10]) are unused in this example, so their port class pools are not parented, by applying the following HS port pool policy to each:

hs-port-pool-policy no-class-pools

Standard port class pools				**Alternative port class pools**			
Class Pool ID	**Parent mid pool**	**Allocation port bw weight**	**Slope policy**	**Class Pool ID**	**Parent mid pool**	**Allocation port bw weight**	**Slope policy**
1-6	None			1-6	None		

```
*A:PE-1>config>qos# hs-port-pool-policy "no-class-pools"
*A:PE-1>config>qos>hs-port-pool-policy# info
----------------------------------------------
            std-port-class-pools
                class-pool 1
                    parent-mid-pool none
                exit
                class-pool 2
                    parent-mid-pool none
                exit
                class-pool 3
                    parent-mid-pool none
                exit
                class-pool 4
                    parent-mid-pool none
                exit
                class-pool 5
                    parent-mid-pool none
                exit
                class-pool 6
                    parent-mid-pool none
                exit
            exit
```

The HS port pool policies applied to the ports on card 3 are configured as follows:

```
configure
    port 3/1/1
        ethernet
            egress
                hs-port-pool-policy "hs-port-pool-pol-1"
    port 3/1/2
        ethernet
            egress
                hs-port-pool-policy "hs-port-pool-pol-1"
    port 3/1/[3..10]
        ethernet
            egress
                hs-port-pool-policy "no-class-pools"
```

The pools created on port 3/1/1, after applying the preceding HS pool policy and HS port pool policies, are shown following. The root and mid pools in this output are the same as in the **show hs-pools 3 fp 1 egress** output; these pools are configured per FP so are the same for all ports. The additional information shows the details of the port class pools on this port and to which mid pools they are parented:

```
*A:PE-1# show hs-pools port 3/1/1 egress

===============================================================================
HS Pools Port Information
===============================================================================
Port               : 3/1/1

-------------------------------------------------------------------------------
System Pool Information
-------------------------------------------------------------------------------

Total Buffers      : 209412 KB          Allocated          : 0 KB
Available          : 209412 KB          High Water Mark    : 0 KB

-------------------------------------------------------------------------------
Buffer Pool Hierarchy Information
-------------------------------------------------------------------------------

Root Pool : 1
|  Total               : 198942 KB   Allocated         : 0 KB
|  Available           : 198942 KB   High Water Mark   : 0 KB
|  Hi-Slope Drop Prob  : 0           Lo-Slope Drop Prob: 0
|  Excd-Slope Drop Prob: 0
|  Hs Slope Policy     : _tmnx_hs_default
|
|--- Mid Pool : 1
|     | Total               : 198942 KB   Allocated         : 0 KB
|     | Available           : 198942 KB   High Water Mark   : 0 KB
|     | Hi-Slope Drop Prob  : 0           Lo-Slope Drop Prob: 0
|     | Excd-Slope Drop Prob: 0
|     | Hs Slope Policy     : _tmnx_hs_default
|     |
|     |--- Std Port Class Pool : 6
|     |      Total               : 99470 KB    Allocated         : 0 KB
|     |      Available           : 99470 KB    High Water Mark   : 0 KB
|     |      Hi-Slope Drop Prob  : 0           Lo-Slope Drop Prob: 0
|     |      Excd-Slope Drop Prob: 0
|     |      Hs Slope Policy     : _tmnx_hs_default
|     |
Root Pool : 2
|  Total               : 397886 KB   Allocated         : 0 KB
|  Available           : 397886 KB   High Water Mark   : 0 KB
|  Hi-Slope Drop Prob  : 0           Lo-Slope Drop Prob: 0
|  Excd-Slope Drop Prob: 0
|  Hs Slope Policy     : _tmnx_hs_default
|
|--- Mid Pool : 2
|     | Total               : 397886 KB   Allocated         : 0 KB
|     | Available           : 397886 KB   High Water Mark   : 0 KB
|     | Hi-Slope Drop Prob  : 0           Lo-Slope Drop Prob: 0
|     | Excd-Slope Drop Prob: 0
|     | Hs Slope Policy     : _tmnx_hs_default
|     |
|     |--- Std Port Class Pool : 5
|     |      Total               : 198942 KB   Allocated         : 0 KB
|     |      Available           : 198942 KB   High Water Mark   : 0 KB
|     |      Hi-Slope Drop Prob  : 0           Lo-Slope Drop Prob: 0
|     |      Excd-Slope Drop Prob: 0
|     |      Hs Slope Policy     : _tmnx_hs_default
|     |
Root Pool : 3
|  Total               : 795772 KB  Allocated         : 0 KB
|  Available           : 795772 KB  High Water Mark   : 0 KB
|  Hi-Slope Drop Prob  : 0          Lo-Slope Drop Prob: 0
|  Excd-Slope Drop Prob: 0
|  Hs Slope Policy     : _tmnx_hs_default
|
|--- Mid Pool : 3
|     | Total               : 795772 KB  Allocated         : 0 KB
|     | Available           : 795772 KB  High Water Mark   : 0 KB
|     | Hi-Slope Drop Prob  : 0          Lo-Slope Drop Prob: 0
|     | Excd-Slope Drop Prob: 0
|     | Hs Slope Policy     : _tmnx_hs_default
|     |
|     |--- Std Port Class Pool : 4
|     |      Total               : 397886 KB  Allocated         : 0 KB
|     |      Available           : 397886 KB  High Water Mark   : 0 KB
|     |      Hi-Slope Drop Prob  : 0          Lo-Slope Drop Prob: 0
|     |      Excd-Slope Drop Prob: 0
|     |      Hs Slope Policy     : _tmnx_hs_default
|     |
Root Pool : 4
|  Total               : 2586262 KB Allocated         : 0 KB
|  Available           : 2586262 KB High Water Mark   : 0 KB
|  Hi-Slope Drop Prob  : 0          Lo-Slope Drop Prob: 0
|  Excd-Slope Drop Prob: 0
|  Hs Slope Policy     : hs-slope-1
|
|--- Mid Pool : 4
|     | Total               : 2586262 KB Allocated         : 0 KB
|     | Available           : 2586262 KB High Water Mark   : 0 KB
|     | Hi-Slope Drop Prob  : 0          Lo-Slope Drop Prob: 0
|     | Excd-Slope Drop Prob: 0
|     | Hs Slope Policy     : hs-slope-1
|     |
|     |--- Std Port Class Pool : 1
|     |      Total               : 4194302 KB Allocated         : 0 KB
|     |      Available           : 4194302 KB High Water Mark   : 0 KB
|     |      Hi-Slope Drop Prob  : 0          Lo-Slope Drop Prob: 0
|     |      Excd-Slope Drop Prob: 0
|     |      Hs Slope Policy     : hs-slope-1
|     |
```

Note: The maximum size of a port class pool is capped here at 4194302 kbytes, as shown for the port class pool parented to mid/root pool 4.

The following show commands are available to display the pool configuration and queue details:

```
show hs-pools port <port-id> egress
show hs-pools port <port-id> egress network-queues
show hs-pools port <port-id> egress queue-group <queue-group-name>
                             [instance <instance-id>]
show hs-pools port <port-id> egress sap <sap-id>
show hs-pools port <port-id> egress subscriber <sub-ident-string>
```

The root and mid pool information in each command is the same as in the command **show hs-pools <card-slot-number> fp <forwarding-plane> egress** but the commands also show the port class pool information on the specified port, together with the queue information for that port.

Pool Sizing and Oversubscription

The logic to size the root, mid, and port class pools is described in the Buffer Management section.

Root pools are sized using their **allocation-weight** parameter, which is divided by the sum of all root pool **allocation-weights** to give the portion of the total user-provisioned buffers allocated to the root pool.

The mid pools are sized using their **allocation-percent** parameter, which is a percentage of their parent root pool size.

The port class pools size calculation has more factors. The following output shows the effect of the different sizing parameters on the port class pools size. The steps are:

- An HS pool policy and HS port pool policies are applied to an HSQ IOM and its ports to configure one root pool with one child mid pool that has one child standard port class pool on port 3/1/1. Each pool has the same size matching the total available buffers:

```
*A:PE-1# show hs-pools port 3/1/1 egress |
              match "Root Pool : 1" post-lines 26 | match "Pool" post-lines 1
Root Pool : 1
|  Total               : 3978866 KB  Allocated         : 0 KB
|--- Mid Pool : 1
|     | Total               : 3978866 KB  Allocated         : 0 KB
|     |--- Std Port Class Pool : 1
|     |      Total               : 3978866 KB  Allocated         : 0 KB
```

- A standard port class pool is added to port 3/1/2, causing the mid pool size to be shared between port class pool 1 on port 3/1/1 and 3/1/2:

```
*A:PE-1# configure port 3/1/2 ethernet egress
                               hs-port-pool-policy "hs-port-pool-policy-test"
*A:PE-1# show hs-pools port 3/1/1 egress |
              match "Root Pool : 1" post-lines 26 | match "Pool" post-lines 1
Root Pool : 1
|  Total               : 3978866 KB  Allocated         : 0 KB
|--- Mid Pool : 1
|     | Total               : 3978866 KB  Allocated         : 0 KB
|     |--- Std Port Class Pool : 1
|     |      Total               : 1989432 KB  Allocated         : 0 KB
*A:PE-1# show hs-pools port 3/1/2 egress | match "Root Pool : 1" post-
lines 26 | match "Pool" post-lines 1
Root Pool : 1
|  Total               : 3978866 KB  Allocated         : 0 KB
|--- Mid Pool : 1
|     | Total               : 3978866 KB  Allocated         : 0 KB
|     |--- Std Port Class Pool : 1
|     |      Total               : 1989432 KB  Allocated         : 0 KB
```

- The egress rate is set on port 3/1/1 (which is a 10 Gb/s port) to 5 Gb/s. This reduces the size of the port class pools on port 3/1/1 to one-third of the mid pool size and increases the size of the port class pools on port 3/1/2 to two-thirds of the mid pool size, after which the egress rate is removed:

```
*A:PE-1# configure port 3/1/1 ethernet egress-rate 5000000
*A:PE-1# show hs-pools port 3/1/1 egress |
              match "Root Pool : 1" post-lines 26 | match "Pool" post-lines 1
Root Pool : 1
|  Total               : 3978866 KB  Allocated         : 0 KB
|--- Mid Pool : 1
|     | Total               : 3978866 KB  Allocated         : 0 KB
|     |--- Std Port Class Pool : 1
|     |      Total               : 1326288 KB  Allocated         : 0 KB
*A:PE-1# show hs-pools port 3/1/2 egress |
              match "Root Pool : 1" post-lines 26 | match "Pool" post-lines 1
Root Pool : 1
|  Total               : 3978866 KB  Allocated         : 0 KB
|--- Mid Pool : 1
|     | Total               : 3978866 KB  Allocated         : 0 KB
|     |--- Std Port Class Pool : 1
|     |      Total               : 2652576 KB  Allocated         : 0 KB
*A:PE-1# configure port 3/1/1 ethernet no egress-rate
```

- The egr-percentage-of-rate is set to 200% on port 3/1/1 to increase its port class pool to two-thirds of the mid pools size and reduce the port class pool on port 3/1/2 to one-third of the mid pools size, after which the egr-percentage-of-rate is removed:

```
*A:PE-1# configure port 3/1/1 modify-buffer-allocation-rate egr-percentage-of-
rate 200
*A:PE-1# show hs-pools port 3/1/1 egress |
              match "Root Pool : 1" post-lines 26 | match "Pool" post-lines 1
Root Pool : 1
|  Total               : 3978866 KB  Allocated         : 0 KB
|--- Mid Pool : 1
|     | Total               : 3978866 KB  Allocated         : 0 KB
|     |--- Std Port Class Pool : 1
|     |      Total               : 2652576 KB  Allocated         : 0 KB
*A:PE-1# show hs-pools port 3/1/2 egress |
              match "Root Pool : 1" post-lines 26 | match "Pool" post-lines 1
Root Pool : 1
|  Total               : 3978866 KB  Allocated         : 0 KB
|--- Mid Pool : 1
|     | Total               : 3978866 KB  Allocated         : 0 KB
|     |--- Std Port Class Pool : 1
|     |      Total               : 1326288 KB  Allocated         : 0 KB
```

```
*A:PE-1# configure port 3/1/1 modify-buffer-allocation-rate
                              no egr-percentage-of-rate
```

- Standard port class pool 2 is parented to the mid pool with a **port-bw-weight** set to 2. The **port-bw-weight** of port class pool 1 is the default of 1. This causes this port mid pool size to be shared in a 1:2 ratio between port class pool 1 and 2 on both ports 3/1/1 and 3/1/2 (only port 3/1/1 is shown). The total port class pool size is shown in the second step preceding, that is, 1989432 kbytes.

```
*A:PE-1# configure qos hs-port-pool-policy "hs-port-pool-policy-test"
*A:PE-1>config>qos>hs-port-pool-policy# std-port-class-pools
                                              class-pool 2 parent-mid-pool 1
*A:PE-1>config>qos>hs-port-pool-policy# std-port-class-pools
                                    class-pool 2 allocation port-bw-weight 2
*A:PE-1>config>qos>hs-port-pool-policy# exit all
*A:PE-1# show hs-pools port 3/1/1 egress |
              match "Root Pool : 1" post-lines 47 | match "Pool" post-lines 1
Root Pool : 1
|  Total               : 3978866 KB  Allocated         : 0 KB
|--- Mid Pool : 1
|     | Total               : 3978866 KB  Allocated         : 0 KB
|     |--- Std Port Class Pool : 1
|     |      Total               : 663144 KB   Allocated         : 0 KB
|     |--- Std Port Class Pool : 2
|     |      Total               : 1326288 KB  Allocated         : 0 KB
```

- "The two port class pools 1 and 2 on port 3/1/1 are modified to use an **explicit-percent** of 40% and 60%, respectively:

```
*A:PE-1# configure qos hs-port-pool-policy "hs-port-pool-policy-test"
*A:PE-1>config>qos>hs-port-pool-policy# std-port-class-pools
                                  class-pool 1 allocation explicit-percent 40
*A:PE-1>config>qos>hs-port-pool-policy# std-port-class-pools
                                  class-pool 2 allocation explicit-percent 60
*A:PE-1>config>qos>hs-port-pool-policy# exit all
*A:PE-1# show hs-pools port 3/1/1 egress |
              match "Root Pool : 1" post-lines 47 | match "Pool" post-lines 1
Root Pool : 1
|  Total               : 3978866 KB  Allocated         : 0 KB
|--- Mid Pool : 1
|     | Total               : 3978866 KB  Allocated         : 0 KB
|     |--- Std Port Class Pool : 1
|     |      Total               : 1591546 KB  Allocated         : 0 KB
|     |--- Std Port Class Pool : 2
|     |      Total               : 2387318 KB  Allocated         : 0 KB
```

To assist with sizing the buffer pools, each pool has a high watermark, which can be displayed using the **show hs-pools** command and cleared using the following commands:

```
clear card <slot-number> fp <[1..2]> hs-pool high-water-mark
clear card <slot-number> fp <[1..2]> hs-pool high-water-mark mid-pool <[1..16]>
clear card <slot-number> fp <[1..2]> hs-pool high-water-mark root-pool <[1..16]>
clear card <slot-number> fp <[1..2]> hs-pool high-water-mark system
clear port <port-id> hs-pool high-water-mark
                                      { [standard <1..6>] | [alternate <1..6>]
```

The HSQ ingress and non-HSQ egress line cards support a **stable-pool-sizing** command under *card>fp*, which avoids pool sizes changing when MDAs and ports are configured.

An equivalent effect can be achieved at the egress of an HSQ by creating two sets of root pools and two sets of mid pools in the **hs-pool-policy** applied to the IOM under *card>fp>egress*. The first set of mid pools parent to the first set of root pools and the second set of mid pools parent to the second set of root pools. Then create two **hs-port-pool-policies**: one applied to the ports on the first MDA with its port class pools parented to the first set of mid pools and the other applied to the ports on the second MDA with its port class pools parented to the second set of mid pools. This provides deterministic pool sizing independent of MDAs being inserted or removed.

Further control at the port class level can be obtained by using port class pool **explicit-percent** based sizing to eliminate the effect of changing port states, including bandwidth changes.

The root pools cannot oversubscribe the user-provisioned buffers, but the mid pools can oversubscribe their root pool and the port class pools can oversubscribe their mid pool.

The following configuration and output shows the oversubscription possibilities.

The default HS pool policy is applied to the HSQ IOM, so root pools 1 and 2 are allocated 75% and 25% of the user-provisioned buffers:

```
*A:PE-1>config>qos>hs-pool-policy# info detail |
                      match expression " root-pool 1$| root-pool 2$" post-lines 1
                root-pool 1
                    allocation-weight 75
                root-pool 2
                    allocation-weight 25

*A:PE-1# show hs-pools 3 fp 1 egress | match "Root Pool" post-lines 1
Root Pool : 1
|  Total               : 2984148 KB  Allocated         : 0 KB
Root Pool : 2
|  Total               : 994716 KB   Allocated         : 0 KB
```

If a new HS pool policy is applied to this IOM, only one root pool is allocated buffers with an allocation weight of 100:

```
*A:PE-1>config>qos>hs-pool-policy$ info detail |
                      match expression " root-pool 1$" post-lines 1
                root-pool 1
                    allocation-weight 100

*A:PE-1# show hs-pools port 3/1/1 egress | match "Root Pool : 1" post-lines 1
Root Pool : 1
|  Total               : 3978866 KB  Allocated         : 0 KB
```

Root pool 16 is configured with the same allocation weight of 100, causing the two root pools to share the available user-provisioned buffers:

```
*A:PE-1>config>qos>hs-pool-policy# info detail |
                      match expression " root-pool 1$| root-pool 16$" post-lines 1
                root-pool 1
                    allocation-weight 100
                root-pool 16
                    allocation-weight 100

*A:PE-1# show hs-pools port 3/1/1 egress | match "Root Pool : 1" post-lines 1
Root Pool : 1
|  Total               : 1989432 KB  Allocated         : 0 KB
Root Pool : 16
|  Total               : 1989432 KB  Allocated         : 0 KB
```

Mid pool 16 is parented to root pool 16 with an allocation percent of 100, so it has the same number of allocated buffers as root pool 16:

```
*A:PE-1>config>qos>hs-pool-policy# info detail |
      match expression " root-pool 16$| mid-pool 16$" post-lines 2 |
      match invert-match slope
                root-pool 16
                    allocation-weight 100
                mid-pool 16
                    parent-root-pool 16
                    allocation-percent 100.00

*A:PE-1# show hs-pools port 3/1/1 egress |
               match "Root Pool : 16" post-lines 19 | match "Pool" post-lines 1
Root Pool : 16
|  Total               : 1989432 KB  Allocated         : 0 KB
|--- Mid Pool : 16
|     | Total               : 1989432 KB  Allocated         : 0 KB
```

Mid pool 15 is also parented to root pool 16 with an allocation percent of 100, causing both mid pools 15 and 16 to have the same number of allocated buffers as root pool 16; therefore, the root pool is oversubscribed two times:

```
*A:PE-1>config>qos>hs-pool-policy# info detail |
     match expression " root-pool 16$| mid-pool 15$| mid-pool 16$" post lines 2 |
     match invert-match slope
                root-pool 16
                    allocation-weight 100
                mid-pool 15
                    parent-root-pool 16
                    allocation-percent 100.00
                mid-pool 16
                    parent-root-pool 16
                    allocation-percent 100.00

*A:PE-1# show hs-pools port 3/1/1 egress |
              match "Root Pool : 16" post-lines 19 |
              match "Pool" post-lines 1
Root Pool : 16
|  Total               : 1989432 KB  Allocated         : 0 KB
```

```
|--- Mid Pool : 15
|     | Total               : 1989432 KB  Allocated         : 0 KB
|--- Mid Pool : 16
|     | Total               : 1989432 KB  Allocated         : 0 KB
```

An HS port pool policy is applied to port 3/1/1 with standard port class pool 1 parented to mid pool 16 with an **allocation port-bw-weight** of 1. This port class pool has the same number of allocated buffers as mid pool 16:

```
*A:PE-1>config>qos>hs-port-pool-policy>std-port-class-pools# info
----------------------------------------------
                class-pool 1
                    parent-mid-pool 16
                exit

*A:PE-1# show hs-pools port 3/1/1 egress |
                  match "Root Pool : 16" post-lines 26 | match "Pool" post-lines 1
Root Pool : 16
|  Total               : 1989432 KB  Allocated         : 0 KB
|--- Mid Pool : 15
|     | Total               : 1989432 KB  Allocated         : 0 KB
|--- Mid Pool : 16
|     | Total               : 1989432 KB  Allocated         : 0 KB
|     |--- Std Port Class Pool : 1
|     |      Total               : 1989432 KB  Allocated         : 0 KB
```

The **port-bw-oversub-factor** is set to 2 for mid pool 16 in the HS pool policy, after which the size of mid pool 16 does not change. However, its apparent size for the calculation of its port class pool doubles, which causes the size of port class pool 1 to be twice that of mid pool 16, thereby oversubscribing it:

```
*A:PE-1>config>qos>hs-pool-policy# info
----------------------------------------------
            root-tier
                root-pool 16
                    allocation-weight 100
                exit
            exit
            mid-tier
                mid-pool 15
                    parent-root-pool 16
                    allocation-percent 100.00
                exit
                mid-pool 16
                    parent-root-pool 16
                    allocation-percent 100.00
                    port-bw-oversub-factor 2
                exit
            exit

*A:PE-1# show hs-pools port 3/1/1 egress |
              match "Root Pool : 16" post-lines 26 |
              match "Pool" post-lines 1
Root Pool : 16
|  Total               : 1989432 KB  Allocated         : 0 KB
|--- Mid Pool : 15
```

```
|     | Total               : 1989432 KB  Allocated         : 0 KB
|--- Mid Pool : 16
|     | Total               : 1989432 KB  Allocated         : 0 KB
|     |--- Std Port Class Pool : 1
|     |      Total               : 3978864 KB  Allocated         : 0 KB
```

A second standard port class pool, pool 2, on port 3/1/1 is parented to mid pool 16. The two port class pools share the buffer allocation equivalent to two times that of mid pool 16:

```
*A:PE-1>config>qos>hs-port-pool-policy>std-port-class-pools# info
----------------------------------------------
                class-pool 1
                    parent-mid-pool 16
                exit
                class-pool 2
                    parent-mid-pool 16
                exit
*A:PE-1# show hs-pools port 3/1/1 egress |
              match "Root Pool : 16" post-lines 33  |
              match "Pool" post-lines 1
Root Pool : 16
|  Total               : 1989432 KB  Allocated         : 0 KB
|--- Mid Pool : 15
|     | Total               : 1989432 KB  Allocated         : 0 KB
|--- Mid Pool : 16
|     | Total               : 1989432 KB  Allocated         : 0 KB
|     |--- Std Port Class Pool : 1
|     |      Total               : 1989432 KB  Allocated         : 0 KB
|     |--- Std Port Class Pool : 2
|     |      Total               : 1989432 KB  Allocated         : 0 KB
```

The proportion of buffers available to the port class pools can be modified by configuring their **allocation port-bw-weight**. If the **allocation port-bw-weight** of port class pool 2 is set to 2, the port class pools will be allocated buffers in a 2:1 ratio:

```
*A:PE-1>config>qos>hs-port-pool-policy>std-port-class-pools# info
----------------------------------------------
                class-pool 1
                    parent-mid-pool 16
                exit
                class-pool 2
                    parent-mid-pool 16
                    allocation port-bw-weight 2
                exit

*A:PE-1# show hs-pools port 3/1/1 egress |
              match "Root Pool : 16" post-lines 33 |
              match "Pool" post-lines 1
Root Pool : 16
|  Total               : 1989432 KB  Allocated         : 0 KB
|--- Mid Pool : 15
|     | Total               : 1989432 KB  Allocated         : 0 KB
|--- Mid Pool : 16
|     | Total               : 1989432 KB  Allocated         : 0 KB
|     |--- Std Port Class Pool : 1
|     |      Total              : 1326288 KB  Allocated         : 0 KB
|     |--- Std Port Class Pool : 2
|     |      Total              : 2652576 KB  Allocated         : 0 KB
```

A third standard port class pool, pool 3, is parented to mid pool 16 with an **allocation port-bw-weight** of 2. Port class pools 1, 2, and 3 share the oversubscribed mid pool 16 size in a ratio of 1:2:2:

```
*A:PE-1>config>qos>hs-port-pool-policy>std-port-class-pools# info
----------------------------------------------
                class-pool 1
                    parent-mid-pool 16
                exit
                class-pool 2
                    parent-mid-pool 16
                    allocation port-bw-weight 2
                exit
                class-pool 3
                    parent-mid-pool 16
                    allocation port-bw-weight 2
                exit
```

```
*A:PE-1# show hs-pools port 3/1/1 egress |
              match "Root Pool : 16" post-lines 40 |
              match "Pool" post-lines 1
Root Pool : 16
|  Total               : 1989432 KB  Allocated         : 0 KB
|--- Mid Pool : 15
|     | Total               : 1989432 KB  Allocated         : 0 KB
|--- Mid Pool : 16
|     | Total               : 1989432 KB  Allocated         : 0 KB
|     |--- Std Port Class Pool : 1
|     |      Total              : 795772 KB   Allocated         : 0 KB
|     |--- Std Port Class Pool : 2
|     |      Total              : 1591544 KB  Allocated         : 0 KB
|     |--- Std Port Class Pool : 3
|     |      Total              : 1591544 KB  Allocated         : 0 KB
```

Standard port class pool 1, 2, and 3 are now configured with an allocation **explicit-percent** of 80%, 60%, and 60% respectively. This allocates these percentages of mid pool 16 real size to port class pools 1, 2, and 3, which oversubscribed the mid pool by 100%. The mid pool 16 oversubscription factor is not applied.

```
*A:PE-1>config>qos>hs-port-pool-policy>std-port-class-pools# info
----------------------------------------------
                class-pool 1
                    parent-mid-pool 16
                    allocation explicit-percent 80.00
                exit
                class-pool 2
                    parent-mid-pool 16
                    allocation explicit-percent 60.00
                exit
                class-pool 3
                    parent-mid-pool 16
                    allocation explicit-percent 60.00
```

```
               exit

*A:PE-1# show hs-pools port 3/1/1 egress |
              match "Root Pool : 16" post-lines 40 |
              match "Pool" post-lines 1
Root Pool : 16
|  Total               : 1989432 KB  Allocated         : 0 KB
|--- Mid Pool : 15
|     | Total               : 1989432 KB  Allocated         : 0 KB
|--- Mid Pool : 16
|     | Total               : 1989432 KB  Allocated         : 0 KB
|     |--- Std Port Class Pool : 1
|     |      Total               : 1591544 KB  Allocated         : 0 KB
|     |--- Std Port Class Pool : 2
|     |      Total               : 1193658 KB  Allocated         : 0 KB
|     |--- Std Port Class Pool : 3
|     |      Total               : 1193658 KB  Allocated         : 0 KB
```

Shaping and Scheduling

HSQ Queue Groups

Each configuration uses the same four queues:

- Queue 7 at scheduling class 5
- Queue 6 at scheduling class 4
- Queues 1 and 2 in WRR group at scheduling class 1

The **low-burst-max-class** is configured to be class 1, to match the scheduling class of WRR group 1. This results in queues 1 and 2 being subject to the low burst limit threshold.

