

Cisco Nexus 9000 Series NX-OS VXLAN Configuration Guide, Release 10.6(x)

First Published: 2025-08-13

Americas Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
http://www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 527-0883

THE SPECIFICATIONS AND INFORMATION REGARDING THE PRODUCTS REFERENCED IN THIS DOCUMENTATION ARE SUBJECT TO CHANGE WITHOUT NOTICE. EXCEPT AS MAY OTHERWISE BE AGREED BY CISCO IN WRITING, ALL STATEMENTS, INFORMATION, AND RECOMMENDATIONS IN THIS DOCUMENTATION ARE PRESENTED WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED.

The Cisco End User License Agreement and any supplemental license terms govern your use of any Cisco software, including this product documentation, and are located at: https://www.cisco.com/c/en/us/about/legal/cloud-and-software/software-terms.html. Cisco product warranty information is available at https://www.cisco.com/c/en/us/products/warranty-listing.html. US Federal Communications Commission Notices are found here https://www.cisco.com/c/en/us/products/us-fcc-notice.html.

IN NO EVENT SHALL CISCO OR ITS SUPPLIERS BE LIABLE FOR ANY INDIRECT, SPECIAL, CONSEQUENTIAL, OR INCIDENTAL DAMAGES, INCLUDING, WITHOUT LIMITATION, LOST PROFITS OR LOSS OR DAMAGE TO DATA ARISING OUT OF THE USE OR INABILITY TO USE THIS MANUAL, EVEN IF CISCO OR ITS SUPPLIERS HAVE BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

Any products and features described herein as in development or available at a future date remain in varying stages of development and will be offered on a when-and if-available basis. Any such product or feature roadmaps are subject to change at the sole discretion of Cisco and Cisco will have no liability for delay in the delivery or failure to deliver any products or feature roadmap items that may be set forth in this document.

Any Internet Protocol (IP) addresses and phone numbers used in this document are not intended to be actual addresses and phone numbers. Any examples, command display output, network topology diagrams, and other figures included in the document are shown for illustrative purposes only. Any use of actual IP addresses or phone numbers in illustrative content is unintentional and coincidental.

The documentation set for this product strives to use bias-free language. For the purposes of this documentation set, bias-free is defined as language that does not imply discrimination based on age, disability, gender, racial identity, ethnic identity, sexual orientation, socioeconomic status, and intersectionality. Exceptions may be present in the documentation due to language that is hardcoded in the user interfaces of the product software, language used based on RFP documentation, or language that is used by a referenced third-party product.

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: https://www.cisco.com/c/en/us/about/legal/trademarks.html. Third-party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1721R)

© 2025 Cisco Systems, Inc. All rights reserved.



CONTENTS

Trademarks ?

PREFACE

Preface xxiii

Audience xxiii

Document Conventions xxiii

Related Documentation for Cisco Nexus 9000 Series Switches xxiv

Documentation Feedback xxiv

Communications, services, and additional information xxiv

Cisco Bug Search Tool xxv

Documentation feedback xxv

CHAPTER 1

New and Changed Information 1

New and Changed Information 1

CHAPTER 2

Overview 5

Licensing Requirements 5

Supported Platforms 5

VXLAN Overview 5

Cisco Nexus 9000 as Hardware-Based VXLAN Gateway 6

VXLAN Encapsulation and Packet Format 6

VXLAN Tunnel 7

VXLAN Tunnel Endpoint 7

Underlay Network 7

Overlay Network 7

Distributed Anycast Gateway 7

Control Plane 8

CHAPTER 3 **Configuring the Underlay** 11

IP Fabric Underlay 11

Underlay Considerations 11

Unicast routing and IP addressing options 14

OSPF Underlay IP Network 14

IS-IS Underlay IP Network 19

eBGP Underlay IP Network 25

Multicast Routing in the VXLAN Underlay 30

Underlay Deployment Without Multicast (Ingress Replication) 42

CHAPTER 4 **Configuring VXLAN** 43

Guidelines and Limitations for VXLAN 43

Considerations for VXLAN Deployment 58

vPC Considerations for VXLAN Deployment 61

Network Considerations for VXLAN Deployments 65

Considerations for the Transport Network 66

Considerations for Tunneling VXLAN 67

Configuring VXLAN 69

Enabling VXLANs 69

Mapping VLAN to VXLAN VNI 69

Creating and Configuring an NVE Interface and Associate VNIs 70

Creating and Configuring an NVE Interface Loopback 71

Migration from Single NVE Source Loopback Interface to Separate Source Loopback 73

Configuring a VXLAN VTEP in vPC 73

Configuring Static MAC for VXLAN VTEP 77

Disabling VXLANs 78

Configuring BGP EVPN Ingress Replication 78

Configuring Static Ingress Replication 79

VXLAN and IP-in-IP Tunneling 80

Configuring VXLAN Static Tunnels 83

About VXLAN Static Tunnels 83

Guidelines and Limitations for VXLAN Static Tunnels 83

Enabling VXLAN Static Tunnels 84

Configuring VRF Overlay for Static Tunnels 85
Configuring a VRF for VXLAN Routing 86
Configuring the L3 VNI for Static Tunnels 86
Configuring the Tunnel Profile 88
Verifying VXLAN Static Tunnels 89
Example Configurations for VXLAN Static Tunnels 89