HSQ queue group queues can be attached to scheduling classes or to WRR groups, which can then be attached to a scheduling class. The eight queues and two WRR groups in an HSQ queue group can also be unattached (not attached to any scheduling class, or for queues, to any WRR group). This is configured using an **hs-attachment-policy**:

```
configure
    qos
        hs-attachment-policy <policy-name> [create]
            description <description-string>
            low-burst-max-class <class>
            queue <queue-id> sched-class <class-id>
            queue <queue-id> unattached
            queue <queue-id> wrr-group <wrr-group-id>
            wrr-group <group-id> sched-class <class-id>
```

Where (in order):

```
<policy-name>        : [32 chars max]
<description-string> : [80 chars max]
<class>              : [1..6]
<queue-id>           : [1..8]
<class-id>           : [1..6]
<wrr-group-id>       : [1..2]
<group-id>           : [1..2]
<class-id>           : [1..6]
```

When a queue or WRR group is unattached, the related queues discard all received packets.

When a queue is attached to a WRR group, the weight of that queue within the group is configured under the queue in the SAP egress QoS policy, network queue policy, and egress queue group template.

When a queue is attached to a WRR group, the following queue parameters are ignored. The corresponding configuration is applied to the entire WRR group in the SAP egress QoS policy for services, network queue policy for network interfaces, and egress queue group templates for both access and network egress queue group instances:

- **adaptation-rule**
- **hs-class-weight**
- **percent-rate**
- **rate**

A default **hs-attachment-policy** is created by the system and applied by default to all SAP egress QoS policy for services, to all network queue policy for network interfaces, and to all egress queue group templates for both access and network egress queue group instances. The default policy is not configurable.

```
*A:PE-1>config>qos# hs-attachment-policy "default"
*A:PE-1>config>qos>hs-attachment-policy# info detail
----------------------------------------------
            no description
            low-burst-max-class 6
            queue 1 wrr-group 1
            queue 2 wrr-group 1
            queue 3 wrr-group 1
            queue 4 sched-class 2
            queue 5 sched-class 3
            queue 6 sched-class 4
            queue 7 sched-class 5
            queue 8 sched-class 6
            wrr-group 1 sched-class 1
            wrr-group 2 unattached
```

Note: Queues 1, 2, and 3 are attached by default to WRR group 1, so their configured rates are ignored.

To use a user-defined policy, a new HS attachment policy must be created and applied to the appropriate SAP egress QoS policy, network queue policy, or egress queue group template. A newly created policy has all queues and WRR groups unattached and the **low-burst-max-class** set to 6:

```
*A:PE-1# configure qos hs-attachment-policy hs-att-policy-new create
*A:PE-1>config>qos>hs-attachment-policy$ info detail
----------------------------------------------
            no description
            low-burst-max-class 6
            queue 1 unattached
            queue 2 unattached
            queue 3 unattached
            queue 4 unattached
            queue 5 unattached
            queue 6 unattached
            queue 7 unattached
            queue 8 unattached
            wrr-group 1 unattached
            wrr-group 2 unattached
```

The hs-attachment-policy used for this example is as follows (queue 8 is reserved to be attached to scheduling class 6 for network protocol traffic, but is not used in this example):

```
*A:PE-1>config>qos# hs-attachment-policy "hs-att-pol-1"
*A:PE-1>config>qos>hs-attachment-policy# info detail
----------------------------------------------
            no description
            low-burst-max-class 1
            queue 1 wrr-group 1
            queue 2 wrr-group 1
            queue 3 unattached
            queue 4 unattached
            queue 5 unattached
            queue 6 sched-class 4
            queue 7 sched-class 5
            queue 8 sched-class 6
            wrr-group 1 sched-class 1
            wrr-group 2 unattached
```

An HS attachment policy is shown as follows:

```
*A:PE-1# show qos hs-attachment-policy "hs-att-pol-1"

===============================================================================
HS Attachment Policy Information
===============================================================================
Policy Name             : hs-att-pol-1
```

```
Description             : (Not Specified)
Low Burst Max Class     : 1

-------------------------------------------------------------------------------
Queue               Scheduling Class              WRR Group
-------------------------------------------------------------------------------
1                   (Not-Applicable)              1
2                   (Not-Applicable)              1
3                   unattached                    unattached
4                   unattached                    unattached
5                   unattached                    unattached
6                   4                             (Not-Applicable)
7                   5                             (Not-Applicable)
8                   6                             (Not-Applicable)

-------------------------------------------------------------------------------
WRR Group           Scheduling Class
-------------------------------------------------------------------------------
1                   1
2                   unattached
===============================================================================
*A:PE-1#
```

It is also possible to show the associations for each policy:

```
*A:PE-1# show qos hs-attachment-policy "hs-att-pol-1" association

===============================================================================
HS Attachment Policy Information
===============================================================================
Policy Name             : hs-att-pol-1
Description             : (Not Specified)
Low Burst Max Class     : 1

-------------------------------------------------------------------------------
Associations
-------------------------------------------------------------------------------
Network-Queue Policy
-------------------------------------------------------------------------------
10

Sap-Egress Policy
-------------------------------------------------------------------------------
10
20

Egress Queue-Group Templates
-------------------------------------------------------------------------------
queue-group-1

-------------------------------------------------------------------------------
===============================================================================
```

HS Secondary Shapers

HS secondary shapers are only applicable to SAP egress in the context of this chapter (not to network egress or egress access and network queue group instances). However, because secondary shapers can also be used by subscribers, they are included with the generic configuration aspects.

Secondary shapers are aimed at providing QoS control for traffic forwarded to a specific downstream device, such as an access node. Multiple HS secondary shapers can be configured on an egress port.

An aggregate rate and per-scheduling class rates are configurable for each secondary shaper. A **low-burst-max-class** parameter is also available to provide granular control over the scheduling behavior of which queues (via a WRR group, if used) use the low burst limit threshold and which use the high burst limit threshold.

Secondary shapers are configured on each port under the *config>port>ethernet>egress* context:

```
configure
    port <port-id>
        ethernet
            egress
                hs-secondary-shaper <secondary-shaper-name>
                    description <description-string>
                    aggregate
                        low-burst-max-class <class>
                        rate <rate>
                    class <class-number>
                        rate <rate>
```

Where (in order):

```
<port-id>            : slot/mda/port
<secondary-shaper-*> : [32 chars max]
<description-string> : [80 chars max]
<class>              : [1..6]
<class-number>       : [1..6]
<rate>               : [1..100000000|max] Kbps
<rate>               : [1..100000000|max] Kbps
```

A default HS secondary shaper is applied to all egress HSQ ports with the rates set to **max** and the **low-burst-max-class** set to 6. It is possible to modify the configuration of the default HS secondary shaper.

```
*A:PE-1>config>port>ethernet>egress# hs-secondary-shaper "default"
*A:PE-1>config>port>ethernet>egress>hs-sec-shaper# info detail
----------------------------------------------
                    no description
                    aggregate
                        rate max
                        low-burst-max-class 6
                    exit
                    class 1
                        rate max
                    exit
                    class 2
                        rate max
                    exit
                    class 3
                        rate max
                    exit
                    class 4
                        rate max
                    exit
                    class 5
                        rate max
                    exit
                    class 6
                        rate max
                    exit
```

An HS secondary shaper is configured on port 3/1/2 with a rate of 100 Mb/s for scheduling class 1 and a **low-burst-max-class** set to 1:

```
*A:PE-1>config>port>ethernet>egress# hs-secondary-shaper "hs-sec-shaper-1"
*A:PE-1>config>port>ethernet>egress>hs-sec-shaper# info detail
----------------------------------------------
                    no description
                    aggregate
                        rate max
                        low-burst-max-class 1
                    exit
                    class 1
                        rate 100000
                    exit
                    class 2
                        rate max
                    exit
                    class 3
                        rate max
                    exit
                    class 4
                        rate max
                    exit
                    class 5
                        rate max
                    exit
                    class 6
                        rate max
                    exit
```

This is shown, in this case with its associations, as follows:

```
*A:PE-1# show port 3/1/2 hs-secondary-shaper "hs-sec-shaper-1" associations

===============================================================================
Ethernet Port 3/1/2 Egress HS Secondary Shaper Information
```

```
===============================================================================
Policy Name        : hs-sec-shaper-1
Description        : (Not Specified)
Rate               : max
Low Burst Max Class: 1

-------------------------------------------------------------------------------
Class                                   Rate
-------------------------------------------------------------------------------
1                                       100000 Kbps
2                                       max
3                                       max
4                                       max
5                                       max
6                                       max
-------------------------------------------------------------------------------

-------------------------------------------------------------------------------
Service Associations
-------------------------------------------------------------------------------
Service ID                Service Type                 SAP
-------------------------------------------------------------------------------
1                         IES                          3/1/2:1
-------------------------------------------------------------------------------

-------------------------------------------------------------------------------
Subscriber Associations
-------------------------------------------------------------------------------
Subscriber ID
-------------------------------------------------------------------------------
No Subscriber Associations Found.
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

All HS secondary shapers on a port can be shown using the same command, but omitting the shaper name and following parameter.

It is also possible to show the forwarding statistics related to an HS secondary shaper:

```
*A:PE-1# show port 3/1/2 hs-secondary-shaper "hs-sec-shaper-1" statistics

===============================================================================
Ethernet Port 3/1/2 Egress HS Secondary Shaper Information
===============================================================================
Policy Name        : hs-sec-shaper-1

------------------------------------------------------------------------
Statistics Information
------------------------------------------------------------------------

------------------------------------------------------------------------
                        Packets                 Octets
Class 1
    Forwarded         : 4000                    592000
```

```
Class 2
    Forwarded         : 0                       0

Class 3
    Forwarded         : 0                       0

Class 4
    Forwarded         : 1000                    148000

Class 5
    Forwarded         : 1000                    148000

Class 6
    Forwarded         : 0                       0

Aggregate
    Forwarded         : 6000                    888000

------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

These statistics are cleared using the following command:

```
clear port 3/1/2 hs-secondary-shaper "hs-sec-shaper-1" statistics
```

The SAP egress queues in an HSQ queue group can be associated with a secondary shaper by configuring an egress queue override under the SAP and specifying the name of the secondary shaper. In addition, when using PW-SAPs, an HS secondary shaper can be applied to the egress of a PW port to control the PW-SAP egress traffic over that PW. See the SAP Egress section for configuration details.

If the user configures an HS secondary shaper on a port, the system instantiates a default primary shaper for that secondary shaper (which is used by all HSQ queue groups sending traffic to the secondary shaper) when the first egress SAP or PW-SAP is associated with that HS secondary shaper.

The current traffic rates through the secondary shapers on a port are shown, as follows, where the default interval is 1 second:

```
show qos hs-scheduler-hierarchy port <port-id>
              [hs-secondary-shaper <shaper-name>] [interval <time-in-seconds>]
show qos hs-scheduler-hierarchy port <port-id>
              [interval <time-in-seconds>] hs-secondary-shapers
```

The output shows the current aggregate traffic rate and the current traffic rate for each scheduling class:

```
*A:PE-1# show qos hs-scheduler-hierarchy port 3/1/2
                      hs-secondary-shaper "hs-sec-shaper-1"

===============================================================================
Hs Scheduler Hierarchy Information
```

```
===============================================================================
Hs Sched Policy Name         : default

Port Max-Rate : 137 Mbps
Hs-Sec-Shaper:hs-sec-shaper-1 Agg-Rate : 57516 Kbps

Scheduler Priority 6
  Scheduler Class 6  Rate : 0 Mbps
    Hs-Sec-Shaper:hs-sec-shaper-1 Class 6 Rate : 0 Kbps

Scheduler Priority 5
  Scheduler Class 5  Rate : 22 Mbps
    Hs-Sec-Shaper:hs-sec-shaper-1 Class 5 Rate : 11422 Kbps

Scheduler Priority 4
  Scheduler Class 4  Rate : 45 Mbps
    Hs-Sec-Shaper:hs-sec-shaper-1 Class 4 Rate : 22797 Kbps

Scheduler Priority 3
  Scheduler Class 3  Rate : 0 Mbps
    Hs-Sec-Shaper:hs-sec-shaper-1 Class 3 Rate : 0 Kbps

Scheduler Priority 2
  Scheduler Class 2  Rate : 0 Mbps
    Hs-Sec-Shaper:hs-sec-shaper-1 Class 2 Rate : 0 Kbps

Scheduler Priority 1
  Scheduler Class 1  Rate : 69 Mbps
    Hs-Sec-Shaper:hs-sec-shaper-1 Class 1 Rate : 23296 Kbps
===============================================================================
*A:PE-1#
```

Ports

A single HS scheduler policy can be applied to an egress HSQ port to configure an aggregate rate (**max-rate**) and per-scheduling class rates on that port. In addition, contiguous scheduling classes can be configured with weights in a WRR group, which can also be configured with a rate. The WRR group is scheduled at the scheduling class of its highest member scheduling class. The **max-rate** caps the scheduling class and group rates if its rate is lower.

The rates configured within an HS scheduler policy are applicable to all types of traffic (SAP egress, network egress, and egress queue group instances) exiting that port.

An HS scheduler policy is configured as follows:

```
configure
    qos
        hs-scheduler-policy <policy-name> [create]
            description <description-string>
            group <group-id> rate <rate>
            max-rate <rate>
            scheduling-class <class-id> group <group-id> [weight <weight-in-group>]
            scheduling-class <class-id> rate <rate>
```

Where (in order):

```
<policy-name>        : [32 chars max]
<description-string> : [80 chars max]
<group-id>           : [1]
<rate>               : [1..100000|max] Mbps
<rate>               : [1..100000|max] Mbps
<class-id>           : [1..6]
<group-id>           : [1]
<weight-in-group>    : [1..127]
<rate>               : [1..100000|max] Mbps
```

Note: The rates configured in an HS scheduler policy are in Mb/s (not kb/s).

A default HS scheduler policy is applied to all egress HSQ ports and all its rates are set to **max**. It is not possible to modify the default HS scheduler policy.

```
*A:PE-1>config>qos# hs-scheduler-policy "default"
*A:PE-1>config>qos>hs-scheduler-policy# info detail
----------------------------------------------
            description "Default hs scheduler QoS policy"
            max-rate max
            group 1 rate max
            scheduling-class 1 rate max
            scheduling-class 2 rate max
            scheduling-class 3 rate max
            scheduling-class 4 rate max
            scheduling-class 5 rate max
            scheduling-class 6 rate max
```

A newly created HS scheduler policy has the same configuration as the default policy.

An HS scheduler policy with a rate of 5 Gb/s for scheduling class 1 is applied to port 3/1/1:

```
*A:PE-1# configure qos
*A:PE-1>config>qos# hs-scheduler-policy "hs-sched-pol-1"
*A:PE-1>config>qos>hs-scheduler-policy# info detail
----------------------------------------------
            no description
            max-rate max
            group 1 rate max
            scheduling-class 1 rate 5000
            scheduling-class 2 rate max
            scheduling-class 3 rate max
            scheduling-class 4 rate max
            scheduling-class 5 rate max
            scheduling-class 6 rate max
----------------------------------------------
*A:PE-1>config>qos>hs-scheduler-policy# exit all
*A:PE-1# configure port 3/1/1 ethernet egress hs-scheduler-policy "hs-sched-pol-1"
*A:PE-1#
```

The preceding policy is shown as follows:

```
*A:PE-1# show qos hs-scheduler-policy "hs-sched-pol-1"

===============================================================================
HS Scheduler Policy Information
===============================================================================
Policy Name             : hs-sched-pol-1
Description             : (Not Specified)
Max Rate                : max

-------------------------------------------------------------------------------
Scheduling Class    Rate                Group               Weight in Group
-------------------------------------------------------------------------------
1                   5000 Mbps           0                   1
2                   max                 0                   1
3                   max                 0                   1
4                   max                 0                   1
5                   max                 0                   1
6                   max                 0                   1
-------------------------------------------------------------------------------
Group               Rate
-------------------------------------------------------------------------------
1                   max
===============================================================================
*A:PE-1#
```

The ports associated with this policy are as follows:

```
*A:PE-1# show qos hs-scheduler-policy "hs-sched-pol-1" association

===============================================================================
HS Scheduler Policy Information
===============================================================================
Policy Name             : hs-sched-pol-1
Description             : (Not Specified)
Max Rate                : max

-------------------------------------------------------------------------------
Port Ethernet Egress Associations
-------------------------------------------------------------------------------
3/1/1
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

The current traffic rates through the port aggregate, scheduling class, and WRR group shapers are shown, as follows, where the default interval is 1 second:

```
show qos hs-scheduler-hierarchy port <port-id> [interval <time-in-seconds>]
          queue-group <queue-group-name> instance <instance-id> {access|network}
show qos hs-scheduler-hierarchy sap <sap-id> egress [interval <time-in-seconds>]
show qos hs-scheduler-hierarchy subscriber <sub-ident> egress
                                                     [interval <time-in-seconds>]
```

The output shows the current aggregate traffic rate and the current traffic rates for each scheduling class:

```
*A:PE-1# show qos hs-scheduler-hierarchy port 3/1/1

===============================================================================
Hs Scheduler Hierarchy Information
===============================================================================
Hs Sched Policy Name         : hs-sched-pol-1

Port Max-Rate : 659 Mbps

Scheduler Priority 6
  Scheduler Class 6  Rate : 0 Mbps

Scheduler Priority 5
  Scheduler Class 5  Rate : 127 Mbps

Scheduler Priority 4
  Scheduler Class 4  Rate : 254 Mbps

Scheduler Priority 3
  Scheduler Class 3  Rate : 0 Mbps

Scheduler Priority 2
  Scheduler Class 2  Rate : 0 Mbps

Scheduler Priority 1
  Scheduler Class 1  Rate : 277 Mbps
===============================================================================
*A:PE-1#
```

The HS scheduler policy parameters can be overridden under the port policy configuration:

```
configure
    port <port-id>
        ethernet
            egress
                hs-scheduler-overrides [create]
                    group <group-id> rate <rate>
                    max-rate <rate>
                    scheduling-class <class> rate <rate>
                    scheduling-class <class> weight <weight-in-group>
```

Where (in order):

```
<port-id>            : slot/mda/port
<group-id>           : [1..1]
<rate>               : [1..100000|max] Mbps
```

```
<rate>               : [1..100000|max] Mbps
<class>              : [1..6]
<rate>               : [1..100000|max] Mbps
<class>              : [1..6]
<weight-in-group>    : [1..127]
```

SAP Egress

A SAP configured on an HSQ IOM port uses an HSQ queue group for its egress queues. This occurs automatically and dedicates one HSQ queue group, so eight egress queues, to each SAP egress. Only the queues to be used need to be configured within the SAP egress QoS policy.

The operation of classification, policing, and marking within a SAP egress QoS policy when applied to a SAP on an HSQ port is unchanged. For example, it is possible to use egress policers and direct the post-policer traffic to either a local HSQ queue group queue or to an HSQ queue group queue in an access egress queue group instance.

Most of the commands in a SAP egress QoS policy apply to the HSQ egress SAPs, with the following exceptions (including their SAP egress related overrides) being ignored:

- HSMDA commands
- **parent-location sla**
- Policer commands
  - **policers-hqos-manageable**
  - **policer scheduler-parent**
- Queue related
  - **adaptation-rule cir <adaptation-rule>**
  - **adv-config-policy**
  - **avg-frame-overhead**
  - **burst-limit**
  - **cbs**
  - **drop tail**
  - **parent**
  - **percent-rate cir**
  - **percent-rate local-limit**
  - **pool**
  - **port-parent**

  - **rate cir**
  - **wred-queue**
- Subscriber commands
  - **dynamic-policer** commands
  - **sub-insert-shared-pccrule**

The following SAP commands are not configurable on HSQ SAPs:

- **"ingress qos shared-queuing**
- **"ingress qos multipoint-shared**
- **"egress agg-rate limit-unused-bandwidth**
- **"egress agg-rate queue-frame-based-accounting**
- **"multi-service-site**

As mentioned, an HS attachment policy is applied to the SAP egress QoS policy to define the attachment of the queues to scheduling classes or WRR groups, with the WRR group then being attached to a scheduling class:

```
configure
    qos
        sap-egress <policy-id>
            hs-attachment-policy <policy-name>
```

Where:

```
<policy-id>          : [1..65535]|<name:64 char max>
<policy-name>        : [32 chars max]
```

When queues are attached to one of the HSQ queue group WRR groups, the relative weight of each queue within the group is configured under the queue, with the default weight being 1:

```
configure
    qos
        sap-egress <policy-id>
            queue <queue-id>
                hs-wrr-weight <weight>
```

Where:

```
<policy-id>          : [1..65535]|<name:64 char max>
<queue-id>           : [1..8]
<weight>             : [1..127]
```

The rate-related configuration of the two WRR groups in the HSQ queue group is defined within the SAP egress QoS policy, as follows, with the defaults being **rate max** and **adaptation-rule closest**:

```
configure
    qos
        sap-egress <policy-id>
            hs-wrr-group <group-id>
                adaptation-rule [pir <adaptation-rule>]
                percent-rate <percent>
                rate <rate>
```

Where:

```
<policy-id>          : [1..65535]|<name:64 char max>
<group-id>           : [1..2]
<adaptation-rule>    : max|min|closest
<percent>            : [0.01..100.00]
<rate>               : [1..2000000000|max] Kbps
```

The **percent-rate** configured within the **hs-wrr-group** and under a queue is relative to the port rate, so is equivalent to the queue **port-limit** and includes both the **egress-rate** and HS scheduler policy **max-rate**, if configured.

The SAP egress queue default slope policy is *_tmnx_hs_default*. A user-defined slope policy can be configured on a queue, as follows:

```
configure
    qos
        sap-egress <policy-id>
            queue <queue-id>
                hs-wred-queue [policy <slope-policy-name>]
```

Where:

```
<policy-id>          : [1..65535]|<name:64 char max>
<queue-id>           : [1..8]
<slope-policy-name>  : [32 chars max]
```

The *highplus* slope and time average factor in the applied slope policy are ignored on HSQ queue group queues.

By default, SAP egress HSQ queue group queues use buffers from the standard port class pools on their associated port. Each queue can be configured to use the port alternative class pools, as follows:

```
configure
    qos
        sap-egress <policy-id>
            queue <queue-id>
                hs-alt-port-class-pool
```

Where:

```
<policy-id>          : [1..65535]|<name:64 char max>
<queue-id>           : [1..8]
```

WRR group scheduling between queues and WRR groups in different HSQ queue groups is available at a primary shaper scheduling class. This is configured within a SAP egress QoS policy, as follows:

```
configure
    qos
        sap-egress <policy-id>
            hs-wrr-group <group-id>
                hs-class-weight <weight>
            queue <queue-id>
                hs-class-weight <weight>
```

Where (in order):

```
<policy-id>          : [1..65535]|<name:64 char max>
<group-id>           : [1..2]
<weight>             : 1|2|4|8
```

The **hs-class-weight** parameter under the **queue** or **hs-wrr-group** statement specifies the relative weight of the respective **queue** or **hs-wrr-group** for scheduling opportunities when their parent primary shaper scheduling class is serviced. By default, the **hs-class-weight** is 1.

Note: This parameter should not be confused with the **hs-wrr-weight** parameter, which specifies the relative weights of different queues within the same HSQ queue group WRR group.

The HSQ queue group aggregate rate is applied to a SAP egress using the **agg-rate rate** command:

```
configure
    service
        {ipipe <service-id>|epipe <service-id>|vpls <service-id>|
         ies <service-id> interface <ip-int-name>|
         vprn <service-id> interface <ip-int-name>}
            sap
                egress
                    agg-rate
                        rate <kilobits-per-second>
```

Where:

```
<service-id>         : [1..2147483647]|<svc-name:64 char max>
<ip-int-name>        : [32 chars max]  (must start with a letter)
<kilobits-per-seco*> : [1..3200000000|max] Kbps
```

The following HSQ-specific overrides are available under a SAP egress corresponding to the preceding commands:

```
configure
```

```
    service
        {ipipe <service-id>|epipe <service-id>|vpls <service-id>|
         ies <service-id> interface <ip-int-name>|
         vprn <service-id> interface <ip-int-name>}
            sap
                egress
                    queue-override
                        hs-secondary-shaper <policy-name>
                        hs-wrr-group <group-id> [create]
                        hs-wrr-group <group-id> class-weight <weight>
                        hs-wrr-group <group-id> percent-rate <percent>
                        hs-wrr-group <group-id> rate <rate>
                        queue <queue-id> hs-class-weight <weight>
                        queue <queue-id> hs-wred-queue policy <slope-policy-name>
                        queue <queue-id> hs-wrr-weight <weight>
```

Where (in order):

```
<service-id>          : [1..2147483647]|<svc-name:64 char max>
<ip-int-name>         : [32 chars max]  (must start with a letter)
<policy-name>         : [32 chars max]
<group-id>            : [1..2]
<weight>              : 1|2|4|8
<percent>             : [0.01..100.00]
<rate>                : [1..2000000000|max] Kbps
<queue-id>            : [1..8]
<weight>              : 1|2|4|8
<slope-policy-name>   : [32 chars max]
<weight>              : [1..127]
```

Note: Queue depth monitoring is supported for SAP egress HSQ queue group queues. This is enabled by configuring the queue override **monitor-depth** command under SAP egress with the associated **show** output displaying buffer occupancy in ranges of 10% of the queue depth for each configured queue.

An additional SAP egress override is provided to redirect the traffic from an HSQ queue group to a user-configured secondary shaper:

```
configure
    service
        {ipipe <service-id>|epipe <service-id>|vpls <service-id>|
         ies <service-id> interface <ip-int-name>|
         vprn <service-id> interface <ip-int-name>}
            sap
                egress
                    queue-override
                        hs-secondary-shaper <policy-name>
```

Where:

```
<service-id>          : [1..2147483647]|<svc-name:64 char max>
<ip-int-name>         : [32 chars max]  (must start with a letter)
<policy-name>         : [32 chars max]
```

When using pseudowire SAPs (PW-SAPs), an HS secondary shaper can be configured under the SDP binding to apply QoS control to the PW used by the SAPs, as follows:

```
configure
    service
        sdp <sdp-id>
            binding
                pw-port <pw-port-id>
                    egress
                        shaper
                        pw-sap-secondary-shaper <pw-sap-sec-shaper-name>
```

Where:

```
<sdp-id>             : [1..32767]
<pw-port-id>         : [1..32767]
<pw-sap-sec-shaper*> : [32 chars max]
```

When the first egress SAP or PW-SAP is associated with a user-configured HS secondary shaper, the system instantiates a default primary shaper for that secondary shaper, which is used by all HSQ queue groups sending traffic to that secondary shaper.

An IES interface SAP is configured with the HSQ queue group having an aggregate rate of 50 Mb/s (using the **agg-rate rate** command). The SAP egress traffic is directed to an HS secondary shaper, which is applied to port 3/1/2:

```
*A:PE-1>config>service>ies# info
----------------------------------------------
            description "HSQ egress SAP queues"
            interface "PE-1-IES-1" create
                address 192.168.11.1/30
                sap 3/1/2:1 create
                    egress
                        qos 10
                        queue-override
                            hs-secondary-shaper "hs-sec-shaper-1"
                        exit
                        agg-rate
                            rate 50000
                        exit
                    exit
                exit
            exit
            no shutdown
```

The SAP egress QoS policy contains the applied HS attachment policy described for the queue attachment. Queue 1 is configured with a WRR weight of 2. Rates of 20 Mb/s and 10 Mb/s are configured on queues 6 and 7, respectively. WRR group 1 is configured with a rate of 40 Mb/s. DSCP values are used to classify the egress traffic to the forwarding classes mapped to the queues:

```
*A:PE-1>config>qos# sap-egress 10
*A:PE-1>config>qos>sap-egress# info
----------------------------------------------
            hs-attachment-policy "hs-att-pol-1"
            queue 1 create
                hs-wrr-weight 2
            exit
            queue 2 create
            exit
            queue 6 create
                rate 20000
            exit
            queue 7 create
                rate 10000
            exit
            hs-wrr-group 1
                rate 40000
            exit
            fc af create
                queue 2
            exit
            fc ef create
                queue 6
            exit
            fc h1 create
                queue 7
            exit
            dscp cs1 fc "af"
            dscp be fc "be"
            dscp cs2 fc "ef"
            dscp cs3 fc "h1"
```

The queue information is shown as follows:

```
*A:PE-1# show hs-pools port 3/1/2 egress sap 3/1/2:1 |
               match "Queue Information" pre-lines 1 post-lines 40
-------------------------------------------------------------------------------
Queue Information
-------------------------------------------------------------------------------
Queue Name         : 1->3/1/2:1->1
FC Map             : be l2 l1 h2 nc
Admin PIR          : 40000              Oper PIR               : 0
Admin MBS          : 64 KB              Oper MBS               : 64 KB
HS Wrr Group       : 1
HS Wrr Class Weight: 1                  HS Wrr Weight          : 2
Depth              : 0
HS Class           : 1                  HS Alt Port Class Pool : No
HS Slope Policy    : _tmnx_hs_default

Queue Name         : 1->3/1/2:1->2
FC Map             : af
Admin PIR          : 40000              Oper PIR               : 0
Admin MBS          : 64 KB              Oper MBS               : 64 KB
HS Wrr Group       : 1
HS Wrr Class Weight: 1                  HS Wrr Weight          : 1
Depth              : 0
HS Class           : 1                  HS Alt Port Class Pool : No
HS Slope Policy    : _tmnx_hs_default
```

```
Queue Name         : 1->3/1/2:1->6
FC Map             : ef
Admin PIR          : 20000              Oper PIR               : 20000
Admin MBS          : 64 KB              Oper MBS               : 64 KB
HS Wrr Group       : (not-applicable)
HS Wrr Class Weight: 1                  HS Wrr Weight          : 0
Depth              : 0
HS Class           : 4                  HS Alt Port Class Pool : No
HS Slope Policy    : _tmnx_hs_default

Queue Name         : 1->3/1/2:1->7
FC Map             : h1
Admin PIR          : 10000              Oper PIR               : 10000
Admin MBS          : 64 KB              Oper MBS               : 64 KB
HS Wrr Group       : (not-applicable)
HS Wrr Class Weight: 1                  HS Wrr Weight          : 0
Depth              : 0
HS Class           : 5                  HS Alt Port Class Pool : No
HS Slope Policy    : _tmnx_hs_default
```

The current scheduler traffic rates, including the port and secondary shaper current aggregate traffic rate and current traffic rates for each scheduling class, together with queue current traffic rates on the SAP specified, are shown as follows:

```
*A:PE-1# show qos hs-scheduler-hierarchy sap 3/1/2:1 egress

===============================================================================
Hs Scheduler Hierarchy Information
===============================================================================
Hs Sched Policy Name         : default
PortId                       : 3/1/2

Port Max-Rate : 138 Mbps
Hs-Sec-Shaper:hs-sec-shaper-1 Agg-Rate : 57728 Kbps

Scheduler Priority 6
  Scheduler Class 6  Rate : 0 Mbps
    Hs-Sec-Shaper:hs-sec-shaper-1 Class 6 Rate : 0 Kbps
      sap-3/1/2:1->8            Rate : 0 Kbps

Scheduler Priority 5
  Scheduler Class 5  Rate : 22 Mbps
    Hs-Sec-Shaper:hs-sec-shaper-1 Class 5 Rate : 11454 Kbps
      sap-3/1/2:1->7            Rate : 10040 Kbps

Scheduler Priority 4
  Scheduler Class 4  Rate : 45 Mbps
    Hs-Sec-Shaper:hs-sec-shaper-1 Class 4 Rate : 22898 Kbps
      sap-3/1/2:1->6            Rate : 20080 Kbps

Scheduler Priority 3
  Scheduler Class 3  Rate : 0 Mbps
    Hs-Sec-Shaper:hs-sec-shaper-1 Class 3 Rate : 0 Kbps

Scheduler Priority 2
  Scheduler Class 2  Rate : 0 Mbps
    Hs-Sec-Shaper:hs-sec-shaper-1 Class 2 Rate : 0 Kbps
```

```
Scheduler Priority 1
  Scheduler Class 1  Rate : 69 Mbps
    Hs-Sec-Shaper:hs-sec-shaper-1 Class 1 Rate : 23375 Kbps
      sap-3/1/2:1->1  Group: 1 Rate : 13408 Kbps
      sap-3/1/2:1->2  Group: 1 Rate : 6684 Kbps
===============================================================================
===============================================================================
*A:PE-1#
```

The regular SAP **show** commands are supported with SAPs on an HSQ IOM; for example, the SAP statistics:

```
*A:PE-1# show service id 1 sap 3/1/2:1 stats

===============================================================================
Service Access Points(SAP)
===============================================================================
Service Id         : 1
SAP                : 3/1/2:1                  Encap             : q-tag
Description        : (Not Specified)
Admin State        : Up                       Oper State        : Up
Flags              : None
Multi Svc Site     : None
Last Status Change : 09/28/2017 14:05:52
Last Mgmt Change   : 09/28/2017 14:00:29
-------------------------------------------------------------------------------
Sap per Queue stats
-------------------------------------------------------------------------------
                        Packets                 Octets

Ingress Queue 1 (Unicast) (Priority)
Off. HiPrio           : 0                       0
Off. LowPrio          : 12000                   1536000
Dro. HiPrio           : 0                       0
Dro. LowPrio          : 0                       0
For. InProf           : 0                       0
For. OutProf          : 12000                   1536000

Egress Queue 1
For. In/InplusProf    : 0                       0
For. Out/ExcProf      : 2000                    256000
Dro. In/InplusProf    : 0                       0
Dro. Out/ExcProf      : 0                       0

Egress Queue 2
For. In/InplusProf    : 0                       0
For. Out/ExcProf      : 2000                    256000
Dro. In/InplusProf    : 0                       0
Dro. Out/ExcProf      : 0                       0

Egress Queue 6
For. In/InplusProf    : 0                       0
For. Out/ExcProf      : 1000                    128000
Dro. In/InplusProf    : 0                       0
Dro. Out/ExcProf      : 0                       0

Egress Queue 7
```

```
For. In/InplusProf     : 0                       0
For. Out/ExcProf       : 1000                    128000
Dro. In/InplusProf     : 0                       0
Dro. Out/ExcProf       : 0                       0
===============================================================================
*A:PE-1#
```

Network Egress

Each network egress port uses one HSQ queue group for its egress queues. This occurs automatically and dedicates one HSQ queue group, so eight egress queues, to each network egress port, which are used by all network interfaces configured on that port.

The HSQ-specific configuration of network egress queues on an HSQ IOM is similar to that for SAP egress.

The operation of classification, policing, and marking within the network QoS and network queue policies when applied to a network interface and egress HSQ port, respectively, is unchanged. Only the queues to be used need to be configured within the network queue policy.

Most of the commands in a network queue policy apply to the HSQ egress network interfaces, with the following exceptions being ignored:

- Queue commands
  - *adaptation-rule cir <adaptation-rule>*
  - **avg-frame-overhead**
  - **mbs**
  - **cbs**
  - **drop tail**
  - **port-parent**
  - **pool**
  - **rate cir**

The network queue policy **mbs** parameter is ignored, and replaced for HSQ queue group queues with the **hs-mbs** parameter. This is required to allow a more suitable default value to be assigned for the operation of HSQ queues. Both are configured as fractional percentages, with the default for the **mbs** parameter being 50% of the network egress pool, which is not used for HSQ queues, whereas the default for **hs-mbs** is 100% of one second of the queue PIR, converted to bytes. If the queue rate is **max**, the port rate is used (including the HS scheduler policy **max-rate** and the **egress-rate**, if configured on the port).

The network-queue policy has the same HS-specific configuration as in the SAP egress QoS policy, so is not repeated here, but includes:

- The application of an HS attachment policy to define the attachment of the queues to scheduling classes or WRR groups, with the WRR group then being attached to a scheduling class.
- The queue **hs-wrr-weight** to configure the relative weight of each queue within its parented WRR group.
- The rate-related configuration of the two WRR groups in the HSQ queue group.
- The slope policy configured on each HS WRED queue, again with the *highplus* slope and time average factor being ignored.
- The use of the port alternative class pools by each queue.
- The HS class weight for WRR group scheduling between queues and WRR groups in different HSQ queue groups at a primary shaper scheduling class.

HS WRR group and queue rates are configured as a percentage of the port rate, which includes both the **egress-rate** and HS scheduler policy **max-rate**, if configured.

Secondary shapers and HSQ queue group aggregate rates are not applicable to network egress HSQ queue group queues.

The related network queue policy configuration is as follows:

```
configure
    qos
        network-queue <policy-name> [create]
            hs-attachment-policy <policy-name>
            hs-wrr-group <group-id>
                adaptation-rule [pir <adaptation-rule>]
                hs-class-weight <weight>
                rate <percent>
            queue <queue-id> [multipoint] [<queue-type>] [create]
                hs-alt-port-class-pool
                hs-class-weight <weight>
                hs-mbs <percent-of-queue-rate>
                hs-wred-queue [policy <slope-policy-name>]
                hs-wrr-weight <weight>
```

See the SAP Egress section for details of the preceding parameters.

A network interface is configured on port 3/1/1:1:

```
*A:PE-1>config>router# info | match "IP Configuration" pre-lines 1 post-lines 10
#--------------------------------------------------
echo "IP Configuration"
#--------------------------------------------------
        interface "PE-1-Network-1"
            address 192.168.10.1/30
            description "HSQ network egress queues"
            port 3/1/1:1
            qos 10
            no shutdown
        exit
```

The network QoS policy 10 applied to the interface only contains the necessary egress **dscp** statements to classify the egress traffic to the forwarding classes mapped to the queues.

The network queue policy configured on port 3/1/1 contains the applied HS attachment policy described for the queue attachment. Queue 1 is configured with a WRR weight of 2. Rates of 2% and 1% are configured on queues 6 and 7, respectively. WRR group 1 is configured with a rate of 2%:

```
*A:PE-1>config>qos>network-queue# info
----------------------------------------------
            hs-attachment-policy "hs-att-pol-1"
            queue 1 create
                hs-wrr-weight 2
            exit
            queue 2 create
            exit
            queue 6 create
                rate 2
            exit
            queue 7 create
                rate 1
            exit
            hs-wrr-group 1
                rate 2
            exit
```

The network queue information is shown as follows:

```
*A:PE-1# show hs-pools port 3/1/1 egress network-queues |
                            match "Queue Information" pre-lines 1 post-lines 40
-------------------------------------------------------------------------------
Queue Information
-------------------------------------------------------------------------------
Queue Name         : 1 Net=be Port=3/1/1
FC Map             : be l2 l1 h2 nc
Admin PIR          : 200000             Oper PIR               : 0
Admin MBS          : 25000000 B         Oper MBS               : 24416 KB
HS Wrr Group       : 1
HS Wrr Class Weight: 1                  HS Wrr Weight          : 2
Depth              : 0
HS Class           : 1                  HS Alt Port Class Pool : No
HS Slope Policy    : _tmnx_hs_default

Queue Name         : 2 Net=af Port=3/1/1
FC Map             : af
Admin PIR          : 200000             Oper PIR               : 0
Admin MBS          : 25000000 B         Oper MBS               : 24416 KB
HS Wrr Group       : 1
HS Wrr Class Weight: 1                  HS Wrr Weight          : 1
```

```
Depth              : 0
HS Class           : 1                  HS Alt Port Class Pool : No
HS Slope Policy    : _tmnx_hs_default

Queue Name         : 6 Net=ef Port=3/1/1
FC Map             : ef
Admin PIR          : 200000             Oper PIR               : 200000
Admin MBS          : 25000000 B         Oper MBS               : 24416 KB
HS Wrr Group       : (not-applicable)
HS Wrr Class Weight: 1                  HS Wrr Weight          : 0
Depth              : 0
HS Class           : 4                  HS Alt Port Class Pool : No
HS Slope Policy    : _tmnx_hs_default

Queue Name         : 7 Net=h1 Port=3/1/1
FC Map             : h1
Admin PIR          : 100000             Oper PIR               : 100000
Admin MBS          : 12500000 B         Oper MBS               : 12208 KB
HS Wrr Group       : (not-applicable)
HS Wrr Class Weight: 1                  HS Wrr Weight          : 0
Depth              : 0
HS Class           : 5                  HS Alt Port Class Pool : No
HS Slope Policy    : _tmnx_hs_default
```

The current port-based scheduler traffic rates for network egress are shown as follows:

```
*A:PE-1# show qos hs-scheduler-hierarchy port 3/1/1

===============================================================================
Hs Scheduler Hierarchy Information
===============================================================================
Hs Sched Policy Name         : hs-sched-pol-1

Port Max-Rate : 660 Mbps

Scheduler Priority 6
  Scheduler Class 6  Rate : 0 Mbps

Scheduler Priority 5
  Scheduler Class 5  Rate : 127 Mbps

Scheduler Priority 4
  Scheduler Class 4  Rate : 255 Mbps

Scheduler Priority 3
  Scheduler Class 3  Rate : 0 Mbps

Scheduler Priority 2
  Scheduler Class 2  Rate : 0 Mbps

Scheduler Priority 1
  Scheduler Class 1  Rate : 277 Mbps
===============================================================================
*A:PE-1#
```

The regular **show** commands are supported with network interfaces on an HSQ IOM; for example, the port queue statistics:

```
*A:PE-1# show port 3/1/1 detail |
     match expression "Ethernet Interface|Egress Queue" pre-lines 1 post-lines 6
===============================================================================
Ethernet Interface
===============================================================================
Description        : 10-Gig Ethernet
Interface          : 3/1/1                   Oper Speed       : 10 Gbps
Link-level         : Ethernet                Config Speed     : N/A
Admin State        : up                      Oper Duplex      : full
Oper State         : up                      Config Duplex    : N/A

Egress Queue  1              Packets                 Octets
     In/Inplus Prof fwded  :    0                       0
     In/Inplus Prof dropped:    0                       0
     Out/Exc Prof fwded    :    2000                    256000
     Out/Exc Prof dropped  :    0                       0
Egress Queue  2              Packets                 Octets
     In/Inplus Prof fwded  :    0                       0
     In/Inplus Prof dropped:    0                       0
     Out/Exc Prof fwded    :    2000                    256000
     Out/Exc Prof dropped  :    0                       0
Egress Queue  6              Packets                 Octets
     In/Inplus Prof fwded  :    1000                    128000
     In/Inplus Prof dropped:    0                       0
     Out/Exc Prof fwded    :    0                       0
     Out/Exc Prof dropped  :    0                       0
Egress Queue  7              Packets                 Octets
     In/Inplus Prof fwded  :    1000                    128000
     In/Inplus Prof dropped:    0                       0
     Out/Exc Prof fwded    :    0                       0
     Out/Exc Prof dropped  :    0                       0
===============================================================================
*A:PE-1#
```

Access and Network Egress Queue Groups

Each access and network egress queue group instance configured on an HSQ IOM uses an HSQ queue group for its queues. This occurs automatically and dedicates one HSQ queue group, so eight egress queues, to each egress queue group instance. Only the queues to be used need to be configured within the egress queue group template.

The operation of classification, policing, and marking related to egress queue group instances on an HSQ port is unchanged.

Most of the commands in an egress queue group template apply to the HSQ egress queue group instances, with the following exceptions (including their related overrides) being ignored:

- HSMDA commands
- **queues-hqos-manageable**
- Queue related
  - **adaptation-rule cir <adaptation-rule>**
  - **adv-config-policy**
  - **avg-frame-overhead**
  - **burst-limit**
  - **cbs**
  - **drop tail**
  - **dynamic-mbs**
  - **parent**
  - **percent-rate cir**
  - **pool**
  - **port-parent**
  - **rate cir**
  - **wred-queue**

The following commands are not configurable under port access and network egress queue group instances:

- **egress agg-rate limit-unused-bandwidth**
- **egress agg-rate queue-frame-based-accounting**

The configuration of egress queue groups using HSQ queue groups is unchanged. The egress queue group template is applied under *config>port>ethernet>access>egress* or *config>port>ethernet>network>egr* to create the queue group instances, and traffic is redirected to these instances in either a SAP egress QoS policy or network QoS policy.

The system-created egress queue group instances each use an HSQ queue group; for example, the post-policer access egress *policer-output-queues* queue groups.

The egress queue group template has the same HS-specific configuration as in the SAP egress QoS policy, so is not repeated here, but includes:

- The application of an HS attachment policy to define the attachment of the queues to scheduling classes or WRR groups, with the WRR group then being attached to a scheduling class.
- The queue **hs-wrr-weight** to configure the relative weight of each queue within its parented WRR group.
- The rate-related configuration of the two WRR groups in the HSQ queue group.

- The slope policy configured on each HS WRED queue, again with the *highplus* slope and time average factor being ignored.
- The use of the port alternative class pools by each queue.
- The HS class weight for WRR group scheduling between queues and WRR groups in different HSQ queue groups at a primary shaper scheduling class.

Secondary shapers are not applicable to both access and network egress queue group instance HSQ queue groups.

The related egress queue group template configuration syntax is as follows:

```
configure
    qos
        queue-group-templates
            egress
                queue-group <queue-group-name> [create]
                    hs-attachment-policy <policy-name>
                    hs-wrr-group <group-id>
                        adaptation-rule [pir <adaptation-rule>]
                        hs-class-weight <weight>
                        percent-rate <percent>
                        rate <rate>
                    queue <queue-id> [queue-type] [create]
                        hs-alt-port-class-pool
                        hs-class-weight <weight>
                        hs-wred-queue [policy <slope-policy-name>]
                        hs-wrr-weight <weight>
```

See the SAP Egress section for details of the preceding parameters.

The **percent-rate** configured within the **hs-wrr-group** and under a queue is relative to the port rate, and includes both the **egress-rate** and HS scheduler policy **max-rate**, if configured.

The HSQ queue group aggregate rate is applied to an egress queue group instance using the **agg-rate rate** command under the application of the queue group template on the port:

```
configure
    port <port-id>
        ethernet
            access
                egress
                    queue-group <queue-group-name> [instance <instance-id>]
                        agg-rate
                            rate <kilobits-per-second>
            network
                egress
                    queue-group <queue-group-name> [instance <instance-id>]
                        agg-rate
                            rate <kilobits-per-second>
```

Where:

```
<port-id>            : slot/mda/port
<queue-group-name>   : [32 chars max]
<instance-id>        : [1..65535]
<kilobits-per-seco*> : [1..3200000000|max] Kbps
```

When using HSQ queue groups with access or network egress queue group instances on 100G ports, the **hs-turbo** parameter can be configured under the port queue group instance to allow the corresponding HSQ queue group queues to achieve a higher throughput. The default is **no hs-turbo**. The **hs-turbo** parameter is not applicable to 10G ports and is ignored when configured under a queue group instance on a 10G port.

```
configure
    port <port-id>
        ethernet
            access
                egress
                    queue-group <queue-group-name> [instance <instance-id>]
                        hs-turbo
            network
                egress
                    queue-group <queue-group-name> [instance <instance-id>]
                        hs-turbo
```

Where:

```
<port-id>            : slot/mda/port
<queue-group-name>   : [32 chars max]
<instance-id>        : [1..65535]
```

Note: Queue depth monitoring is supported for access and network egress queue groups HSQ queues. This is enabled by configuring the queue override **monitor-depth** command under the queue group instance with the associated output showing buffer occupancy in ranges of 10% of the queue depth for each configured queue.

An egress queue group template is configured containing the applied HS attachment policy described for the queue attachment. Queue 1 is configured with a WRR weight of 2. Rates of 20 Mb/s and 10 Mb/s are configured on queues 6 and 7, respectively. WRR group 1 is configured with a rate of 40 Mb/s:

```
A:PE-1# configure qos queue-group-templates egress
A:PE-1>cfg>qos>qgrps>egr# info
----------------------------------------------
                queue-group "queue-group-1" create
                    hs-attachment-policy "hs-att-pol-1"
                    queue 1 best-effort create
                        hs-wrr-weight 2
                    exit
                    queue 2 best-effort create
                    exit
                    queue 6 best-effort create
                        rate 20000
                    exit
                    queue 7 best-effort create
                        rate 10000
                    exit
                    hs-wrr-group 1
                        rate 40000
                    exit
                exit
```

The queue group template is applied to the network port 3/1/1 and access port 3/1/2, each with an aggregate rate of 100 Mb/s:

```
A:PE-1# configure port 3/1/1
A:PE-1>config>port# info
----------------------------------------------
        ethernet
            network
                egress
                    queue-group "queue-group-1" instance 1 create
                        agg-rate
                            rate 100000
                        exit
                    exit
                exit
            exit
        exit
        no shutdown
----------------------------------------------
A:PE-1>config>port# exit all
A:PE-1# configure port 3/1/2
A:PE-1>config>port# info
----------------------------------------------
        ethernet
            access
                egress
                    queue-group "queue-group-1" instance 1 create
                        agg-rate
                            rate 100000
                        exit
                    exit
                exit
            exit
        exit
        no shutdown
```

The configured aggregate rates are shown as follows:

```
*A:PE-1# show port 3/1/[1,2] queue-group queue-group-1 instance 1 |
                                  match "Ethernet port" pre-lines 1 post-line 7
===============================================================================
Ethernet port 3/1/1 Network Egress queue-group
===============================================================================
Group Name       : queue-group-1      Instance-Id  : 1
Description      : (Not Specified)
```

```
Sched Policy     : None              Acct Pol     : None
Collect Stats    : disabled          Agg. Limit   : 100000
Limit Unused BW  : Disabled
HS Turbo Queues  : Disabled
===============================================================================
Ethernet port 3/1/2 Access Egress queue-group
===============================================================================
Group Name       : queue-group-1     Instance-Id  : 1
Description      : (Not Specified)
Sched Policy     : None              Acct Pol     : None
Collect Stats    : disabled          Agg. Limit   : 100000
Limit Unused BW  : Disabled
HS Turbo Queues  : Disabled
*A:PE-1#
```

An IES interface SAP is configured on port 3/1/2 with a SAP egress QoS policy redirecting the traffic to the access egress queue group instance, and a network interface is configured on port 3/1/1 with a network QoS policy redirecting the traffic to the network egress queue group instance.

The queue information of the access egress queue group HSQ queue group queues is shown as follows:

```
A:PE-1# show hs-pools port 3/1/2 egress queue-group "queue-group-1" |
                            match "Queue Information" pre-lines 1 post-lines 40
-------------------------------------------------------------------------------
Queue Information
-------------------------------------------------------------------------------
Queue Name         : accQGrp->queue-group-1:1(3/1/2)->1
FC Map             : not-applicable
Admin PIR          : 40000              Oper PIR               : 0
Admin MBS          : 64 KB              Oper MBS               : 64 KB
HS Wrr Group       : 1
HS Wrr Class Weight: 1                  HS Wrr Weight          : 2
Depth              : 0
HS Class           : 1                  HS Alt Port Class Pool : No
HS Slope Policy    : _tmnx_hs_default

Queue Name         : accQGrp->queue-group-1:1(3/1/2)->2
FC Map             : not-applicable
Admin PIR          : 40000              Oper PIR               : 0
Admin MBS          : 64 KB              Oper MBS               : 64 KB
HS Wrr Group       : 1
HS Wrr Class Weight: 1                  HS Wrr Weight          : 1
Depth              : 0
HS Class           : 1                  HS Alt Port Class Pool : No
HS Slope Policy    : _tmnx_hs_default

Queue Name         : accQGrp->queue-group-1:1(3/1/2)->6
FC Map             : not-applicable
Admin PIR          : 20000              Oper PIR               : 20000
Admin MBS          : 64 KB              Oper MBS               : 64 KB
HS Wrr Group       : (not-applicable)
HS Wrr Class Weight: 1                  HS Wrr Weight          : 0
Depth              : 0
HS Class           : 4                  HS Alt Port Class Pool : No
HS Slope Policy    : _tmnx_hs_default
```

```
Queue Name         : accQGrp->queue-group-1:1(3/1/2)->7
FC Map             : not-applicable
Admin PIR          : 10000              Oper PIR              : 10000
Admin MBS          : 64 KB              Oper MBS              : 64 KB
HS Wrr Group       : (not-applicable)
HS Wrr Class Weight: 1                  HS Wrr Weight         : 0
Depth              : 0
HS Class           : 5                  HS Alt Port Class Pool : No
HS Slope Policy    : _tmnx_hs_default
```

The equivalent information can be displayed for the network egress queue group HSQ queue group queues by replacing 3/1/2 by 3/1/1.

The current port scheduler aggregate traffic rate and the current traffic rates for each scheduling class, together with current queue traffic rates in the specified access or network egress queue group instance, are shown as follows:

```
*A:PE-1# show qos hs-scheduler-hierarchy port 3/1/2
                  queue-group "queue-group-1" instance 1 access

===============================================================================
Hs Scheduler Hierarchy Information
===============================================================================
Hs Sched Policy Name         : default

Port Max-Rate : 138 Mbps

Scheduler Priority 6
  Scheduler Class 6  Rate : 0 Mbps
      Queue 8            Rate : 0 Kbps

Scheduler Priority 5
  Scheduler Class 5  Rate : 22 Mbps
      Queue 7            Rate : 10062 Kbps

Scheduler Priority 4
  Scheduler Class 4  Rate : 45 Mbps
      Queue 6            Rate : 20124 Kbps

Scheduler Priority 3
  Scheduler Class 3  Rate : 0 Mbps

Scheduler Priority 2
  Scheduler Class 2  Rate : 0 Mbps

Scheduler Priority 1
  Scheduler Class 1  Rate : 69 Mbps
      Queue 1            Rate : 26837 Kbps
      Queue 2            Rate : 13397 Kbps
===============================================================================
*A:PE-1#
```

Conclusion

The HSQ IOM provides high scale QoS in terms of the number of ingress policers and egress queues supported. It supports six scheduling classes across multiple hierarchical levels of hardware egress shaping encompassing HSQ queue groups, primary shapers, secondary shapers, and port schedulers. A flexible buffer allocation mechanism permits both buffer isolation and buffer oversubscription for the queue buffer allocation.

Pseudowire QoS

This chapter describes pseudowire QoS configurations.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

The information and configuration is based on release 11.0.R4. There are no specific prerequisites for this configuration.

Overview

A pseudowire (PW) provides a virtual connection across an IP or MPLS network between services configured on provider edge (PE) devices. In SR OS release R10.0.R1, and later, it is possible to provide specific QoS to either a single pseudowire or a multiple pseudowires. This is supported for the following:

- SDP
  - MPLS
  - GRE
- Epipe
  - Including vc-switching and dynamic MS-PW
  - PBB-epipe
  - BGP-VPWS (release 11.0.R1 and later)
- VPLS
  - Mesh and spoke SDP
  - LDP signaled pseudowires
  - BGP-AD signaled pseudowires
  - I-VPLS, B-VPLS

   - R-VPLS
   - BGP-VPLS
- Spoke termination on IES/VPRN (both Epipe and Ipipe)
- Apipe (from R10.0.R4)
- Cpipe (from R10.0.R4)
- Fpipe (from R10.0.R4)
- Ipipe (from R10.0.R4)

It is supported at ingress on both Ethernet and POS/TDM ports and only on Ethernet ports at egress.

Bandwidth control is achieved using queue-groups which are implemented per flexpath (FP) at the ingress and per port at the egress (these being relative to the data path through the system), as shown in Figure 223 and Figure 224, respectively.

Figure 223 **Ingress PW QoS**

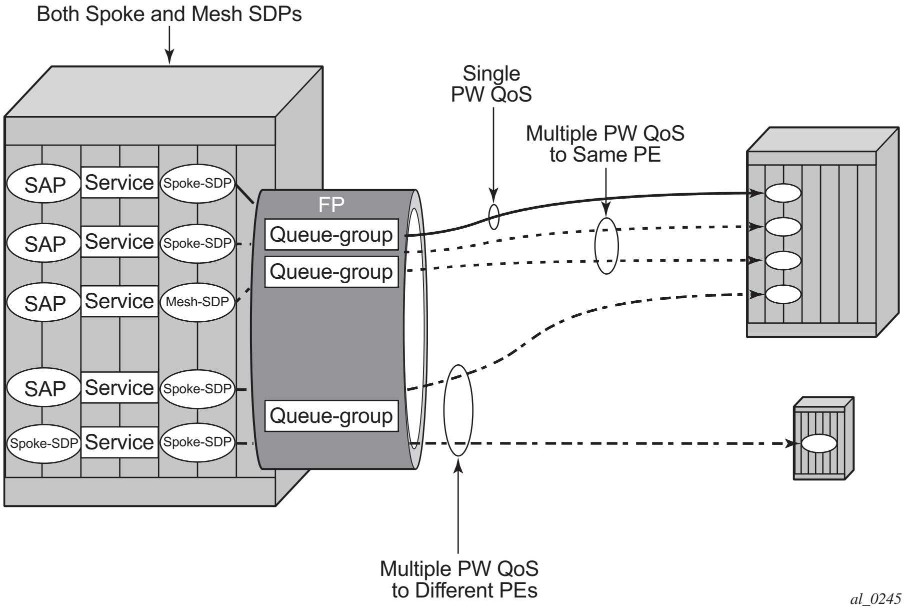
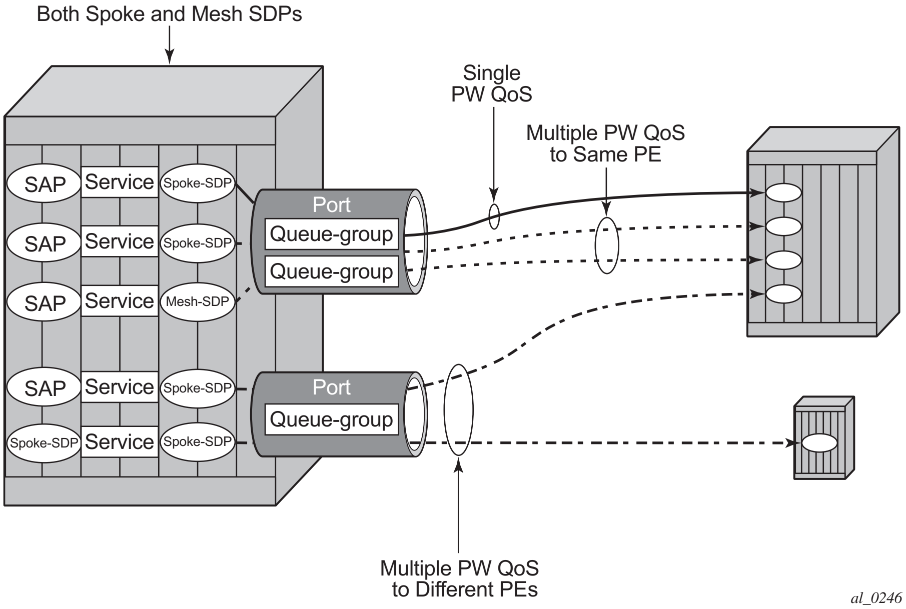


Figure 224 **Egress PW QoS**



Bandwidth control is applied independently for ingress and egress, and can be set up for a single pseudowire or for multiple pseudowires where the remote services are located on a single PE or on multiple PEs.

It is possible to benefit from Hierarchical QoS which can be configured under the queue-groups, but this is beyond the scope of this chapter.

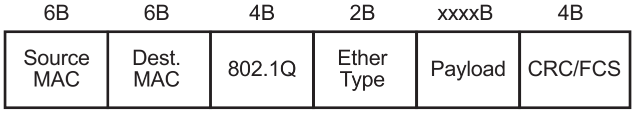
The ingress and egress classification and egress marking is configured by applying a network QoS policy to each pseudowire.

Ingress QoS

Ingress QoS is achieved using a queue group which is applied to an ingress FP on a card. Queue groups applied to an FP can only contain policers, not queues. The network QoS policy applied to the pseudowire redirects forwarding classes (FCs) to the individual queue group (unicast or multipoint) policers. The actual queue group to be used is defined separately to the network QoS policy, thereby allowing the network QoS policies to be independent from the queue groups used and therefore both are reusable.

Ingress bandwidth control does not take into account the outer Ethernet header, the MPLS labels/control word or GRE headers, or the FCS of the incoming frame. The configuration allows an offset to be added or subtracted from the received frame size in order to change the actual length used for the bandwidth control.
For example: if the same ingress rate is configured on a pseudowire (without a control word) and a dot1q SAP, what packet-byte-offset needs to be used on the pseudowire in order to achieve the same throughput as on the SAP?

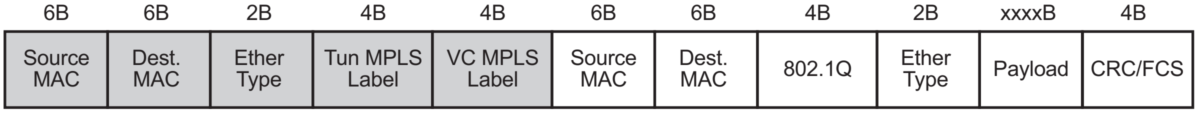
- SAP — The following shows the bytes in the frame that are used by default on a policer for the rate at a SAP ingress.



6B	6B	4B	2B	xxxxB	4B
Source MAC	Dest. MAC	802.1Q	Ether Type	Payload	CRC/FCS

al_0247

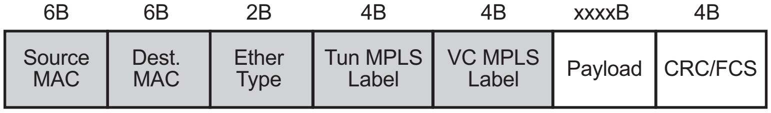
- VPLS Pseudowire — For a tagged (**vc-type vlan**) pseudowire, it would be necessary to add 4 bytes using the packet-byte-offset applied to the ingress policer in order to achieve the same throughput as on the SAP. This compensates for the omission of the FCS that is included on the SAP and so needs to be added.



6B	6B	2B	4B	4B	6B	6B	4B	2B	xxxxB	4B
Source MAC	Dest. MAC	Ether Type	Tun MPLS Label	VC MPLS Label	Source MAC	Dest. MAC	802.1Q	Ether Type	Payload	CRC/FCS

al_0248

- VPRN Pseudowire — For an Ipipe (**vc-type** ipipe) pseudowire, it would be necessary to add 22 bytes using the packet-byte-offset to the ingress policer to achieve the same throughput as on the SAP. This compensates for the omission of the source and destination MAC addresses (12 bytes), Ether type (2 bytes), VLAN tag (4 bytes) and the FCS (4 bytes) that are included on the SAP and so needs to be added.



6B	6B	2B	4B	4B	xxxxB	4B
Source MAC	Dest. MAC	Ether Type	Tun MPLS Label	VC MPLS Label	Payload	CRC/FCS

al_0249

The ingress classification is configured in the ingress section of the network QoS policy and is based on the outer encapsulation header only, the outer Ethernet header (dot1p/DE), MPLS labels (EXP), or GRE headers (DSCP). At an egress LER, the ler-use-dscp is applicable only to IES and VPRN pseudowires.

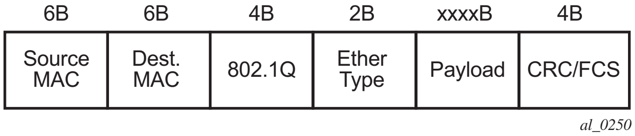
Egress QoS

Egress QoS is achieved using a queue group which is applied to an egress port. Queue groups applied to a port can contain both policers and queues. The network QoS policy applied to the pseudowire redirects forwarding classes (FCs) to the individual queue group policers/queues. The actual queue group to be used is defined separately to the network QoS policy, thereby allowing the network QoS policies to be independent from the queue groups used and therefore both are reusable.

Egress bandwidth control takes into account the outer Ethernet header, MPLS labels/control word, or GRE headers, and the FCS of the outgoing frame. The configuration allows an offset to be added or subtracted from the sent frame size in order to affect the actual length used for the bandwidth control.

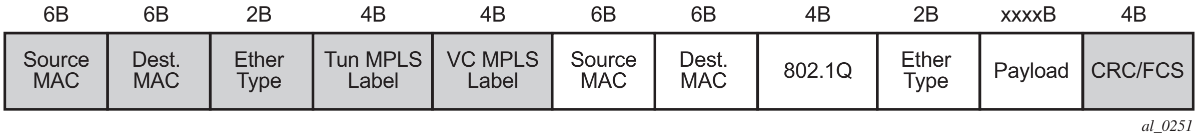
For example, if the same egress rate is configured on a pseudowire (without a control word) and a dot1q SAP, what packet-byte-offset needs to be used on the pseudowire in order to achieve the same throughput as on the SAP?

- SAP — The following shows the bytes in the frame that are used by default on a policer/queue at a SAP egress.



6B	6B	4B	2B	xxxxB	4B
Source MAC	Dest. MAC	802.1Q	Ether Type	Payload	CRC/FCS

- VPLS Pseudowire — For a tagged (**vc-type vlan**) pseudowire, it would be necessary to subtract 22 bytes using the packet-byte-offset applied to the egress policer/queue applied to achieve the same throughput as on the SAP. This compensates for the MPLS header (source and destination MAC addresses (12 bytes), Ether type (2 bytes), two labels (8 bytes)) that is not included on the SAP and needs to be subtracted.



6B	6B	2B	4B	4B	6B	6B	4B	2B	xxxxB	4B
Source MAC	Dest. MAC	Ether Type	Tun MPLS Label	VC MPLS Label	Source MAC	Dest. MAC	802.1Q	Ether Type	Payload	CRC/FCS

- VPRN Pseudowire — For an Ipipe (**vc-type ipipe**) pseudowire, it would be necessary to subtract 4 bytes using the packet-byte-offset applied to the egress policer/queue applied to achieve the same throughput as on the SAP. This compensates for the MPLS header (source and destination MAC addresses (12 bytes), Ether type (2 bytes), two labels (8 bytes)) that is not included on the SAP so is subtracted, and the source and destination MAC addresses (12 bytes), dot1q header (4 bytes) and Ether type (2 bytes) of the SAP frame which needs to be added. This results in subtracting 4 bytes.

6B	6B	2B	4B	4B	xxxxB	4B
Source MAC	Dest. MAC	Ether Type	Tun MPLS Label	VC MPLS Label	Payload	CRC/FCS

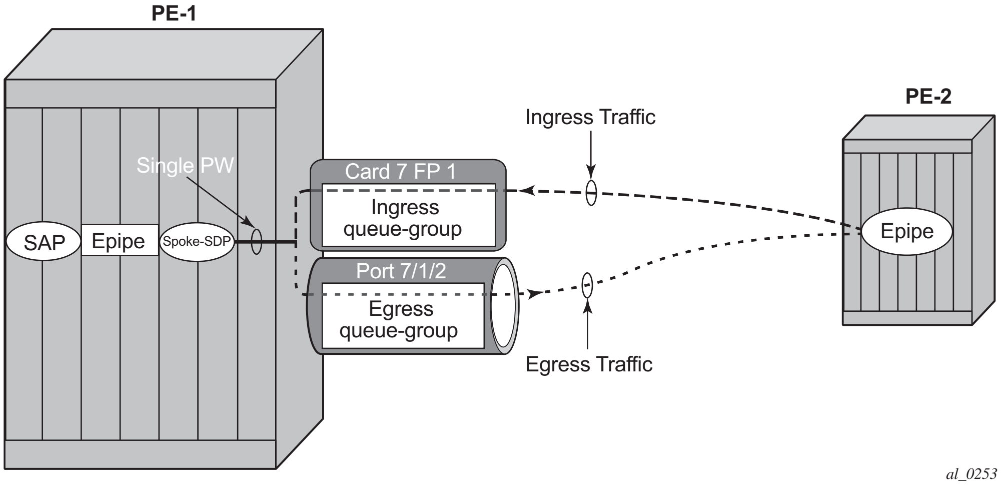
al_0252

The egress classification and marking is configured in the egress section of the network QoS policy. DSCP/prec egress reclassification is supported in SR OS release R10.0.R4, and later, for IES and VPRN spoke SDPs. The egress marking affects the outer encapsulation header, the outer Ethernet header (dot1p/DE), MPLS labels (EXP) or GRE headers (DSCP).

Configuration

The configuration of pseudowire QoS is described using an Epipe pseudowire. The topology is shown in Figure 225.

Figure 225 **Example Epipe Pseudowire Topology**



The following prerequisite configuration is assumed to be in place:

- Hardware provisioning
- IP address and routing
- MPLS protocols
- SDP
- Epipe service, including the SAP
- SAP QoS policies

Traffic is sent across a virtual leased line between PE-1 and PE-2 using Epipes with a pseudowire configured as a spoke SDP on each PE. The QoS is applied to the pseudowire at the ingress and egress of PE-1.

The following configuration is required for applying pseudowire QoS:

- Create the ingress and egress queue groups.
  These contain the ingress policer and egress policer/queue definitions.
- Create an instance of the ingress queue group on the ingress FP and instance of the egress queue group on the port that will be used for the pseudowire traffic.
- Create a network QoS policy to redirect the traffic to the ingress and egress queue groups, and to perform the ingress classification and egress marking.
- Apply the network QoS policy, together with the reference to the ingress and egress queue group instances, to the spoke SDP representing the pseudowire.

The traffic consists of two bidirectional flows, one in FC BE and one in FC EF. At the ingress of the pseudowire, each FC is assigned to its own policer, whereas at the egress of the pseudowire, FC BE is assigned to a queue and FC EF is assigned to a policer.

Although this example makes use of both ingress and egress queue groups, the focus is pseudowire QoS, so the full details of queue group configuration are not covered.

Create the Ingress and Egress Queue Groups

Queue groups are created using templates, which are separate for ingress and egress. The following shows the queue group templates configured.

```
configure qos
        queue-group-templates
            ingress
                queue-group "ingress-queue-group" create
                    policer 1 create
                        rate 6000
                        packet-byte-offset add 4
                    exit
                    policer 2 create
                        rate 4000
                        packet-byte-offset add 4
                    exit
                exit
            exit
            egress
                queue-group "egress-queue-group" create
                    queue 1 best-effort create
                        rate 6000
                        xp-specific
                            packet-byte-offset subtract 22
                        exit
                    exit
                    policer 1 create
                        rate 4000
                        packet-byte-offset subtract 22
                    exit
                exit
            exit
        exit
```

The ingress queue group has two policers associated with it; policer 1 will be used for the FC BE traffic and policer 2 will be used for the FC EF traffic. The configuration of policers in an ingress queue group is the same as that in a sap-ingress QoS policy, with the exception that the percent-rate is not supported within the queue group.

In order to achieve the same ingress throughput as that when applying the same rates to policers on a dot1q tagged SAP, the packet-byte-offset adds 4 bytes to the packet length for both policers.

The egress queue group has one queue (queue 1) that will be used for the FC BE traffic and one policer (policer 1) that will be used for the FC EF traffic. The configuration of policers in an egress queue group is the same as that in a sap-egress QoS policy, with the exception that the percent-rate is not supported within the queue group. The configuration of queues in an egress queue group is the same as in a sap-egress QoS policy, with the exception that the avg-frame-overhead is not supported within the queue group.

In order to achieve the same egress throughput as that when applying the same rates to policers/queues on a dot1q tagged SAP, the packet-byte-offset subtracts 22 bytes from the packet length for both the policer and queue.

Rates have been configured such that the ingress and egress capacity of the BE traffic is 6Mb/s and 4Mb/s for the EF traffic.

Create the Ingress FP and Egress Port Queue Group Instances

The queue group templates are then applied as individual instances to the ingress FP and egress port; using instances allows the reuse of the same template.

Following is the ingress FP configuration. From a QoS perspective, it is also possible to configure a policer-control-policy under the ingress queue group in order to perform hierarchical policing. In release R11.0R4, and later, the configuration supports overrides for both the policer-control-policy parameters and some of the queue group policer parameters.

```
configure
    card 7
        card-type imm5-10gb-xfp
        mda 1
            no shutdown
        exit
        fp 1
            ingress
                network
                    queue-group "ingress-queue-group" instance 1 create
                    exit
                exit
            exit
        exit
        no shutdown
    exit
```

Following is the egress port configuration. From a QoS perspective, it is also possible to configure under the egress queue group a policer-control-policy in order to perform hierarchical policing, a scheduler-policy in order to perform hierarchical shaping and overrides for some of the queue group queue parameters.

```
configure
    port 7/1/2
        ethernet
            network
                egress
                    queue-group "egress-queue-group" instance 1 create
                    exit
                exit
            exit
        exit
        no shutdown
    exit
```

If there are redundant network interfaces over which the pseudowire traffic can enter or exit the system, it is necessary to configure any ingress FP and egress port queue groups consistently across all possible interfaces to be used by the pseudowire to ensure the QoS is always applied. If a queue group configuration was omitted, the pseudowire would not be subject to the QoS defined in that queue group.

If a LAG is used, the system only allows the egress port queue group to be added or removed from the LAG primary port, thereby keeping the LAG configuration consistent. However, this is not possible at the ingress as the queue-group is applied at the FP, so it is necessary to ensure that the ingress queue group is applied consistently on all FPs corresponding to the configured LAG.

Create the Network QoS Policy

A network QoS policy is created to redirect ingress and egress traffic to the respective queue groups, and perform ingress classification (in this example).

The redirection to the queue group policer/queue is performed per FC.

At ingress, traffic can be redirected to policers (being the same or different policers) based on the traffic type. Unicast traffic is redirected to a policer specified by the policer command and will use the ingress shared policer-output-queues to access the switch fabric. All multipoint traffic is redirected to the policer specified by the multicast-policer command (for example, with a pseudowire configured in a VPLS service, all broadcast, unknown, and multicast traffic will use this policer). The multipoint traffic accesses the switch fabric using the Ingress Multicast Path Management queues. It is possible to individually redirect one traffic type (unicast or multipoint) within an FC to a queue group policer while allowing the other traffic type to use default network queues.

At egress, traffic can be redirected to a queue or to a policer. The policed traffic will exit the egress port using one of the default network queues (with the queue chosen by FC assignment) or optionally can use a queue in the egress queue group if configured in the port-redirect-group command following the policer parameter.

Any FC not redirected to a queue-group, will continue to use the regular default network ingress and egress queues.

The syntax for the FC redirection is as follows.

```
config# qos
  network <network-policy-id> [create]
      ingress
          fc <fc-name>
              fp-redirect-group multicast-policer <policer-id>
              fp-redirect-group policer <policer-id>
      egress
          fc <fc-name>
              port-redirect-group {queue <queue-id>|
policer <policer-id> [queue <queue-id>]}
```

The required commands are shown below.

```
configure qos
        network 10 create
            ingress
                lsp-exp 5 fc ef profile in
                fc be
                    fp-redirect-group policer 1
                exit
                fc ef
                    fp-redirect-group policer 2
                exit
            exit
            egress
                fc be
                    port-redirect-group queue 1
                exit
                fc ef
                    port-redirect-group policer 1
                exit
            exit
        exit
```

At ingress, the FC BE and FC EF traffic are redirected to the two policers in the queue-group applied to the FP. At egress, the two FCs are redirected to the queue and policer in the queue group applied to the egress port.

The ingress classification required here is for the traffic which is received with exp=5 to be in FC EF.

Apply Network QoS Policy with Queue Group Instances to the Spoke SDP

To apply the QoS to the pseudowire, the following commands can be used, dependent on the service type.

```
config# service {apipe|cpipe|epipe|fpipe|ipipe} <service-id>
  spoke-sdp <sdp-id:vc-id>
     ingress
       qos <network-policy-id> fp-redirect-group <queue-group-name>
                               instance <instance-id>
     egress
       qos <network-policy-id> port-redirect-group <queue-group-name>
                               instance <instance-id>
```

```
config# service {ies|vprn} <service-id>
  interface <ip-int-name>
    spoke-sdp <sdp-id:vc-id>
      ingress
        qos <network-policy-id> fp-redirect-group <queue-group-name>
                                instance <instance-id>
      egress
        qos <network-policy-id> port-redirect-group <queue-group-name>
                                instance <instance-id>
```

```
config# service vpls <service-id>
  {spoke-sdp|mesh-sdp} <sdp-id:vc-id>
     ingress
       qos <network-policy-id> fp-redirect-group <queue-group-name>
                               instance <instance-id>
     egress
       qos <network-policy-id> port-redirect-group <queue-group-name>
                               instance <instance-id>
```

For services using BGP auto-discovery to signal the pseudowire, the QoS configuration is included in the pseudowire template.

```
config# service pw-template <policy-id>
     ingress
       qos <network-policy-id> fp-redirect-group <queue-group-name>
                               instance <instance-id>
     egress
       qos <network-policy-id> port-redirect-group <queue-group-name>
                               instance <instance-id>
```

To propagate changes in a pw-template to existing BGP-AD pseudowires, it is necessary to use the following command:

```
tools perform service eval-pw-template policy-id
```

The allow-service-impact parameter is not required for changing the ingress or egress QoS definition as these do not affect the operational state of the pseudowire.

QoS applied directly to a pseudowire, using the preceding commands, takes precedence over any QoS applied to the network interface (using a network QoS policy with or without queue group redirection).

Each time a pseudowire uses a network egress port queue group, an FP resource is allocated. This only requires that the pseudowire egress QoS is configured with a port-redirect-group, and will occur even if there are no FCs redirected using a port-redirect-group within the configured network QoS policy. The resources used can be seen using the **tools dump system-resources** command and is listed under Egr Network Queue Group Mappings which is part of the total for the “Dynamic Service Entries “.

As an Epipe is used in this example, QoS is configured directly under a spoke SDP.

```
configure service
        epipe 1 customer 1 create
            spoke-sdp 1:1 vc-type vlan create
                ingress
                    qos 10 fp-redirect-group "ingress-queue-group" instance 1
                exit
                egress
                    qos 10 port-redirect-group "egress-queue-group" instance 1
                exit
                no shutdown
            exit
            no shutdown
        exit
```

The created network QoS policy is applied at both ingress and egress, with the ingress referencing the ingress queue group instance applied to the FP and the egress referencing the egress queue group instance applied to the port.

Show Output

The configured ingress queue group can be shown, including the details of the configured policers and where it is applied, as follows.

```
*A:PE-1# show qos queue-group "ingress-queue-group" ingress detail
===============================================================================
QoS Queue-Group Ingress
===============================================================================
-------------------------------------------------------------------------------
QoS Queue Group
-------------------------------------------------------------------------------
Group-Name     : ingress-queue-group
Description    : (Not Specified)
-------------------------------------------------------------------------------
---snip---
===============================================================================
Queue Group FP Maps
===============================================================================
Card Num       Fp Num              Instance            Type
-------------------------------------------------------------------------------
7              1                   1                   Network
-------------------------------------------------------------------------------
Entries found: 1
-------------------------------------------------------------------------------
===============================================================================
Queue Group Policer
===============================================================================
Policer Id     : 1
Description    : (Not Specified)
PIR Adptn      : closest                      CIR Adptn    : closest
Parent         : none                         Level        : 1
Weight         : 1                            Adv. Cfg Plcy: none
Admin PIR      : 6000                         Admin CIR    : 0
CBS            : def                          MBS          : def
Hi Prio Only   : def                          Pkt Offset   : 4
```

```
Profile Capped : Disabled
StatMode       : minimal
===============================================================================
Policer Id     : 2
Description    : (Not Specified)
PIR Adptn      : closest                      CIR Adptn    : closest
Parent         : none                         Level        : 1
Weight         : 1                            Adv. Cfg Plcy: none
Admin PIR      : 4000                         Admin CIR    : 0
CBS            : def                          MBS          : def
Hi Prio Only   : def                          Pkt Offset   : 4
Profile Capped : Disabled
StatMode       : minimal
```

Similar information can be shown for the egress queue group, including the details of the configured queue and policer and again where it is applied.

```
*A:PE-1# show qos queue-group "egress-queue-group" egress detail
===============================================================================
QoS Queue-Group Egress
===============================================================================
-------------------------------------------------------------------------------
QoS Queue Group
-------------------------------------------------------------------------------
Group-Name     : egress-queue-group
Description    : (Not Specified)

-------------------------------------------------------------------------------
Q  CIR Admin PIR Admin CBS         HiPrio PIR Lvl/Wt    Parent    BurstLimit(B)
   CIR Rule  PIR Rule  MBS                CIR Lvl/Wt    Wred-Queue     Slope
   Named-Buffer Pool                      Adv Config Policy Name
-------------------------------------------------------------------------------
1  0         6000      def         def    1/1           None          default
   closest   closest   def                0/1           disabled      default
   (not-assigned)                         (not-assigned)
---snip---
===============================================================================
Queue Group Ports (network)
===============================================================================
Port   Sched Pol    Policer-Ctrl-Pol  Acctg Pol Stats Description QGrp-Instance
-------------------------------------------------------------------------------
7/1/2                                           No                1
-------------------------------------------------------------------------------
---snip---
===============================================================================
Queue Group Policer
===============================================================================
Policer Id     : 1
Description    : (Not Specified)
PIR Adptn      : closest                      CIR Adptn    : closest
Parent         : none                         Level        : 1
Weight         : 1                            Adv. Cfg Plcy: none
Admin PIR      : 4000                         Admin CIR    : 0
CBS            : def                          MBS          : def
Hi Prio Only   : def                          Pkt Offset   : -22
Profile Capped : Disabled
StatMode       : minimal
```

```
---snip---
```

The following command shows where the ingress queue group has been applied.

```
*A:PE-1# show qos queue-group ingress association
===============================================================================
QoS Queue-Group Ingress
===============================================================================
---snip---
-------------------------------------------------------------------------------
QoS Queue Group
-------------------------------------------------------------------------------
Group-Name     : ingress-queue-group
Description    : (Not Specified)
---snip---
===============================================================================
Queue Group FP Maps
===============================================================================
Card Num       Fp Num              Instance            Type
-------------------------------------------------------------------------------
7              1                   1                   Network
-------------------------------------------------------------------------------
Entries found: 1
-------------------------------------------------------------------------------
---snip---
===============================================================================
```

The following command shows where the egress queue group has been applied.

```
*A:PE-1# show qos queue-group egress association
===============================================================================
QoS Queue-Group Egress
===============================================================================
-------------------------------------------------------------------------------
QoS Queue Group
-------------------------------------------------------------------------------
Group-Name     : egress-queue-group
Description    : (Not Specified)
---snip---
===============================================================================
Queue Group Ports (network)
===============================================================================
Port   Sched Pol    Policer-Ctrl-Pol  Acctg Pol Stats Description QGrp-Instance
-------------------------------------------------------------------------------
7/1/2                                           No                1
-------------------------------------------------------------------------------
---snip---
===============================================================================
```

The following command shows the ingress queue group applied to the FP on card 7.

```
*A:PE-1# show card 7 fp 1 ingress queue-group "ingress-queue-group" instance 1
mode network
===============================================================================
Card:7  Net.QGrp: ingress-queue-group  Instance: 1
===============================================================================
Group Name    : ingress-queue-group
Description   : (Not Specified)
Pol Ctl Pol   : None                    Acct Pol      : None
Collect Stats : disabled
```

The following command show the details of the policers in the ingress FP queue group.

```
*A:PE-1# show qos policer card 7 fp 1 queue-group "ingress-queue-group" instance 1
network detail
===============================================================================
Policer Info (Net-FPQG-1-ingress-queue-group:1->1), Slot 7
===============================================================================
Policer Name       : Net-FPQG-1-ingress-queue-group:1->1
Direction          : Ingress            Fwding Plane       : 1
Depth PIR          : 0 Bytes            Depth CIR          : 0 Bytes
Depth FIR          : 0 Bytes
MBS                : 7680 B             CBS                : 0 KB
Hi Prio Only       : 768 B              Pkt Byte Offset    : 4
Admin PIR          : 6000 Kbps          Admin CIR          : 0 Kbps
Oper PIR           : 6000 Kbps          Oper CIR           : 0 Kbps
Oper FIR           : 6000 Kbps
Stat Mode          : minimal
PIR Adaption       : closest            CIR Adaption       : closest
Adv.Cfg Plcy       : None               Profile Capped     : disabled
Parent Arbiter Name: (Not Specified)
-------------------------------------------------------------------------------
Arbiter Member Information
-------------------------------------------------------------------------------
Offered Rate       : 0 Kbps
Level              : 0                  Weight             : 0
Parent PIR         : 0 Kbps             Parent FIR         : 0 Kbps
Consumed           : 0 Kbps
-------------------------------------------------------------------------------
===============================================================================
Policer Info (Net-FPQG-1-ingress-queue-group:1->2), Slot 7
===============================================================================
Policer Name       : Net-FPQG-1-ingress-queue-group:1->2
Direction          : Ingress            Fwding Plane       : 1
Depth PIR          : 0 Bytes            Depth CIR          : 0 Bytes
Depth FIR          : 0 Bytes
MBS                : 5 KB               CBS                : 0 KB
Hi Prio Only       : 512 B              Pkt Byte Offset    : 4
Admin PIR          : 4000 Kbps          Admin CIR          : 0 Kbps
Oper PIR           : 4000 Kbps          Oper CIR           : 0 Kbps
Oper FIR           : 4000 Kbps
Stat Mode          : minimal
PIR Adaption       : closest            CIR Adaption       : closest
Adv.Cfg Plcy       : None               Profile Capped     : disabled
Parent Arbiter Name: (Not Specified)
-------------------------------------------------------------------------------
Arbiter Member Information
-------------------------------------------------------------------------------
Offered Rate       : 0 Kbps
Level              : 0                  Weight             : 0
Parent PIR         : 0 Kbps             Parent FIR         : 0 Kbps
Consumed           : 0 Kbps
```

```
-------------------------------------------------------------------------------
===============================================================================
Network Interface Association
-------------------------------------------------------------------------------
No Association Found.
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
SDP Association
-------------------------------------------------------------------------------
Policer Info (1->1:1->10), Slot 7
-------------------------------------------------------------------------------
SDP Association Count : 1
-------------------------------------------------------------------------------
```

The details of the queue and policer in the egress queue group applied to port 7/1/2 can also be shown as follows.

```
*A:PE-1# show port 7/1/2 queue-group egress "egress-queue-group" network instance 1
===============================================================================
Ethernet port 7/1/2 Network Egress queue-group
===============================================================================
Group Name    : egress-queue-group Instance-Id   : 1
Description   : (Not Specified)
Sched Policy  : None               Acct Pol      : None
Collect Stats : disabled           Agg. Limit    : -1

Queues
-------------------------------------------------------------------------------
Queue-Group   : egress-queue-group Instance-Id   : 1      Queue-Id   : 1
Description   : (Not Specified)
Admin PIR     : 6000*              Admin CIR     : 0*
PIR Rule      : closest*           CIR Rule      : closest*
CBS           : def*               MBS           : def*
Hi Prio       : def*

Policers
-------------------------------------------------------------------------------
Queue-Group   : egress-queue-group Instance-Id   : 1      Policer-Id   : 1
Description   : (Not Specified)
Admin PIR     : 4000*              Admin CIR     : 0*
PIR Rule      : closest*           CIR Rule      : closest*
CBS           : def*               MBS           : def*
Hi Prio       : def*

* means the value is inherited
```

The network QoS policy can be shown with the details of the configured FC redirection and ingress classification used on the pseudowire, as follows.

```
*A:PE-1# show qos network 10 detail
===============================================================================
QoS Network Policy
===============================================================================
-------------------------------------------------------------------------------
Network Policy (10)
-------------------------------------------------------------------------------
Policy-id        : 10                      Remark           : False
```

```
Forward Class    : be                     Profile          : Out
LER Use DSCP     : False
Description    : (Not Specified)
---snip---
-------------------------------------------------------------------------------
LSP EXP Bit Map                         Forwarding Class                Profile
-------------------------------------------------------------------------------
5                                       ef                              In
---snip---
-------------------------------------------------------------------------------
Egress Forwarding Class Mapping
-------------------------------------------------------------------------------
FC Value         : 0                      FC Name          : be
- DSCP Mapping
Out-of-Profile   : be                     In-Profile       : be
---snip---
DE Mark          :  None
Redirect Grp Q   :  1                     Redirect Grp Plcr:  None

---snip---
FC Value         : 5                      FC Name          : ef
---snip---
DE Mark          :  None
Redirect Grp Q   :  None                  Redirect Grp Plcr:  1

-------------------------------------------------------------------------------
Ingress Forwarding Class Mapping
-------------------------------------------------------------------------------
FC Value                : 0            FC Name                 : be
Redirect UniCast Plcr   : 1            Redirect MultiCast Plcr : None

---snip---
FC Value                : 5            FC Name                 : ef
Redirect UniCast Plcr   : 2            Redirect MultiCast Plcr : None
---snip---
```

The details of the configuration of the pseudowire QoS can be seen when showing the details of the SDP within the Epipe service, as follows.

```
*A:PE-1# show service id 1 sdp 1:1 detail
===============================================================================
Service Destination Point (Sdp Id : 1:1) Details
===============================================================================
-------------------------------------------------------------------------------
 Sdp Id 1:1  -(192.0.2.2)
-------------------------------------------------------------------------------
Description     : (Not Specified)
SDP Id             : 1:1                      Type              : Spoke
Spoke Descr     : (Not Specified)
VC Type            : VLAN                     VC Tag            : 0
Admin Path MTU     : 0                        Oper Path MTU     : 9190
Delivery           : MPLS
Far End            : 192.0.2.2
Tunnel Far End     : 192.0.2.2                LSP Types         : LDP
Hash Label         : Disabled                 Hash Lbl Sig Cap  : Disabled
Oper Hash Label    : Disabled

Admin State        : Up                       Oper State        : Up
```

```
---snip---
Ingress Qos Policy : 10                       Egress Qos Policy : 10
Ingress FP QGrp    : ingress-queue-group      Egress Port QGrp  : egress-queue*
Ing FP QGrp Inst   : 1                        Egr Port QGrp Inst: 1
```

The usage of the “Egr Network Queue Group Mappings” out of the total number of “Dynamic Service Entries” can be seen with the following command. Only one egress QoS pseudowire redirection has been configured.

```
*A:PE-1# tools dump system-resources
Resource Manager info at 005 m 07/31/13 13:11:03.355:

Hardware Resource Usage for Slot #7, CardType imm5-10gb-xfp, Cmplx #0:
                                 |   Total   | Allocated |    Free
 --------------------------------+-----------+-----------+------------
---snip---
         Dynamic Service Entries |      65535|          1|      65534
                Subscriber Hosts |           |          0|
             Encap Group Members |           |          0|
Egr Network Queue Group Mappings |           |          1|
```

It is possible to show the statistics on the ingress FP queue group used by the pseudowire.

```
*A:PE-1# show card 7 fp 1 ingress queue-group "ingress-queue-group" instance 1 mode
network statistics

===============================================================================
Card:7  Net.QGrp: ingress-queue-group  Instance: 1
===============================================================================
Group Name    : ingress-queue-group
Description   : (Not Specified)
Pol Ctl Pol   : None                    Acct Pol      : None
Collect Stats : disabled
-------------------------------------------------------------------------------
Statistics
-------------------------------------------------------------------------------
                        Packets                 Octets

Ing. Policer:  1  Grp: ingress-queue-group (Stats mode: minimal)
Off. All              :         184275                  23587200
Dro. All              :         36801                   4710528
For. All              :         147474                  18876672

Ing. Policer:  2  Grp: ingress-queue-group (Stats mode: minimal)
Off. All              :         184274                  23587072
Dro. All              :         85955                   11002240
For. All              :         98319                   12584832
```

Similar statistics can be shown for the egress port queue group used by the pseudowire.

```
*A:PE-1# show port 7/1/2 queue-group egress "egress-queue-group" network statistics
instance 1
-------------------------------------------------------------------------------
Ethernet port 7/1/2 Network Egress queue-group
-------------------------------------------------------------------------------
                        Packets                 Octets
Egress Queue:  1    Group: egress-queue-group    Instance-Id:  1
In Profile forwarded  : 0                       0
In Profile dropped    : 0                       0
Out Profile forwarded : 150989                  19326592
Out Profile dropped   : 37123                   4751744

Egress Policer:  1  Group: egress-queue-group  Instance-Id: 1
Stats mode: minimal
Off. All              : 188421                  24117888
Dro. All              : 87894                   11250432
For. All              : 100527                  12867456
```

Monitor commands are available to see the statistics (and rates on egress port queue group). As an example, the following shows the utilization on the queue and policer in the egress queue-group.

```
*A:PE-1# monitor port 7/1/2 queue-group "egress-queue-group" instance 1 egress network
egress-queue 1 repeat 1 rate
===============================================================================
Monitor Port Queue-Group Egress Network Queue Statistics
===============================================================================
-------------------------------------------------------------------------------
At time t = 0 sec (Base Statistics)
-------------------------------------------------------------------------------
                        Packets                 Octets

In Profile forwarded  : 0                       0
In Profile dropped    : 0                       0
Out Profile forwarded : 299113                  38286464
Out Profile dropped   : 74155                   9491840
-------------------------------------------------------------------------------
At time t = 11 sec (Mode: Rate)
-------------------------------------------------------------------------------
                        Packets/sec             Octets/sec              % Port
                                                                        Util.

In Profile forwarded  : 0                       0                       0.00
In Profile dropped    : 0                       0                       0.00
Out Profile forwarded : 5863                    750436                  0.06
Out Profile dropped   : 1466                    187609                  0.01
===============================================================================
```

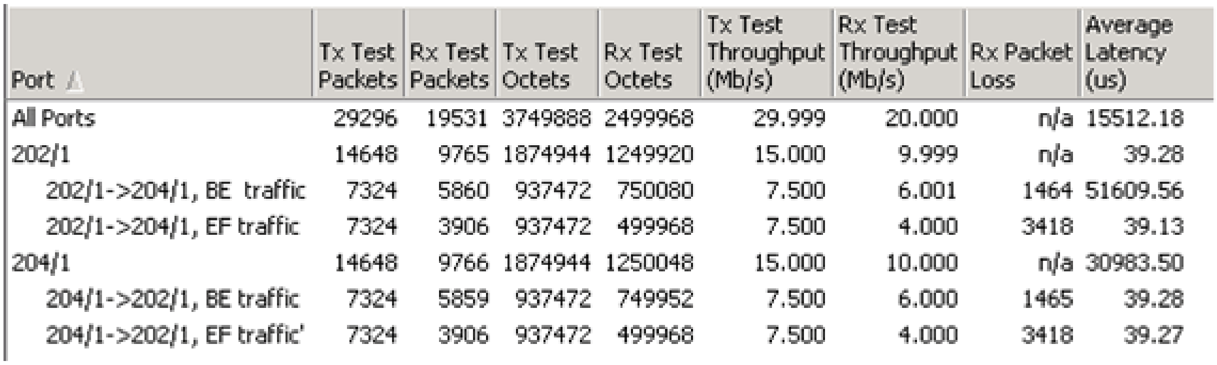
```
*A:PE-1# monitor port 7/1/2 queue-group "egress-queue-group" instance 1 egress network
policer 1 repeat 1 rate
===============================================================================
Monitor Port Queue-Group Egress Network Policer Statistics
===============================================================================
-------------------------------------------------------------------------------
At time t = 0 sec (Base Statistics)
-------------------------------------------------------------------------------
```

```
                        Packets                 Octets

Off. All              : 454750                  58208000
Dro. All              : 212181                  27159168
For. All              : 242569                  31048832

-------------------------------------------------------------------------------
At time t = 11 sec (Mode: Rate)
-------------------------------------------------------------------------------
                        Packets/sec             Octets/sec              % Port
                                                                        Util.

Off. All              : 7326                    937716                  0.07
Dro. All              : 3419                    437609                  0.03
For. All              : 3907                    500108                  0.04
===============================================================================
*A:PE-1#
```

As mentioned, the egress policer (FC EF) traffic exits the egress port by default using the related network queue on the port, as follows.

```
*A:PE-1# show port 7/1/2 detail
===============================================================================
Ethernet Interface
===============================================================================
Description        : 10-Gig Ethernet
Interface          : 7/1/2                       Oper Speed       : 10 Gbps
Link-level         : Ethernet                    Config Speed     : N/A
Admin State        : up                          Oper Duplex      : full
Oper State         : up                          Config Duplex    : N/A
---snip---
===============================================================================
Queue Statistics
===============================================================================
-------------------------------------------------------------------------------
---snip---
Egress Queue  6              Packets                 Octets
     In Profile forwarded  :    0                       0
     In Profile dropped    :    0                       0
     Out Profile forwarded :    102381                  15357150
     Out Profile dropped   :    0                       0
```

The throughput achieved using the preceding configuration can be verified in the traffic generator output. Port 202/1 is connected to PE-1 and port 204/1 is connected to PE-2.



Port	Tx Test Packets	Rx Test Packets	Tx Test Octets	Rx Test Octets	Tx Test Throughput (Mb/s)	Rx Test Throughput (Mb/s)	Rx Packet Loss	Average Latency (us)
All Ports	29296	19531	3749888	2499968	29.999	20.000	n/a	15512.18
202/1	14648	9765	1874944	1249920	15.000	9.999	n/a	39.28
202/1->204/1, BE traffic	7324	5860	937472	750080	7.500	6.001	1464	51609.56
202/1->204/1, EF traffic	7324	3906	937472	499968	7.500	4.000	3418	39.13
204/1	14648	9766	1874944	1250048	15.000	10.000	n/a	30983.50
204/1->202/1, BE traffic	7324	5859	937472	749952	7.500	6.000	1465	39.28
204/1->202/1, EF traffic'	7324	3906	937472	499968	7.500	4.000	3418	39.27

Conclusion

This example has shown the configuration and monitoring of pseudowire QoS, providing a powerful QoS solution for pseudowire applications. QoS can be applied independently to the ingress and/or egress of a single pseudowire or multiple pseudowires.

QoS Architecture and Basic Operation

This chapter provides information about QoS architecture and basic operation.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

The information in this chapter is applicable to all of the Nokia 7x50 platforms and assumes only FP2-and higher-based line cards are used. The chapter was initially written for SR OS release 9.0.R3. The CLI in the current edition corresponds to SROS release 14.0.R2.

Overview

The 7x50 platforms provide an extensive Quality of Service (QoS) capability for service provider solutions. QoS is a system behavior to provide different traffic with different amounts of resources, including buffer memory and queue serving time.

By allocating system resources with certain degrees of guarantee, the bandwidth can be used more efficiently and more controllably. Lack of buffer memory leads to packet drop, while a smaller amount of queue serving time normally means longer delay for the packet and may cause buffer memory to be completely consumed and eventually also lead to packet drop.

In a system, such as the 7x50 platform, different types of traffic contend for the same resources at several major points, such as the ingress to the switch fabric and the egress out of a physical port. In a multi-node network, QoS is achieved on hop by hop basis. Thus, QoS needs to be configured individually but with the consistency across the whole network.

This chapter is focused on the configuration of the basic QoS, namely the use of queues to shape traffic at the ingress and egress of the system. More sophisticated aspects will be referenced where appropriate but their details are beyond the scope of this chapter. Other topics not included are Hierarchical QoS scheduling, egress port-scheduler, queue-groups, named buffer pools, WRED-per-queue, LAGs, high scale MDA, QoS for ATM/FR and Enhanced Subscriber Management.

QoS Components

QoS consists of four main components:

- Classification
- Buffering (enqueuing)
- Scheduling (dequeuing)
- Remarking

These are also the fundamental building blocks of the QoS implementation in the 7x50. Ingress packets, classified by various rules, belong to one of eight Forwarding Classes (FC). An FC can be viewed a set of packets which are treated in a similar manner within the system (have the same priority level and scheduling behavior). Each FC has a queue associated with it and each queue has a set of parameters controlling how buffer memory is allocated for packets; if a packet is enqueued (placed on the queue) a scheduler will control the way the packet gets dequeued (removed from the queue) relative to other queues. When a packet exits an egress port, a remarking action can be taken to guarantee the next downstream device will properly handle the different types of traffic.

Configuration

Policies

QoS policies are used to control how traffic is handled at distinct points in the service delivery model within the device. There are different types of QoS policies catering to the different QoS needs at each point. QoS policies only take effect when applied to a relevant entity (Service Access Point (SAP) or network port/interface) so by default can be seen as templates with each application instantiating a new set of related resources.

The following QoS policies are discussed:

- Ingress/egress QoS Policies — For classification, queue attributes and remarking.
- Slope policies — Define the RED slope definitions.
- Scheduler policies — Determine how queues are scheduled (only the default scheduling is included here).

Access, Network, and Hybrid Ports

The system has two different types of interfaces: access and network.

- A network interface will classify packets received from the network core at ingress and remark packets sent to the core at egress. Aggregated differentiated service QoS is performed on network ports, aggregating traffic from multiple services into a set of queues per FC.
- An access interface connects to one or more customer devices; it receives customer packets, classifies them into different FCs at ingress and remarks packets according to FCs at egress. Since an access interface needs application awareness, it has many more rules to classify the ingress packets. Access and network also differ in how buffer memory is handled, as will be made clear when discussing the buffer management. Here the QoS is performed per SAP.

Access interfaces (SAPs) are configured on access ports and network interfaces are configured on network ports. A third type of port is available, the hybrid port, which supports both access and network interfaces on the same port.

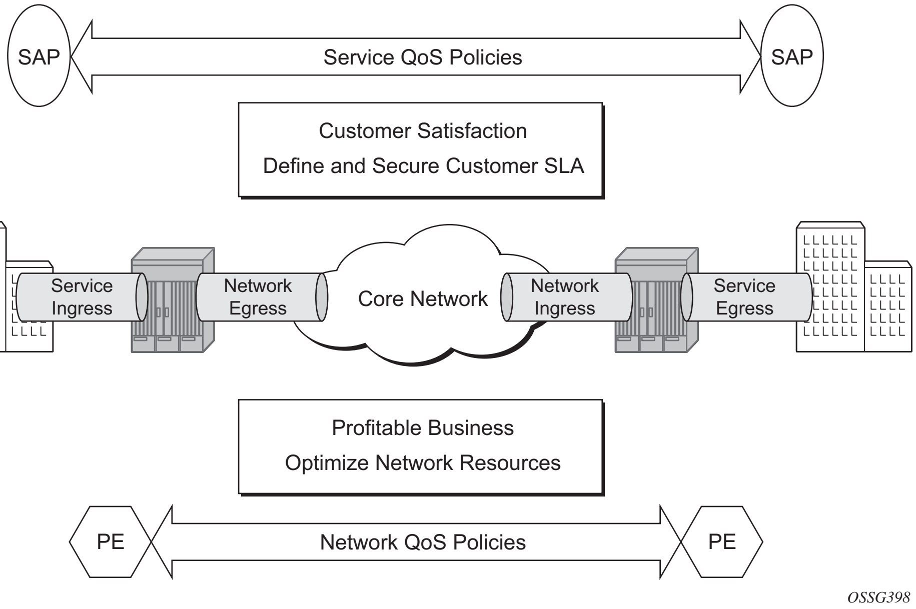
Hybrid ports are only supported on Ethernet ports and optionally with a single-chassis LAG. They must be configured to use VLANs (either single (dot1q encapsulation) or double (QinQ encapsulation) tagging) with each VLAN mapping to either the access or network part of the port. This allows the classification to associate incoming traffic with the correct port type and service. Port based traffic, such as LACP, CCM and EFM, uses a system queue on an access port, but the default network queues on a network or hybrid port.

Customer traffic follows the path shown below:

[service ingress ➔ network egress]	➔	[network ingress ➔ network egress] ➔	[network ingress ➔ service egress]
ingress PE		transit P	egress PE

The network administrator needs to make sure that QoS is configured correctly at each point using the appropriate QoS policies (Figure 226).

Figure 226 **Service and Network QoS Policies**



Service Ingress QoS Policy

The SAP ingress policies are created in the *qos* context of the CLI and require a unique identifier (from 1 to 65535). The default *sap-ingress* policy has identifier 1.

Classification

Services can be delineated at the SAP ingress by

- A physical port (null encapsulated) or
- An encapsulation on the physical port, for example a VLAN ID on an Ethernet port or a DLCI on a Frame Relay port.

The following configuration is an example of an IES service created with an IP interface on VLAN 2 of port 3/2/10 (IOM 3, MDA 2, port 10) and has SAP ingress QoS policy 10 applied.

```
configure
```

```
service
    ies 1 customer 1 create
        interface "int-access" create
            address 192.168.1.1/30
            sap 1/1/1:1 create
                ingress
                    qos 10
                exit
            exit
        exit
        no shutdown
    exit
```

As traffic enters the port, the service can be identified by the VLAN tag (and unwanted packets dropped). The ingress service QoS policy applied to the SAP maps traffic to FCs, and thus to queues, and sets the enqueuing priority. Mapping flows to FCs is controlled by comparing each packet to the match criteria in the QoS policy. The match criteria that can be used in ingress QoS policies can be combinations of those listed in Table 24. When a packet matches two criteria (802.1p priority and DSCP) it is the lowest precedence value that is used to map the packet to the FC.

Table 24 **SAP Ingress Classification Match Criteria**

Match Precedence	Match Criteria		
1	IPv4 fields match criteria: • Destination IP address/prefix including prefix list • Destination port/range • DSCP value • IP fragment • Protocol type (TCP, UDP, etc.) • Source port/range • Source IP address/prefix including prefix list	IPv6 fields match criteria: • Destination IP address/prefix • Destination port/range • DSCP value • Next header • Source port/range • Source IP address/prefix	MAC fields match criteria: • Frame type [802dot3|802dot2-llc|802dot2-snap|ethernetII|atm] • ATM VCI value • IEEE 802.1p value/mask • Source MAC address/mask • Destination MAC address/mask • EtherType value • IEEE 802.2 LLC SSAP value/mask • IEEE 802.2 LLC DSAP value/mask • IEEE 802.3 LLC SNAP OUI zero or non-zero value • IEEE 802.3 LLC SNAP PID value
	Note: For an ingress QoS policy, either IP match criteria or MAC match criteria can be defined, not both.		
2	DSCP		

Table 24 **SAP Ingress Classification Match Criteria**

Match Precedence	Match Criteria
3	IP precedence
4	LSP EXP
5	IEEE 802.1p priority and/or Drop Eligibility Indicator (DEI)
6	Default forwarding class for non-matching traffic

It is possible to match MAC criteria on VPLS/Epipe SAPs and IP criteria on IP interface SAPs. However, it is also possible to classify on MAC criteria on an IP interface SAP and conversely to classify on IP criteria on VPLS/Epipe SAPs. When MPLS labeled traffic is received on a VPLS/Epipe SAP, it is possible to match on either of the LSP EXP bits (outer label) or the MAC criteria.

A SAP can be configured to have no VLAN tag (null encapsulated), one VLAN tag (dot1q encapsulated) or two VLAN tags (QinQ encapsulated). The configuration allows the selection of which VLAN tag to match against for the 802.1p bits, using the keyword **match-qinq-dot1p** with the keyword **top** or **bottom**.

The following example configuration shows match QinQ traffic with dot1p value 1 in the top VLAN tag entering the QinQ SAP in Epipe service 1 and assign it to FC **af** using queue 2.

```
configure
    qos
        sap-ingress 10 create
            queue 2 create
            exit
            fc "af" create
                queue 2
            exit
            dot1p 1 fc "af"
        exit


configure
    service
        epipe 2 customer 1 create
            sap 1/1/2:1.2 create
                ingress
                    qos 10
                exit
                ingress
                    match-qinq-dot1p top
                exit
            exit
            no shutdown
        exit
```

The classification of traffic using the default, **top** and **bottom** keyword parameters is summarized in Table 25. A TopQ SAP is a QinQ SAP where only the outer (top) VLAN tag is explicitly specified (sap 1/1/1:10.* or sap 1/1/1:10.0).

Table 25 **QinQ Dot1p Bit Classification**

Port/SAP Type	Existing Packet Tags	Pbits Used for Match		
		Default	Match Top	Match Bottom
Null	None	None	None	None
Null	Dot1P (VLAN-ID 0)	Dot1P PBits	Dot1P PBits	Dot1P PBits
Null	Dot1Q	Dot1Q PBits	Dot1Q PBits	Dot1Q PBits
Null	TopQ BottomQ	TopQ PBits	TopQ PBits	BottomQ PBits
Null	TopQ (No BottomQ)	TopQ PBits	TopQ PBits	TopQ PBits
Dot1Q	None (Default SAP)	None	None	None
Dot1Q	Dot1P (Default SAP VLAN-ID 0)	Dot1P PBits	Dot1P PBits	Dot1P PBits
Dot1Q	Dot1Q	Dot1Q PBits	Dot1Q PBits	Dot1Q PBits
QinQ/TopQ	TopQ	TopQ PBits	TopQ PBits	TopQ PBits
QinQ/TopQ	TopQ BottomQ	TopQ PBits	TopQ PBits	BottomQ PBits
QinQ/QinQ	TopQ BottomQ	BottomQ PBits	TopQ PBits	BottomQ Pbits

The Drop Eligibility Indicator (DEI) bit (IEEE 802.1ad-2005 and IEEE 802.1ah (PBB)) can be used to indicate the in/out profile state of the packet, this will be covered later in the discussion on profile mode.

Ingress traffic with a local destination (for example, OSPF hellos) is classified by the system automatically and uses a set of dedicated system queues.

After the traffic has been classified, the next step is to assign it to a given FC. There are 8 pre-defined FCs within the system which are shown in Table 26 (the FC identifiers are keywords and do not have a fixed relationfship with the associated Differentiated Services Code Points (DSCP)).

Table 26 **Forwarding Classes**

FC Identifier	FC Name	Default Scheduling Priority
NC	Network Control	Expedited

Table 26 **Forwarding Classes**

FC Identifier	FC Name	Default Scheduling Priority
H1	High-1	Expedited
EF	Expedited	Expedited
H2	High-2	Expedited
L1	Low-1	Best Effort
AF	Assured	Best Effort
L2	Low-2	Best Effort
BE	Best Effort	Best Effort

When an FC is configured for a classification, it must first be created in the configuration. One of the FCs can be also configured to be the default in case there is no explicit classification match and by default this FC is **be**.

Normally, once traffic is assigned to an FC at the ingress it remains in that FC throughout its time within the system. Re-classification of IP traffic at a SAP egress is possible, but is beyond the scope of this chapter. The FC used at egress can also be specified to be different than that used at ingress by configuring **egress-fc** under the FC configuration in the SAP ingress policy.

Packets also have a state of being in-profile or out-of-profile which represents their drop precedence within the system, therefore there can be up to 8 distinct per hop behavior (PHB) classes with two drop precedences.

Buffering (Enqueuing)

Once a packet is assigned to a certain forwarding class, it will try to get a buffer in order to be enqueued. Whether the packet can get a buffer is determined by the instantaneous buffer utilization and several attributes of the queue (such as Maximum Burst Size (MBS), Committed Burst Size (CBS) and high-prio-only) that will be discussed in more detail later in this chapter. If a packet cannot get a buffer for whatever reason, the packet will get dropped immediately.

As traffic is classified at the SAP ingress it is also assigned an enqueuing priority, which can be high or low. This governs the likelihood of a packet being accepted into a buffer and so onto a queue, and is managed using the queue's high-prio-only parameter and the buffer pools weighted random early detection (WRED) slope policies. Traffic having a high enqueuing priority has more chance of getting a buffer than traffic with low enqueuing priority. The enqueuing priority is specified with the classification using the parameter *priority*, and a default enqueuing priority can be configured, its default being low.

Enqueuing priority is a property of a packet and should not to be confused with scheduling priority, expedited or best-effort, which is a property of a queue.

The following configuration shows an example where all packets with dot1p value 3 are classified as **ef** and have their enqueuing priority set to high, while all other packets are classified as **af** with a low enqueuing priority.

```
configure
    qos
        sap-ingress 10 create
            fc "af" create
            exit
            fc "ef" create
            exit
            dot1p 3 fc "ef" priority high
            default-fc "af"
            default-priority low # this is the default
        exit
```

Each forwarding class is associated with at most one unicast queue. In the case of a VPLS service, each FC can also be assigned a single multipoint queue at ingress, or for more granular control, separate queues for broadcast, multicast and unknown traffic. Since each queue maintains forward/drop statistics, it allows the network operator to easily track unicast, broadcast, multicast and unknown traffic load per forwarding class. Separate multicast queues can also be assigned for IES/VPRN services which have IP multicast enabled.

This results in an Epipe SAP having up to 8 ingress queues, an IES/VPRN SAP having up to 16 ingress queues and a VPLS SAP having up to 32 ingress queues. Each queue has a locally significant (to the policy) identifier, which can be from 1 to 32.

The default SAP ingress QoS policy (id=1) has two queues; queue 1 for unicast traffic and queue 11 for multipoint traffic, and is assigned to every ingress SAP at service creation time. Equally, when a new (non-default) SAP ingress policy is created, queue 1 and queue 11 are automatically created with all FCs assigned to both by default. Additional queues must be created before being assigned to a FC, with multipoint queues requiring the **multipoint** keyword. When a SAP ingress policy is applied to a SAP, physical hardware queues on the FP are allocated for each queue with a FC assigned (if no QoS policy is explicitly configured, the default policy is applied). Multipoint queues within the SAP ingress policy are ignored when applied to an Epipe SAP or an IES/VPRN SAP which is not configured for IP multicast.

The mechanism described here uses a separate set of queues per SAP. For cases where per-SAP queuing is not required it is possible to use port based queues, known as *queue-groups*, which reduces the number of queues required, as described in chapter *FP and Port Queue Groups*.

Scheduling (Dequeuing)

A queue has a priority which affects the relative scheduling of packets from it compared to other queues. There are two queue priorities: expedited and best-effort, with expedited being the higher. When creating a queue, one of these priorities can be configured, thereby explicitly setting the queue’s priority. Alternatively, the default is auto-expedite in which case the queue’s priority is governed by the FCs assigned to it, as shown in Table 26. If there is a mix of expedited and best-effort FCs assigned, the queue is deemed to be best-effort.

The following configuration displays an example that ensures that EF traffic is treated as expedited by assigning it to new unicast and multicast queues.

```
configure
    qos
        sap-ingress 10 create
            queue 3 expedite create
            exit
            queue 13 multipoint expedite create
            exit
            fc "ef" create
                queue 3
                multicast-queue 13
            exit
            default-fc "ef"
        exit
```

Once a packet gets a buffer and is queued, it will wait to be served and sent through the switch fabric to its destination port by the hardware scheduler. There are two scheduler priorities: expedited or best-effort, corresponding to the queue’s priority. The expedited hardware schedulers are used to enforce priority access to internal switch fabric destinations with expedited queues having a higher preference than best-effort queues. Queues of the same priority get equally serviced in round robin fashion by the associated scheduler.

When a queue gets its turn to be serviced, the scheduler will use the operational Peak Information Rate (PIR) and Committed Information Rate (CIR) attributes of the queue to determine what to do with the packet.

- The scheduler does not allow queues to exceed their configured PIR. If the packet arrival rate for a given queue is higher than the rate at which it is drained, the queue will fill. If the queue size (in bytes or Kbytes) reaches its defined MBS all subsequent packets will be discarded, this is known as tail drop.
- If the dequeue rate is below or equal to the operational CIR, the packet will be forwarded and marked as **in-profile**.
- If the dequeue rate is below or equal to the operational PIR but higher than the CIR, the packet will be forwarded but marked as **out-of-profile**.

Out-of-profile packets have a higher probability of being dropped when there is congestion somewhere in the downstream path. Packets that are marked as out-of-profile will also be treated differently at the network egress and service egress.

These marking actions are known as color marking (green for in-profile and yellow for out-of-profile). Using the default queue setting of **priority-mode**, as described above, the in/out-of-profile state of a packet is determined from the queue scheduling state (within CIR or above CIR, as described later) at the time that the packet is dequeued. An alternative queue mode is **profile-mode**.

Profile Mode

A queue is created with profile mode when the aim is that the in/out-of-profile state of packets is determined by the QoS bits of the incoming packets, this is known as color-aware (as opposed to color-unaware for priority mode).

As part of the classification, the profile state of the packets is explicitly configured. To provide granular control, it is possible to configure FC sub-classes with each having a different profile state, while inheriting the other parameters from their parent FC (for example the queue, in order to avoid out of order packets). The FC subclasses are named *fc.sub-class*, where *sub-class* is a text string up to 29 characters (though normally the words **in** and **out** are used for clarity). Any traffic classified without an explicit profile state is treated as if the queue were in priority mode.

When using the profile mode, the DEI in the Ethernet header can be used to classify a packet as in-profile (DEI=0) or out-of-profile (DEI=1).

The following configuration shows traffic with dot1p 3 is set to in-profile, dot1p 2 to out-of-profile and the profile state of dot1p 0 depends on the scheduling state of the queue.

```
configure
    qos
        sap-ingress 20 create
            queue 2 profile-mode create
            exit
            fc "af" create
                queue 2
            exit
            fc "af.in" create
                profile in
            exit
            fc "af.out" create
                profile out
            exit
            dot1p 0 fc "af"
            dot1p 2 fc "af.out"
            dot1p 3 fc "af.in"
        exit
```

The difference between a queue configured in priority (default) and profile mode is summarized in Table 27 (within/above CIR is described later).

Table 27 **Queue Priority vs. Profile Mode**

	Priority Mode	**Profile Mode**
Packet In-Profile/ Out-of-Profile state	Determined by state of the queue at scheduling time. Within CIR – In Profile Above CIR – Out Profile	Explicitly stated in FC or subclass classification. If not, then defaults to state of the queue at scheduling time

Packet High/Low Enqueuing Priority	Explicitly stated in FC classification. If not, then defaults to Low priority	Always follows state of in-profile/out-of-profile determined above
		In-profile = High Priority
		Out-Profile = Low Priority
		If not set = High Priority

Remarking

Remarking at the service ingress is possible when using an IES or VPRN service. The DSCP/precedence field can be remarked for in-profile (**in-remark**) and out-of-profile (**out-remark**) traffic as defined above for queues in either priority mode or profile mode. If configured for other services, the remarking is ignored. If remarking is performed at the service ingress, then the traffic is not subject to any egress remarking on the same system.

The following configuration displays an example classifying traffic to 10.0.0.0/8 as FC **ef** in-profile and remark its DSCP to **ef**.

```
configure
    qos
        sap-ingress 30 create
            queue 2 profile-mode create
            exit
            fc "ef" create
                queue 2
                in-remark dscp ef
                profile in
            exit
            ip-criteria
                entry 10 create
                    match
                        dst-ip 10.0.0.0/8
                    exit
                    action fc "ef"
                exit
            exit
        exit
```

Service Egress QoS Policy

The service egress uses a SAP egress QoS policy to define how FCs map to queues and how a packet of an FC is remarked. SAP egress policies are created in the CLI qos context and require a unique identifier (from 1 to 65535). The default SAP egress policy has identifier 1.

Once a service packet is delivered to the egress SAP, it has following attributes:

- Forwarding class, determined from classification at the ingress of the node.
- In/out-of-profile state from the service ingress or network ingress.

Similar to the service ingress enqueuing process, it is possible that a packet cannot get a buffer and thus gets dropped. Once on an egress queue, a packet is scheduled from the queue based on priority of the queue (expedited or best-effort) and the scheduling state with respect to the CIR/PIR rates (the profile state of the packet [in/out] is not modified here). Egress queues do not have a priority/profile mode and have no concept of multipoint.

Only one queue exists in the default SAP egress QoS policy (id=1) and also when a new *sap-egress* policy is created, this being queue 1 which is used for both unicast traffic and multipoint traffic. All FCs are assigned to this queue unless otherwise explicitly configured to a different configured queue. When a SAP egress policy is applied to a SAP, physical hardware queues on the FP are allocated for each queue with FC assigned (if no QoS policy is explicitly configured, the default policy is applied).

As mentioned earlier, re-classification at a SAP egress is possible based on the packet's dot1p, DSCP or precedence values or using IP or IPv6 criteria matching, similar to the functionality at SAP ingress.

SAP egress also supports two additional profiles, inplus-profile and exceed-profile. Both can be assigned to a packet using egress reclassification and the exceed-profile can be assigned to a packet in an egress policer configured with the **enable-exceed-pir** command.

Traffic originated by the system (known as self generated traffic) has its FC and marking configured under router/sgt-qos (for the base routing) or under service/vprn/sgt-qos (for a VPRN service). This is beyond the scope of this chapter.

Remarking

At the service egress, the dot1p/DEI can be remarked for any service per FC with separate marking for in/out/exceed profile if required (inplus-profile packets are marked with the same value as in-profile packets and exceed-profile packets are marked with the same value as out-of-profile packets if not explicitly configured). The DEI bit can also be forced to a specific value (using the **de-mark force** command). When no **dot1p/de-mark** is configured, the ingress dot1p/DEI is preserved; if the ingress was untagged, the dot1p/DEI bit is set to 0.

The following configuration shows a remark example with different FCs with different dot1p values. FC **af** also differentiates between in/out-of-profile and then remarks the DEI bit accordingly based on the packet's profile.

```
configure
    qos
        sap-egress 10 create
            queue 1 create
                rate 20000
            exit
            queue 2 create
                rate 10000 cir 5000
            exit
            queue 3 create
                rate 2000 cir 2000
            exit
            fc af create
                queue 2
                dot1p in-profile 3 out-profile 2
                de-mark
            exit
            fc be create
                queue 1
                dot1p 0
            exit
            fc ef create
```

```
                queue 3
                dot1p 5
            exit
        exit
```

If QinQ encapsulation is used, the default is to remark both tags in the same way. However, it is also possible to remark only the top tag using the **qinq-mark-top-only** parameter configured under the SAP egress.

The following configuration shows a remark example with only the dot1p/DEI bits in top tag of a QinQ SAP.

```
configure
    service
        vpls 3 customer 1 create
            sap 1/1/2:2.2 create
                egress
                    qos 20
                    qinq-mark-top-only
                exit
            exit
            no shutdown
        exit
```

For IES and VPRN services, the DSCP/precedence field can also be remarked based on the in/out/exceed-profile state (with inplus-profile packets marked with the same values as in-profile packets, and exceed-profile packets marked with the same value as out-of-profile packets if not explicitly configured) of the packets (and only if no ingress remarking was performed).

The following configuration shows DSCP values for FC **af** based on in/out-of-profile traffic.

```
configure
    qos
        sap-egress 20 create
            queue 2 create
            exit
            fc af create
                queue 2
                dscp in-profile af41 out-profile af43
            exit
        exit
```

Network Ports

The QoS policies relating to the network ports are divided into a network and a network-queue policy. The network policy covers the ingress and egress classification into FCs and the egress remarking based on FCs, while the network-queue policy covers the queues parameters and the FC to queue mapping. The logic behind this is that there is only one set of queues provisioned on a network port, whereas the use of these queues is configured per network IP interface. This in turn determines where the two policies can be applied. Network ports are used for IP routing and switching, and for GRE/MPLS tunneling.

Network QoS Policy

The network QoS policy has an ingress section and an egress section. It is created in the *qos* context of the CLI and requires a unique identifier (from 1 to 65535). The default network policy has identifier 1. Network QoS policies are applied to IP interfaces configured on a network port.

The following configuration shows an example to apply different network QoS policies to two network interfaces.

```
configure
    router
        interface "int-PE-1-PE-2"
            address 192.168.12.1/30
            port 1/1/3:1
            qos 28
        exit
        interface "int-PE-1-PE-3"
            address 192.168.13.1/30
            port 1/1/4
            qos 18
        exit
```

Classification

Classification is available at both ingress and egress.

The ingress section defines the classification rules for IP/MPLS packets received on a network IP interface. The rules for classifying traffic are based on the incoming QoS bits (Dot1p, DSCP, EXP [MPLS experimental bits]). The order in which classification occurs relative to these fields is:

1. IPv4 and IPv6 match criteria for IP packets

2. EXP (for MPLS packets) or DSCP (for IP packets)
   Dot1p/DEI bit (network ports do not support QinQ encapsulation)
3. default action (default = fc be profile out)

The configuration specifies the QoS bits to match against the incoming traffic together with the FC and profile (in/out) to be used (it is analogous to the SAP profile-mode in that the profile of the traffic is determined from the incoming traffic, rather than the CIR configured on the queue). A **default-action** keyword configures a default FC and profile state.

The IPv4 and IPv6 criteria matching only applies to the outer IP header of *non-tunneled* traffic, except for traffic received on an RFC 6037 MVPN tunnel for which classification on the outer IP header only is supported, and is only supported on network interfaces.

For tunneled traffic (GRE or MPLS), the match is based on the outer encapsulation header unless the keyword **ler-use-dscp** is configured. In this case, traffic received on the router terminating the tunnel that is to be routed to the base router or a VPRN destination is classified based on the encapsulated packet DSCP value (assuming it is an IP packet) rather than its EXP bits.

The ability exists for an egress LER to signal an implicit-null label (numeric value 3). This informs the previous hop to send MPLS packets without an outer label and so is known as penultimate hop popping (PHP). This can result in MPLS traffic being received at the termination of an LSP without any MPLS labels. In general, this would only be the case for IP encapsulated traffic, in which case the egress LER would need to classify the incoming traffic using IP criteria.

The egress section also defines the classification rules based on both the DSCP and precedence values in a packet to re-assign the packet’s FC and profile (inplus/in/out/exceed).

Remarking

The egress section of the network policy defines the remarking of the egress packets, there is no remarking possible at the network ingress. The egress remarking is configured per FC and can set the related dot1p/DEI (explicitly or dependent on in/out-of-profile), DSCP (dependent on in/out-of-profile) and EXP (dependent on in/out-of-profile; inplus-profile packets are marked with the same values as in-profile packets and exceed-profile packets are marked with the same value as out-of-profile packets).

The traffic exiting a network port is either tunneled (in GRE or MPLS) or IP routed.

For tunneled traffic exiting a network port, the remarking applies to the DSCP/EXP bits in any tunnel encapsulation headers (GRE/MPLS) pushed onto the packet by this system, together with the associated dot1p/DEI bits if the traffic has an outer VLAN tag. For MPLS tunnels, the EXP bits in the entire label stack are remarked.

Note: Strictly speaking this is marking (as opposed to remarking) as the action is adding QoS information rather than changing it.

A new outer encapsulation header is pushed onto traffic at each MPLS transit label switched router as part of the label swap operation.

For VPLS/Epipe services no additional remarking is possible. However, for IES/VPRN/base-routing traffic, the remarking capabilities at the network egress are different at the first network egress (egress on the system on which the traffic entered by a SAP ingress) and subsequent network egress in the network (egress on the systems on which the traffic entered through another network interface).

At the first network egress, the DSCP of the routed/tunneled IP packet can be remarked, but this is dependent on two configuration settings:

- The trusted state of the ingress (service/network) interface and
- The **remarking** keyword in the network QoS policy at the network egress. The configuration combinations are summarized in Table 28.

This is in addition to the remarking of any encapsulation headers and, as stated earlier, is not performed if the traffic was remarked at the service ingress.

For traffic exiting a subsequent network egress in the network, only the IP routed traffic can be remarked, again this is dependent on the ingress trusted state and egress remarking parameter.

There is one addition to the above to handle the marking for IP-VPN Option-B in order to remark the EXP, DSCP and dot1p/DEI bits at a network egress, this being **remarking force**. Without this, only the EXP and dot1p/DEI bits are remarked. This does not apply to label switched path traffic switched at a label switched router.

Table 28 **Network QoS Policy DSCP Remarking**

Ingress	Trusted State	Remarking Configuration	Marking Performed
IES	Untrusted (default)	remarking	Yes
		no remarking (default)	Yes
	Trusted	remarking	Yes
		no remarking (default)	No

Table 28 Network QoS Policy DSCP Remarking (Continued)

Ingress	Trusted State	Remarking Configuration	Marking Performed
Network	Untrusted	remarking	Yes
		no remarking (default)	Yes
	Trusted (default)	remarking	Yes
		no remarking (default)	No
VPRN	Untrusted	remarking	Yes
		no remarking (default)	Yes
	Trusted (default)	remarking	No
		no remarking (default)	No

The following configuration shows a ingress network classification for DSCP EF explicitly, with a default action for the remainder of the traffic and use the DSCP from the encapsulated IP packet if terminating a tunnel. Remark the DSCP values for FC **af** and **ef** and remark all traffic (except incoming VPRN traffic) at the egress. Apply this policy to a network interface.

```
configure
    qos
        network 20 create
            ingress
                default-action fc af profile out
                ler-use-dscp
                dscp ef fc ef profile in
            exit
            egress
                remarking
                fc af
                    no dscp-in-profile
                    dscp-out-profile af13
                    lsp-exp-in-profile 6
                    lsp-exp-out-profile 5
                exit
                fc ef
                    dscp-in-profile af41
                exit
            exit
        exit

configure
    router
        interface "int-PE-1-PE-4"
            address 192.168.14.1/30
            port 1/2/1
            qos 20
        exit
```

The following configuration shows the trusted IES interface.

```
configure
    service
        ies 1 customer 1 create
            interface "int-access" create
                address 192.168.1.1/30
                tos-marking-state trusted
                sap 1/1/1:1 create
                exit
            exit
            no shutdown
        exit
```

The network QoS ingress and egress sections also contain the configuration for the use of FP-based policers and port-based queues by queue-groups which are out of scope of this chapter.

Network Queue Policy

The network queue QoS policy defines the queues and their parameters together with the FC to queue mapping. The policies are named, with the default policy having the name **default**. Network queues policies are applied under **config > card > mda > network > ingress** for the network ingress queues though only one policy is supported per MDA, so when a new policy is applied under one MDA, it is automatically applied under the other MDA on the same FP. At egress, network queue policies are applied under Ethernet: **config > port > ethernet > network**, POS: **config > port > sonet-sdh > path > network**, TDM: **config > port > tdm > e3 | ds3 > network** for the egress.

The following configuration shows an ingress and egress network-queue policy.

```
configure
    card 1
        card-type iom3-xp
        mda 1
            mda-type m4-10gb-xp-xfp
            network
                ingress
                    queue-policy "network-queue-1"
                exit
            exit
            no shutdown
        exit

configure
    port 1/1/3
        ethernet
            encap-type dot1q
            network
```

```
                queue-policy "network-queue-1"
            exit
        exit
        no shutdown
    exit
```

Up to 16 queues can be configured in a network-queue policy, each with a queue-type of best-effort, expedite, or auto-expedite. A new network-queue policy contains two queues, queue 1 for unicast traffic and queue 9 for multipoint traffic and by default all FCs are mapped to these queues. There is no differentiation for broadcast, multicast and unknown traffic. If the policy is applied to the egress, then any multipoint queues are ignored. As there are 8 FCs, there would be up to 8 unicast queues and 8 multipoint queues, resulting in 16 ingress queues and 8 egress queues. Normally, the network queue configuration is symmetric (the same queues/FC-mapping at the ingress and egress).

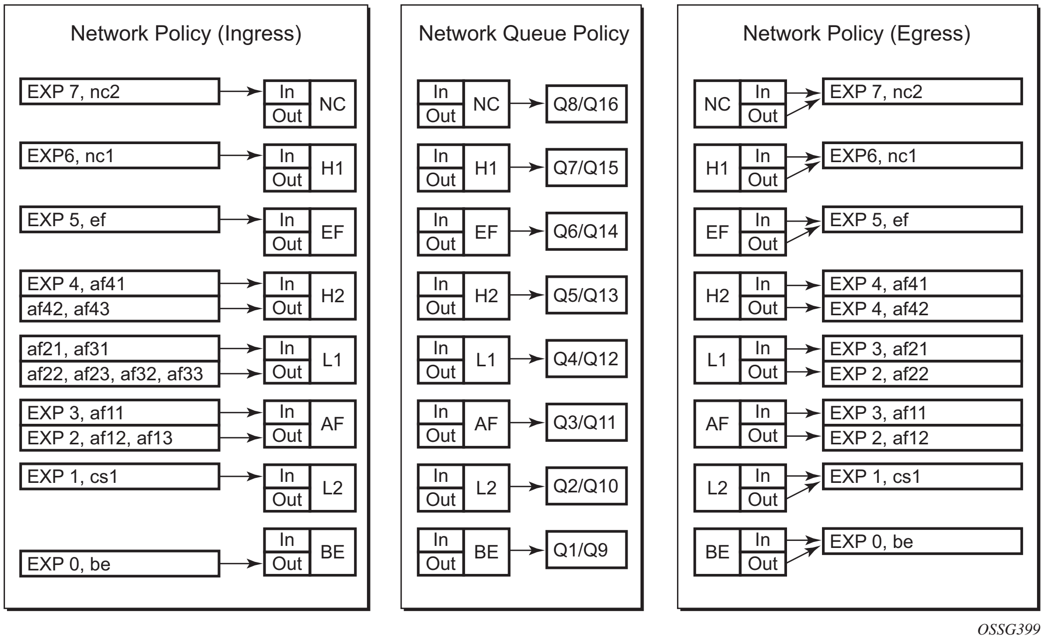
The following configuration defines a network-queue policy with FC **af** and **ef** assigned to queues 2 and 3 for unicast traffic, and queue 9 for multipoint traffic.

```
configure qos
        network-queue "network-queue-1" create
            queue 1 create
                mbs 50
                high-prio-only 10
            exit
            queue 2 create
            exit
            queue 3 create
            exit
            queue 9 multipoint create
                mbs 50
                high-prio-only 10
            exit
            fc af create
                multicast-queue 9
                queue 2
            exit
            fc ef create
                multicast-queue 9
                queue 3
            exit
        exit
```

Summary of Network Policies

Figure 227 displays the default network policies with respect to classification, FC to queue mapping and remarking.

Figure 227 Visualization of Default Network Policies



Queue Management

The policies described so far define queues but not the characteristics of those queues which determine how they behave. This section describes the detailed configuration associated with these queues. There are two aspects:

- Enqueuing packets onto a queue
  - buffer pools
  - queue sizing
  - Weighted Random Early Detection (WRED)
- Dequeuing packets from a queue
  - queue rates
  - scheduling

Enqueuing Packets: Buffer Pools

The packet buffer space is divided equally between ingress and egress. For line cards using a 50G FP2 for both ingress and egress traffic, the proportion of ingress versus egress buffer space can be modified using the following command:

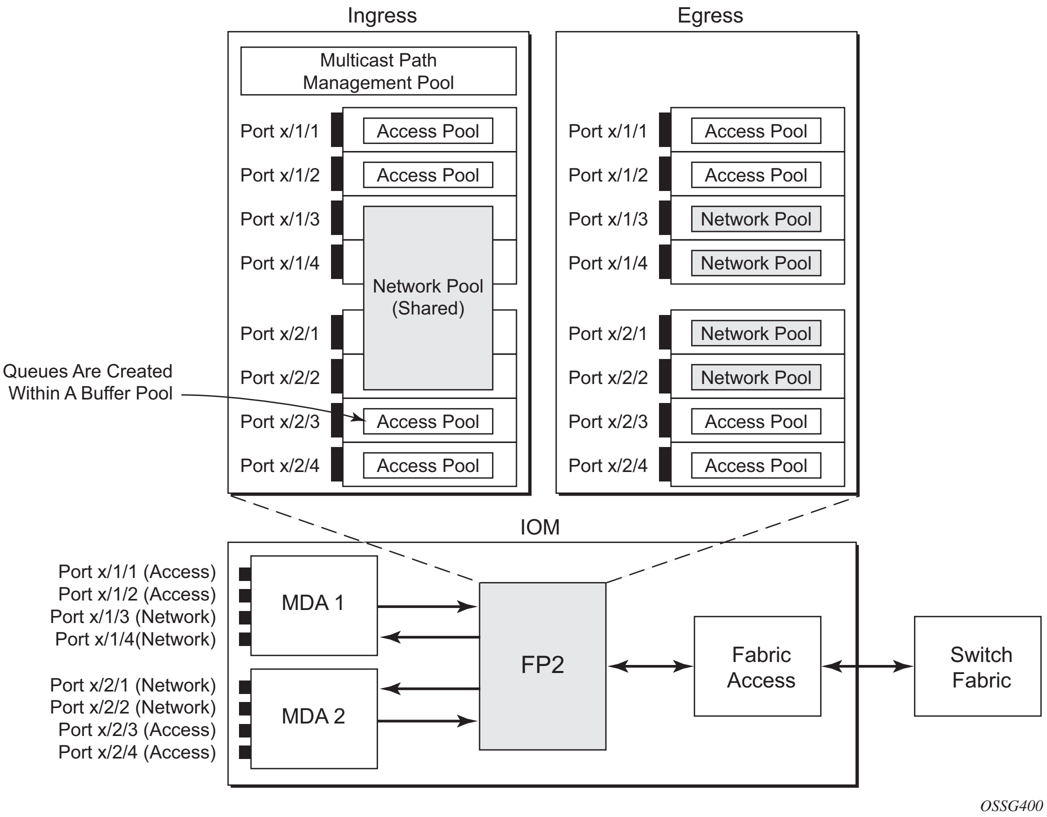
```
configure
    card <slot-number>
        ingress-buffer-allocation <percentage>
```

The ingress buffer allocation percentage can be configured from 20% up to 80%.

Beyond that, by default there is one pool for network ingress per FP2/IOM, with one pool per access ingress port and one pool per access/network egress port. This is shown in Figure 228. This segregation provides isolation against buffer starvation between the separate pools. An additional ingress pool exists for managed multicast traffic (the multicast path management pool) but this is beyond the scope of this chapter.

The buffer management can be modified using named buffer pools and/or WRED-per-queue pools which are out of scope of this chapter.

Figure 228 **Default Buffer Pools**



The size of the pools is based on the MDA type and the speed/type (access or network) of each port. Buffer space is allocated in proportion to the active bandwidth of each port, which is dependent on:

- The actual speed of the port
- Bandwidth for configured channels only (on channelized cards)
- Zero for ports without queues configured

This calculation can be tuned separately for ingress and egress, without modifying the actual port speed, using the port/modify-buffer-allocation-rate. Changing the port’s ingress-rate or egress-rate will also modify its buffer sizes.

The following configuration changes the relative size for the ingress/egress buffer space on port 1/1/10 to 50% of the default.

```
configure
    port 1/1/3
        modify-buffer-allocation-rate
            ing-percentage-of-rate 50
            egr-percentage-of-rate 50
        exit
```

Each of the buffer pools created is further divided into a section of reserved buffers and another of shared buffers, see Figure 230. The amount of reserved buffers is calculated differently for network and access pools. For network pools, the default is approximately the sum of the CBS (committed burst size) values defined for all of the queues within the pool. The reserved buffer size can also be statically configured to a percentage of the full pool size (ingress: **config > card > mda > network > ingress > pool**; egress: **config > port > network > egress > pool**). For access pools, the default reserved buffer size is 30% of the full pool size and can be set statically to an explicit value (ingress: **config > port > access > ingress > pool**; egress: **config > port > access > egress > pool**).

The following configuration sets the reserved buffer size to 50% of the egress pool space on a network port.

```
configure
    port 1/1/3
        network
            egress
                pool
                    resv-cbs 50
                exit
            exit
        exit
```

On an access port, the reserved buffer size is set to 50% of the egress pool space, as follows:

```
configure
    port 1/1/1
        access
            egress
                pool
                    resv-cbs 50
                exit
            exit
        exit
```

Both the total buffer and the reserved buffer sizes are allocated in blocks (discrete values of Kbytes). The pool sizes can be seen using the **show pools** command.

It is possible to configure alarms to be triggered when the usage of the reserved buffers in the buffer pools reaches a certain percentage. Two alarm percentages are configurable, amber and red, **amber-alarm-threshold** <*percentage*> and **red-alarm-threshold** <*percentage*>. The percentage range is 1 — 1000.

- The percentage for the red must be at least as large as that for the amber.
- The alarms are cleared when the reserved CBS drops below the related threshold.
- When the amber alarm is enabled, dynamic reserved buffer sizing can be used; after the amber alarm is triggered the reserved buffer size is increased or decreased depending on the CBS usage. This requires a non-default resv-cbs to be configured together with a step and max value for the amber-alarm-action parameters. As the reserved CBS usage increases above the amber alarm percentage, the reserved buffer size is increased in increments defined by the step, up to a maximum of the max. If the CBS usage decreases, the reserved buffer size is reduced in steps down to its configured size.
- As the reserved buffer size changes, alarms will continue to be triggered at the same color (amber or red) indicating the new reserved buffer size. The pool sizing is checked at intervals, so it can take up to one minute for the alarms and pool re-sizing to occur.

The following displays a configuration for access ingress and egress pools.

```
configure
    port 1/1/1
        access
            ingress
                pool
                    amber-alarm-threshold 25
                    red-alarm-threshold 50
                    resv-cbs 20 amber-alarm-action step 5 max 50
                exit
            exit
            egress
                pool
                    amber-alarm-threshold 25
                    red-alarm-threshold 25
                    resv-cbs 20 amber-alarm-action step 5 max 50
                exit
            exit
        exit
```

The following is an example alarm that is triggered when the amber percentage has been exceeded and the reserved buffer size has increased from 20% to 25%:

```
82 2016/04/25 14:21:52.42 UTC MINOR: PORT #2050 Base Resv CBS Alarm
"Amber Alarm: CBS over Amber threshold: ObjType=port Owner=1/1/1 Type=accessIngress
 Pool=default NamedPoolPolicy= ResvSize=672 SumOfQ ResvSize=138 Old ResvCBS=20
New ResvCBS=25 Old ResvSize=528"
```

When a port is configured to be a hybrid port, its buffer space is divided into an access portion and a network portion. The split by default is 50:50 but it can be configured on a per port basis.

```
configure port 1/1/1
        ethernet
            mode hybrid
            encap-type dot1q
        exit
        hybrid-buffer-allocation
            ing-weight access 70 network 30
            egr-weight access 70 network 30
        exit
```

Enqueing Packets: Queue Sizing

Queue sizes change dynamically when packets are added to a queue faster than they are removed, without any traffic the queue depth is zero. When packets arrive for a queue there will be request for buffer memory which will result in buffers being allocated dynamically from the buffer pool that the queue belongs to.

A queue has four buffer size related attributes: MBS, CBS, high-prio-only, and hi-low-prio-only, which affect packets only during the enqueuing process.

- Maximum Burst Size (MBS) defines the maximum buffer size that a queue can use. If the actual queue depth is larger than the MBS, any incoming packet will not be able to get a buffer and the packet will be dropped. This is defined in bytes or Kbytes for access queues with a configurable non-zero minimum of 1byte or a default (without configuring the MBS) of the maximum between 10ms of the PIR or 64Kbytes. A value of zero will cause all packets to be dropped. MBS is a fractional percentage (xx.xx%) of pool size for network queues with defaults varying dependent on the queue (see default network-queue policy for default values). The MBS setting is the main factor determining the packet latency through a system when packets experience congestion. Queues within an egress queue group can have their MBS configured with as target packet queue delay in milliseconds.
- Committed Burst Size (CBS) defines the maximum guaranteed buffer size for an incoming packet. This buffer space is effectively reserved for this queue as long as the CBS is not oversubscribed (such the sum of the CBS for all queues using this pool does not exceed its reserved buffer pool size). For access queues, the CBS is defined in Kbytes with a configurable non-zero minimum of 6Kbytes or a default (without configuring the CBS) of the maximum between 10ms of the CIR or 6Kbytes. It is a fractional percentage (xx.xx%) of pool size for network queues with defaults varying dependent on the queue (see default network-queue policy for default values). Regardless of what is configured, the CBS attained normally

will not be larger than the MBS. One case where CBS could be configured larger than MBS is for queues on LAGs, because in some cases the CBS is shared among the LAG ports (LAG QoS is not covered in this chapter). If the MBS and CBS values are configured to be equal (or nearly equal) this will result in the CBS being slightly higher than the value configured.

- High-prio-only. As a queue can accept both high and low enqueuing priority packets, a high enqueuing priority packet should have a higher probability to get a buffer. High-prio-only is a way to achieve this. Within the MBS, high-prio-only defines that a certain amount of buffer space will be exclusively available for high enqueuing priority packets. At network ingress and all egress buffering, highpriority corresponds to in-profile and low priority to out-of-profile. At service ingress, enqueuing priority is part of the classification. The high-prio-only is defined as a percentage of the MBS, with the default being 10%. A queue being used only for low priority/out-of-profile packets would normally have this set to zero. The high-prio-only could be considered to be an MBS for low enqueuing/out-of-profile packets.
- Hi-low-prio-only. There is an additional threshold, hi-low-prio-only, at egress which is equivalent to an MBS for exceed-profile packet. When the queue depth is beyond the hi-low-prio-only depth, the exceed-profile packets are dropped. The hi-low-prio-only is defined as a percentage of the MBS, with the default being 20%.

As with the buffer pools, the MBS, CBS, high-prio-only, and hi-low-prio-only values attained are based on a number of discrete values (not always an increment of 3Kbytes). The values for these parameters can be seen using the **show pools** command.

The MBS changes dynamically for queues in an egress queue group when H-QoS is used and the queue command **dynamic-mbs** is configured. This results in the MBS being modified in the ratio of the operational PIR to the admin PIR which derives an operational MBS. The ratio also affects the high-prio-only and hi-low-prio-only drop tails, and the WRED slopes if a slope policy is applied ot the queue. The configured CBS is used as the minimum operational MBS and the maximum MBS is capped by the maximum admin MBS (1GB).

As packets are added to a queue they will use the available CBS space, in which case they are placed in the reserved portion in the buffer pool. Once the CBS is exhausted, packets use the shared buffer pool space up to the hi-low-prio-only threshold (for exceed-profile packets), the high-prio-only threshold (for out-of-profile packets), or the maximum MBS size (for inplus-profile and in-profile packets).

The following configuration shows a queue with a specific MBS, CBS and disable high-prio-only.

```
configure
    qos
```

```
        sap-ingress 10 create
            queue 1 create
                mbs 10000
                cbs 100
                high-prio-only 0
            exit
        exit
```

Queue depth monitoring aims to give more visibility to the operator of the queue depths when traffic is bursty. It is a polling mechanism that is by default disabled. Queue depth monitoring can be enabled as a queue override on a service SAP or on a queue group.

The following command enables queue depth monitoring on SAP 1/1/1:11 in VPLS 10:

```
configure service vpls 10 sap 1/1/1:11 ingress queue-override queue 1 create monitor-
depth
```

The result of the queue depth monitoring is presented in the form of occupancy ranges of 10% on the queue depth for each configured queue with the percentage of polls seen in each occupancy range, as follows:

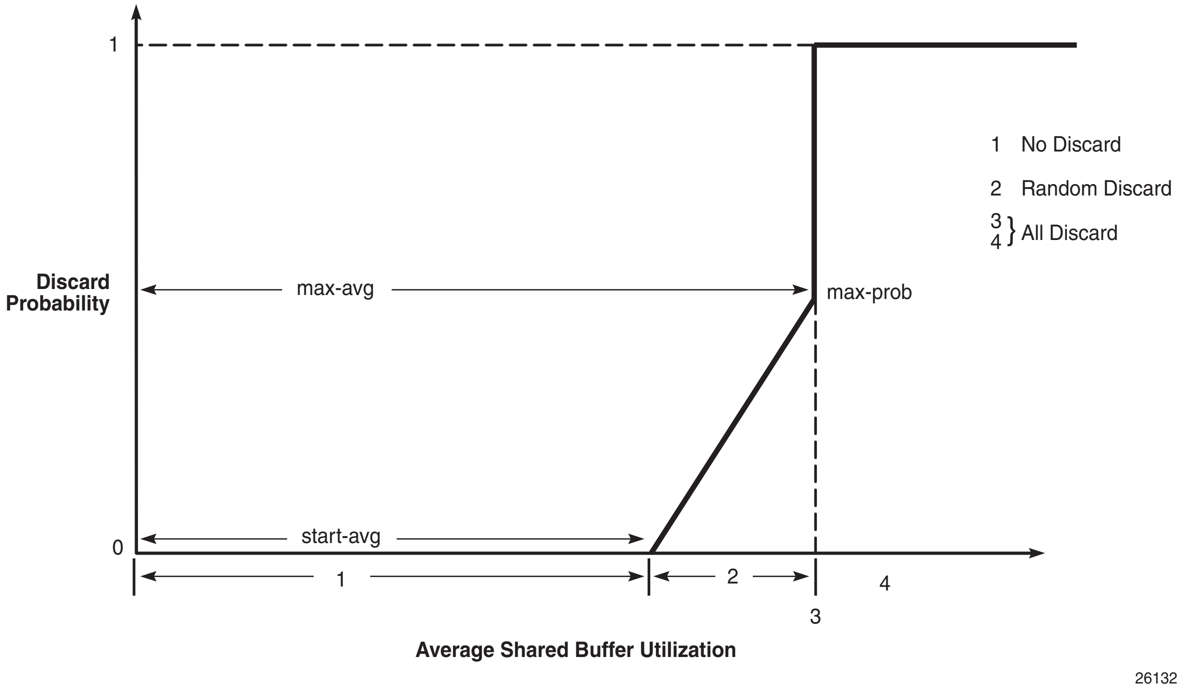
```
*A:PE-1# show service id 10 sap 1/1/1:11 queue-depth
===============================================================================
Queue Depth Information (Ingress SAP)
===============================================================================
No Matching Entries
===============================================================================

===============================================================================
Queue Depth Information (Egress SAP)
===============================================================================
-------------------------------------------------------------------------------
Name                      : 10->1/1/1:11->1
MBS                       : Def

-------------------------------------------------------------------------------
Queue Depths (percentage)
-------------------------------------------------------------------------------
0%-10% 11%-20% 21%-30% 31%-40% 41%-50% 51%-60% 61%-70% 71%-80% 81%-90% 91%-100%
-------------------------------------------------------------------------------
68.21  3.64    3.43    3.47    3.86    3.22    3.86    2.78    3.78    3.66
-------------------------------------------------------------------------------
Average Elapsed Time      : 0d 00:39:11
Wghtd Avg PollEgr Interval: 100 ms
-------------------------------------------------------------------------------
*A:PE-1#
```

Enqueuing Packets: Weighted Random Early Detection (WRED)

In order to gracefully manage the use of the shared portion of the buffer pool, WRED can be configured on that part of the pool, and therefore applies to all queues in the shared pool as it fills. WRED is a congestion avoidance mechanism designed for TCP traffic. This chapter will only focus on the configuration of WRED. WRED-per-queue is an option to have WRED apply on a per egress queue basis, but is not covered here.

WRED is configured by a slope-policy which contains two WRED slope definitions, a high-slope which applies WRED to high enqueuing priority/in-profile packets and a low-slope which applies WRED to low enqueuing priority/out-of-profile packets. Both have the standard WRED parameters: start average (start-avg), maximum average (max-avg) and maximum probability (max-prob), and can be enabled or disabled individually. WRED slope policies also contain definitions for two slopes which are only applicable to access and network egress; a highplus-slope which applies WRED to inplus-profile packets and an exceed-slope which applies WRED to exceed-profile packets. The WRED slope characteristics are shown in Figure 229.

Figure 229 **WRED Slope Characteristics**



A time-average-factor parameter can be configured per slope-policy which determines the sensitivity of the WRED algorithm to shared buffer utilization fluctuations (the smaller the value makes the average buffer utilization more reactive to changes in the instantaneous buffer utilization). The slope-policy is applied on a network port under **config > card > mda > network > ingress > pool** and **config > port > network > egress > pool** and on an access port under **config > port > access > ingress > pool** and **config > port > access > egress > pool**.

WRED is usually configured for assured and best-effort service traffic with premium traffic not typically being subject to WRED as it is always given preferential treatment and should never be dropped.

The following configuration defines a WRED slope policy and applies it to an ingress access port (the highplus and exceed slopes are ignored at ingress).

```
configure
    qos
        slope-policy "slope1" create
            exceed-slope
                shutdown
                start-avg 30
                max-avg 55
                max-prob 80
            exit
            high-slope
                start-avg 80
                max-avg 100
                max-prob 100
                no shutdown
            exit
            highplus-slope
                shutdown
                start-avg 85
                max-avg 100
                max-prob 80
            exit
            low-slope
                max-avg 100
                start-avg 80
                max-prob 100
                no shutdown
            exit
            time-average-factor 12
        exit


configure
    port 1/1/1
        access
            ingress
                pool
                    slope-policy "slope1"
                exit
            exit
        exit
```

The queue sizing parameters and buffer pools layout for ingress is shown in Figure 230.

Figure 230 **Ingress Buffer Pools and Queue Sizing**

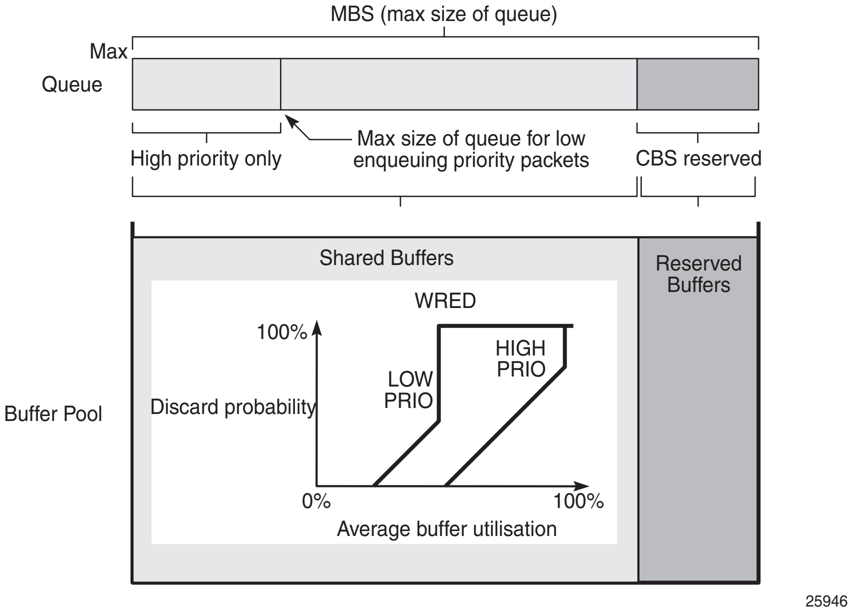
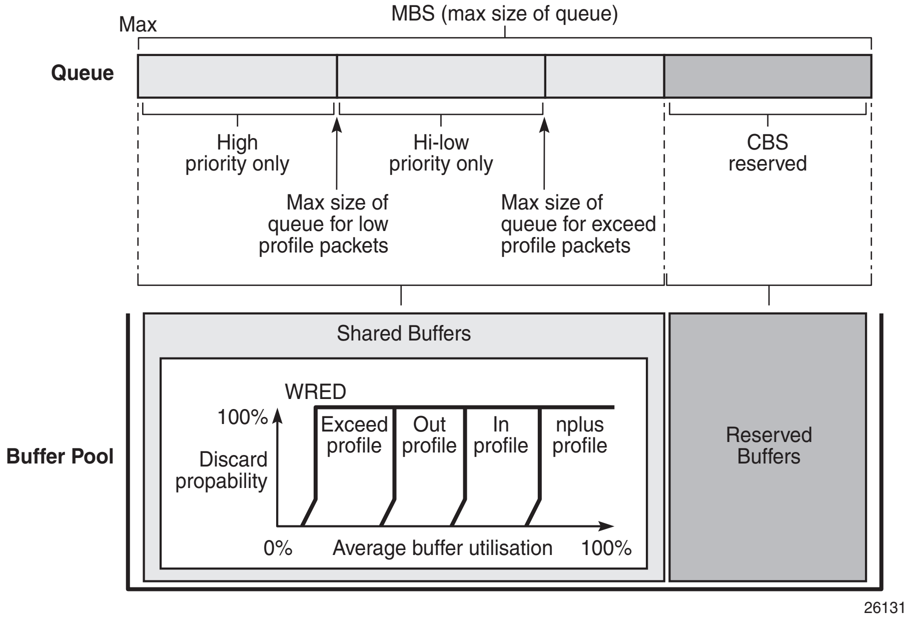


Figure 231 shows the queue sizing parameters and buffer pools layout for egress.

Figure 231 **Egress Buffer Pools and Queue Sizing**



Dequeuing Packets: Queue Rates

A queue has two rate attributes: PIR and CIR. These affect packets only during the dequeue process.

- PIR — If the instantaneous dequeue rate of a queue reaches this rate, the queue is no longer served. Excess packets will be discarded eventually when the queue reaches its MBS/high-prio-only/hi-low-prio-only sizes. The PIR for access ports can be set in Kb/s with a default of **max** or as a percentage (see below). For network ports, the PIR is set as a percentage of 390000Kb/s for ingress queues and of the port speed for egress queues, both with a default of 100%.
- CIR — The CIR is used to determine whether an ingress packet is in-profile or out-of-profile at the SAP ingress. It is also used by the scheduler in that queues operating within their CIRs will be served ahead of queues operating above their CIRs. The CIR for access ports can be set in Kb/s with a default of zero or as a percentage (see below). For network ports, it is set as a percentage of 390000Kb/s for ingress queues and of the port speed for egress queues, with defaults varying dependent on the queue.

A percentage rate can be used in the sap-ingress and sap-egress policies, and can be defined relative to the local-limit (the parent scheduler rate) or the port-limit (the rate of the port on which the SAP is configured, including any egress-rate configured). The parameters rate and percent-rate are mutually exclusive and will overwrite each other when configured in the same policy. The following example shows a percent-rate configured as a port-limit.

```
configure
    qos
        sap-egress 10 create
            queue 1 create
                percent-rate 50.00 cir 10.00 port-limit
            exit
```

The PIR and CIR rates are shown in Figure 232.

The queues operate at discrete rates supported by the hardware. If a configured rate does not match exactly one of the hardware rates an adaptation rule can be configured to control whether the rate is rounded up or down or set to the closest attainable value. The actual rate used can be seen under the operational PIR/CIR (O.PIR/O.CIR) in the **show pools** command output.

The following configuration shows a queue with a PIR, CIR and adaptation rule.

```
configure
    qos
        sap-ingress 20 create
            queue 2 profile-mode create
                adaptation-rule pir max cir min
                rate 10000 cir 5000
            exit
```

By default, the rates apply to packet bytes based on packet accounting, which for Ethernet includes the Layer 2 frame plus the FCS. An alternative is frame accounting which adds the Ethernet inter-frame gap, preamble and start frame delimiter.

Dequeuing Packets: Scheduling

Once a packet is placed on a queue, it is always dequeued from the queue by a scheduler. The scheduling order of the queues dynamically changes depending on whether a queue is currently operating below or above its CIR, with expedited queues being serviced before best-effort queues. This results in a default scheduling order of (in strict priority).

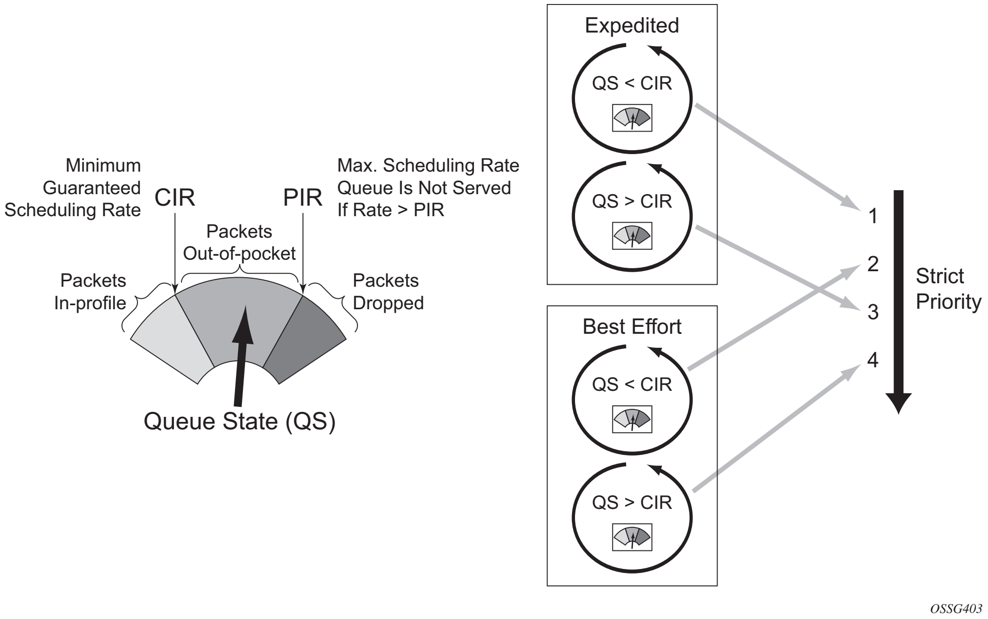
1. Expedited queues operating below CIR
2. Best-effort queues operating below CIR

3. Expedited queues operating above CIR
4. Best-effort queues operating above CIR

This is displayed in Figure 232.

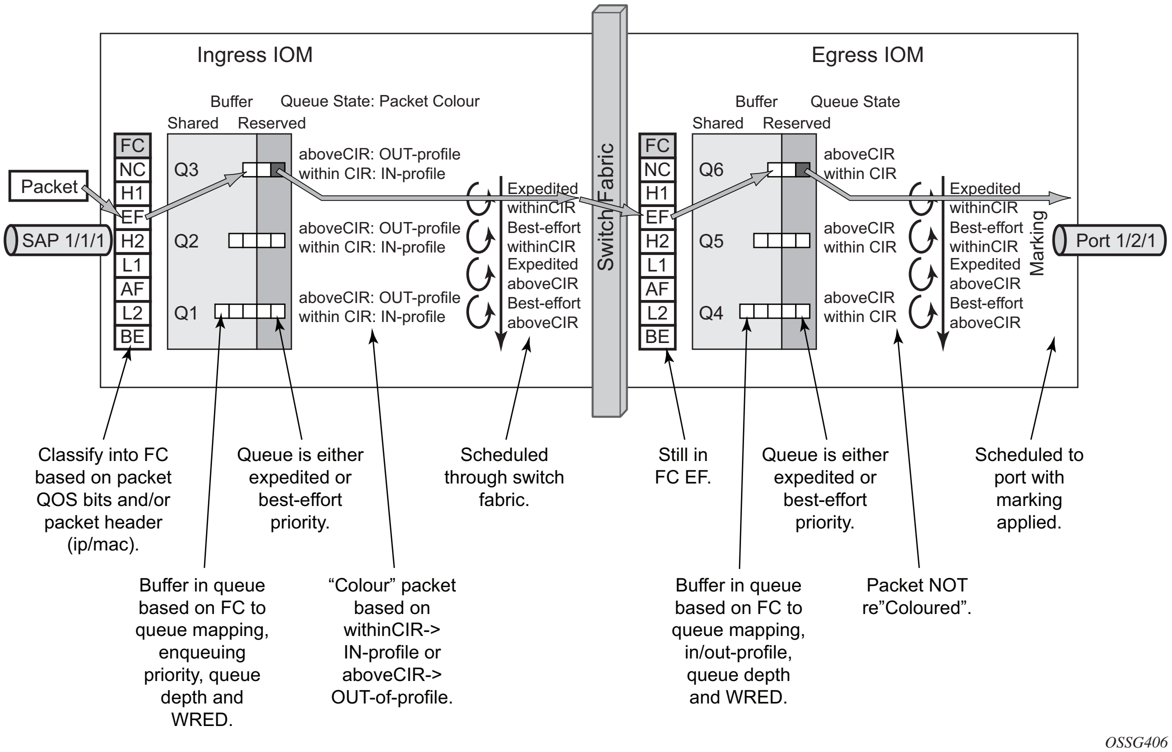
The scheduling operation can be modified using hierarchical QoS (with a scheduler-policy or port-scheduler-policy) which is out of scope of this chapter.

Figure 232 **Scheduling (Dequeuing Packets from the Queue)**



The overall QoS actions at both the ingress and egress IOMs are shown in Figure 233.

Figure 233 **IOM QoS Overview**



Show Output

The following displays **show** command output for:

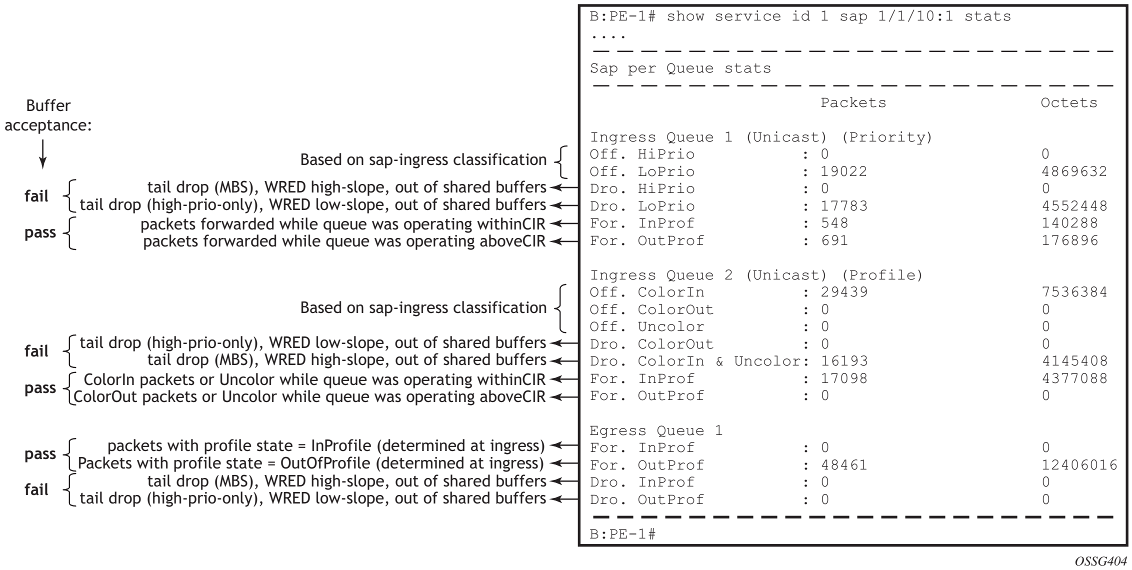
- SAP queue statistics
- port queue statistics
- per-port aggregate egress-queue statistics monitoring
- access-ingress pools

The **show pools** command output for network-ingress and network/access-egress is similar to that of access-ingress and is not included here.

SAP Queue Statistics

The following output shows an example of the ingress and egress statistics on a SAP for an IES service (without multicast enabled, therefore no ingress multicast queue). There are two ingress queues, one being in priority mode and the other in profile mode. An explanation of the statistics is given for each entry.

Figure 234 Ingress and Egress SAP Queue Statistics for an IES Service



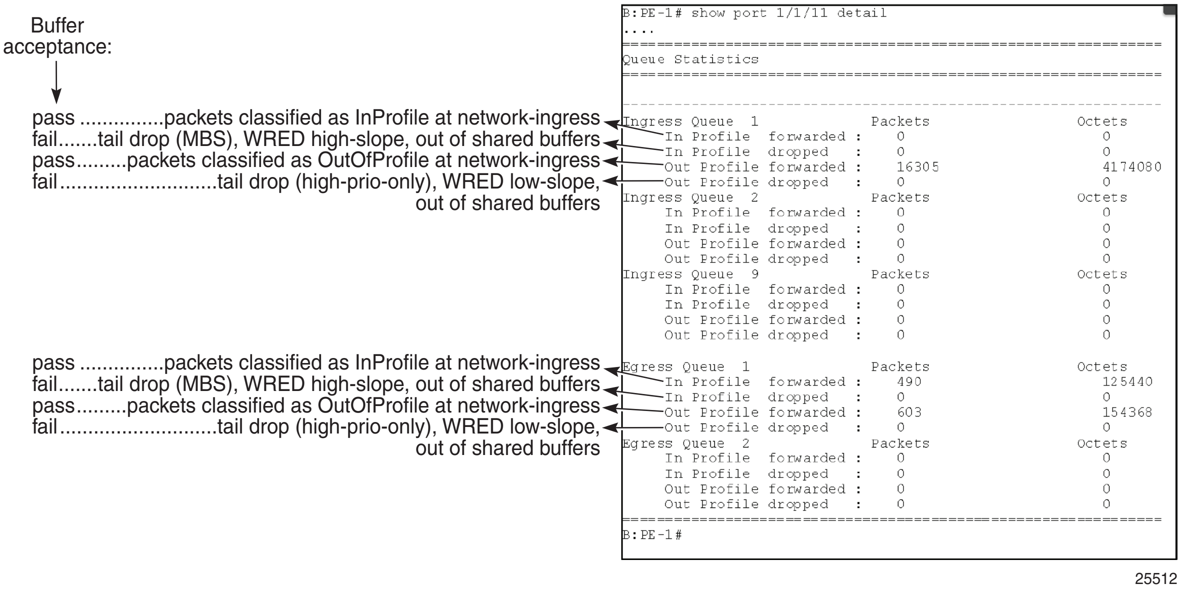
```
B:PE-1# show service id 1 sap 1/1/10:1 stats
....
— — — — — — — — — — — — — — — — — — — — — — — —
Sap per Queue stats
— — — — — — — — — — — — — — — — — — — — — — — —
                        Packets                Octets

Ingress Queue 1 (Unicast) (Priority)
Off. HiPrio          : 0                     0
Off. LoPrio          : 19022                 4869632
Dro. HiPrio          : 0                     0
Dro. LoPrio          : 17783                 4552448
For. InProf          : 548                   140288
For. OutProf         : 691                   176896

Ingress Queue 2 (Unicast) (Profile)
Off. ColorIn         : 29439                 7536384
Off. ColorOut        : 0                     0
Off. Uncolor         : 0                     0
Dro. ColorOut        : 0                     0
Dro. ColorIn & Uncolor: 16193                4145408
For. InProf          : 17098                 4377088
For. OutProf         : 0                     0

Egress Queue 1
For. InProf          : 0                     0
For. OutProf         : 48461                 12406016
Dro. InProf          : 0                     0
Dro. OutProf         : 0                     0
— — — — — — — — — — — — — — — — — — — — — — — —
B:PE-1#
```

Buffer acceptance:
Based on sap-ingress classification
fail
tail drop (MBS), WRED high-slope, out of shared buffers
tail drop (high-prio-only), WRED low-slope, out of shared buffers
pass
packets forwarded while queue was operating withinCIR
packets forwarded while queue was operating aboveCIR
Based on sap-ingress classification
fail
tail drop (high-prio-only), WRED low-slope, out of shared buffers
tail drop (MBS), WRED high-slope, out of shared buffers
pass
ColorIn packets or Uncolor while queue was operating withinCIR
ColorOut packets or Uncolor while queue was operating aboveCIR
pass
packets with profile state = InProfile (determined at ingress)
Packets with profile state = OutOfProfile (determined at ingress)
fail
tail drop (MBS), WRED high-slope, out of shared buffers
tail drop (high-prio-only), WRED low-slope, out of shared buffers
OSSG404

Port Queue Statistics

This output shows an example of the ingress and egress network port statistics. There are two unicast ingress queues (1 and 2) and one multicast ingress queue (9) with two egress queues. An explanation of the statistics is given for each entry.

Figure 235 **Ingress and Egress Network Port Queue Statistics**



```
B:PE-1# show port 1/1/11 detail
....
===============================================================
Queue Statistics
===============================================================

---------------------------------------------------------------
Ingress Queue  1                Packets                 Octets
     In Profile  forwarded :      0                       0
     In Profile  dropped   :      0                       0
     Out Profile forwarded :      16305                   4174080
     Out Profile dropped   :      0                       0
Ingress Queue  2                Packets                 Octets
     In Profile  forwarded :      0                       0
     In Profile  dropped   :      0                       0
     Out Profile forwarded :      0                       0
     Out Profile dropped   :      0                       0
Ingress Queue  9                Packets                 Octets
     In Profile  forwarded :      0                       0
     In Profile  dropped   :      0                       0
     Out Profile forwarded :      0                       0
     Out Profile dropped   :      0                       0

Egress Queue  1                 Packets                 Octets
     In Profile  forwarded :      490                     125440
     In Profile  dropped   :      0                       0
     Out Profile forwarded :      603                     154368
     Out Profile dropped   :      0                       0
Egress Queue  2                 Packets                 Octets
     In Profile  forwarded :      0                       0
     In Profile  dropped   :      0                       0
     Out Profile forwarded :      0                       0
     Out Profile dropped   :      0                       0
===============================================================
B:PE-1#
```

Buffer acceptance:
pass ....packets classified as InProfile at network-ingress
fail.......tail drop (MBS), WRED high-slope, out of shared buffers
pass.........packets classified as OutOfProfile at network-ingress
fail............................tail drop (high-prio-only), WRED low-slope, out of shared buffers
pass ....packets classified as InProfile at network-ingress
fail.......tail drop (MBS), WRED high-slope, out of shared buffers
pass.........packets classified as OutOfProfile at network-ingress
fail............................tail drop (high-prio-only), WRED low-slope, out of shared buffers
25512

Per-Port Aggregate Egress-Queue Statistics Monitoring

Per-port aggregate egress-queue statistics show the number of forwarded and the number of dropped packets for in-profile and out-of-profile packets. All egress queues on the port are monitored: SAP egress, network egress, subscriber egress, egress queue group queues, system queues.

Per-port aggregate egress-queue statistics monitoring is enabled with the following command:

```
configure port 1/1/1 monitor-agg-egress-queue-stats
```

The collected statistics can be displayed as follows:

```
*A:PE-1# show port 1/1/1 statistics egress-aggregate

================================================================================
Port 1/1/1 Egress Aggregate Statistics on Slot 1
================================================================================
                        Forwarded                Dropped                   Total
--------------------------------------------------------------------------------
PacketsIn                       0                      0                       0
PacketsOut                5251690                      0                 5251690
OctetsIn                        0                      0                       0
OctetsOut               357114942                      0               357114942
================================================================================
*A:PE-1#
```

Access-Ingress Pools

This output shows an example of the default pools output for access-ingress. It includes the pools sizes, WRED information and queue parameters for each queue in the pool.

For this particular output, queue 2 on SAP 5/1/1:1 is being over-loaded which is causing its queue depth to be 67087296 bytes, made up of 64509 Kbytes from the shared pool (in use) and 1008 Kbytes from the reserved pool (in use). The output shows the pool total in usage as 65517 Kbytes, which is the sum of the shared and reserved pool in use. Sometimes the sum and total could be different by the size of one buffer, however, this is due to the dynamics of the buffer allocation which uses a ‘sliding-window’ mechanism and may therefore not always be perfectly aligned.

It can be seen that the high, low, and exceed WRED slopes are enabled and their instantaneous drop probability is shown 100% and their max averages are 64512 Kbytes, 46080 Kbytes, and 27648 Kbytes, respectively – this shows that the reserved portion of the buffer pool on this port is exhausted causing WRED to drop the packets for this queue.

The admin and operational PIR on the overloaded queues is 10Mb/s with CIR values of zero.

```
*A:PE-1# show pools 5/1/1 access-ingress

===============================================================================
Pool Information
===============================================================================
Port                 : 5/1/1
Application          : Acc-Ing          Pool Name             : default
CLI Config. Resv CBS : 30%(default)
Resv CBS Step        : 0%               Resv CBS Max          : 0%
Amber Alarm Threshold: 0%               Red Alarm Threshold   : 0%
-------------------------------------------------------------------------------
Utilization                  State       Start-Avg    Max-Avg    Max-Prob
-------------------------------------------------------------------------------
High-Slope                   Up                70%        70%        100%
Low-Slope                    Up                50%        50%         80%
Exceed-Slope                 Up                30%        30%         80%
Time Avg Factor      : 12
Pool Total           : 132096 KB
Pool Shared          : 92160 KB         Pool Resv             : 39936 KB

High Slope Start Avg : 64500 KB         High Slope Max Avg    : 64512 KB
Low Slope Start Avg  : 46068 KB         Low Slope Max Avg     : 46080 KB
Excd Slope Start Avg : 27636 KB         Excd Slope Max Avg    : 27648 KB
```

```
-------------------------------------------------------------------------------

-------------------------------------------------------------------------------
Current Resv CBS    Provisioned    Rising         Falling        Alarm
%age                all Queues     Alarm Thd      Alarm Thd      Color
-------------------------------------------------------------------------------
30%                 1020 KB        NA             NA             Green
Pool Total In Use    : 65517 KB
Pool Shared In Use   : 64509 KB         Pool Resv In Use      : 1008 KB
WA Shared In Use     : 64509 KB

Hi-Slope Drop Prob   : 100              Lo-Slope Drop Prob    : 100
Excd-Slope Drop Prob : 100

===============================================================================
Queue : 1->5/1/1:1->1
===============================================================================
FC Map           : be l2 af l1 h2 h1 nc
Tap              : 5/1
Admin PIR        : 10000000             Oper PIR         : Max
Admin CIR        : 0                    Oper CIR         : 0
Admin MBS        : 12288 KB             Oper MBS         : 12288 KB
Hi Prio Only     : 1344 KB              Hi Low Prio Only : 2496 KB
CBS              : 12 KB                Depth            : 0
Slope            : not-applicable
===============================================================================

===============================================================================
Queue : 1->5/1/1:1->2
===============================================================================
FC Map           : ef
Tap              : 5/1
Admin PIR        : 10000                Oper PIR         : 10000
Admin CIR        : 0                    Oper CIR         : 0
Admin MBS        : 132096 KB            Oper MBS         : 132096 KB
Hi Prio Only     : 0 KB                 Hi Low Prio Only : 27648 KB
CBS              : 1008 KB              Depth            : 67087296 B
Slope            : not-applicable
===============================================================================

===============================================================================
Queue : 28->5/1/1:28->1
===============================================================================
FC Map           : be l2 af l1 h2 ef h1 nc
Tap              : 5/1
Admin PIR        : 10000000             Oper PIR         : Max
Admin CIR        : 0                    Oper CIR         : 0
Admin MBS        : 12288 KB             Oper MBS         : 12288 KB
Hi Prio Only     : 1344 KB              Hi Low Prio Only : 2496 KB
CBS              : 0 KB                 Depth            : 0
Slope            : not-applicable
===============================================================================

===============================================================================
Queue : 28->5/1/1:28->11
===============================================================================
FC Map           : be l2 af l1 h2 ef h1 nc
Tap              : MCast
Admin PIR        : 10000000             Oper PIR         : Max
Admin CIR        : 0                    Oper CIR         : 0
Admin MBS        : 12288 KB             Oper MBS         : 12288 KB
Hi Prio Only     : 1344 KB              Hi Low Prio Only : 2496 KB
CBS              : 0 KB                 Depth            : 0
Slope            : not-applicable
===============================================================================
===============================================================================

*A:PE-1#
```

Conclusion

This chapter has described the basic QoS functionality available on the Nokia 7x50 platforms, specifically focused on the FP2 chipset. This comprises of the use of queues to shape traffic at the ingress and egress of the system and the classification, buffering, scheduling and remarking of traffic on access, network, and hybrid ports.

Customer Document and Product Support

Customer documentation

Customer Documentation Welcome Page

Technical Support

Product Support Portal

Documentation feedback

Customer Documentation Feedback