CHAPTER 5 **Configuring VXLAN with IPv6 in the Underlay (VXLANv6) 91**
Information About Configuring VXLANv6 91
Information About vPC and VXLAN with IPv6 in the Underlay (VXLANv6) 92
Information About vPC Peer Keepalive and VXLAN with IPv6 in the Underlay (VXLANv6) 92
Guidelines and Limitations for VXLAN with IPv6 in the Underlay (VXLANv6) 93
Configuring the VTEP IP Address 96
Configuring vPC for VXLAN with IPv6 in the Underlay (VXLANv6) 98
Configuring VXLAN EVPN with IPv6 Ingress Replication 99
Configuring Ingress Replication for L2-VNI 100
Configuring Multicast Underlay for BUM on L2-VNI 101
Example Configurations for VXLAN with IPv6 in the Underlay (VXLANv6) 102
Verifying VXLAN with IPv6 in the Underlay (VXLANv6) 103

CHAPTER 6 **Configuring VXLAN BGP EVPN 113**
About VXLAN BGP EVPN 113
About RD Auto 113
About Route-Target Auto 114
Guidelines and Limitations for VXLAN BGP EVPN 115
About VXLAN EVPN with Downstream VNI 121
Asymmetric VNIs 121
Shared Services VRFs 122
Multi-Site with Asymmetric VNIs 122
Guidelines and Limitations for VXLAN EVPN with Downstream VNI 123
Configuring VXLAN BGP EVPN 126
Enabling VXLAN 126
Configuring VLAN and VXLAN VNI 126
Configuring New L3VNI Mode 127

Guidelines and Limitations for New L3VNI Mode 127
Configuring New L3VNI Mode 130
Verifying New L3VNI Mode Configuration 131
Configuring VRF for VXLAN Routing 131
Configuring VXLAN UDP Source Port 132
Configuring SVI for Core-facing VXLAN Routing 133
Configuring SVI for Host-Facing VXLAN Routing 134
Configuring the NVE Interface and VNIs Using Multicast 135
Configuring the Delay Timer on NVE Interface 136
Configuring VXLAN EVPN Ingress Replication 136
Configuring BGP on the VTEP 138
Configuring iBGP for EVPN on the Spine 139
Configuring eBGP for EVPN on the Spine 140
Suppressing ARP 142
Disabling VXLANs 142
Duplicate Detection for IP and MAC Addresses 143
Configuring Event History Size for L2RIB 145
Associating VLAN to VRF - L2 to L3 Mapping 146
Verifying the VXLAN BGP EVPN Configuration 147
Verifying the VXLAN EVPN with Downstream VNI Configuration 148
Example of VXLAN BGP EVPN (iBGP) 150
Example of VXLAN BGP EVPN (eBGP) 162
Example Show Commands 174
Configuring ND Suppression 177
ND Suppression on the Overlay 177
Guidelines and Limitations for ND Suppression 177
Configuring ND Suppression 178
Verifying the ND Suppression Configuration 179
VXLAN BGP EVPN with RFC 5549 Underlay 183
VXLAN BGP EVPN with RFC 5549 Underlay Overview 183
Benefits of VXLAN BGP EVPN with RFC 5549 Underlay 183
Supported Platform and Release for VXLAN BGP EVPN with RFC 5549 Underlay 183
Functionalities of VXLAN BGP EVPN with RFC 5549 Underlay 183
VXLAN BGP EVPN with RFC 5549 Underlay on Single-Site 184

Supported Combination of RFC 5549 and IPv4 over Core and DC Fabric 185
Guidelines for VXLAN BGP EVPN with RFC 5549 Underlay 185
Supported and Unsupported Features of VXLAN BGP EVPN with RFC 5549 Underlay 186
Configuring VXLAN BGP EVPN with RFC 5549 Underlay on Single-Site 188
Examples of VXLAN BGP EVPN with RFC 5549 Underlay 192

CHAPTER 7 EVPN Hybrid IRB Mode 195
EVPN Hybrid IRB Mode 195

CHAPTER 8 Default Gateway Coexistence of HSRP and Anycast Gateway (VXLAN EVPN) 199
Default Gateway Coexistence of HSRP and Anycast Gateway (VXLAN EVPN) 199
Guidelines and Limitations for Migrating from Classic Ethernet / FabricPath to VXLAN 200
Configuring Classic Ethernet / FabricPath to VXLAN Migration 202
Configuring an External Port on Border Leaf for Migration 203
Configuring External IP Address for Migration 204

CHAPTER 9 Configuring External VRF Connectivity and Route Leaking 207
Configuring External VRF Connectivity 207
About External Layer-3 Connectivity for VXLAN BGP EVPN Fabrics 207
VXLAN BGP EVPN - VRF-lite brief 207
Guidelines and Limitations for External VRF Connectivity and Route Leaking 208
Configuring VXLAN BGP EVPN with eBGP for VRF-lite 208
VXLAN BGP EVPN - Default-Route, Route Filtering on External Connectivity 214
Configuring VXLAN BGP EVPN with OSPF for VRF-lite 221
Configuring Route Leaking 225
About Centralized VRF Route-Leaking for VXLAN BGP EVPN Fabrics 225
Guidelines and Limitations for Centralized VRF Route-Leaking 225
Centralized VRF Route-Leaking Brief - Specific Prefixes Between Custom VRF 226
Configuring Centralized VRF Route-Leaking - Specific Prefixes between Custom VRF 226
Configuring VRF Context on the Routing-Block VTEP 226
Configuring the BGP VRF instance on the Routing-Block 228
Example - Configuration Centralized VRF Route-Leaking - Specific Prefixes Between Custom VRF 229
Centralized VRF Route-Leaking Brief - Shared Internet with Custom VRF 230

Configuring Centralized VRF Route-Leaking - Shared Internet with Custom VRF 231
Configuring Internet VRF on Border Node 231
Configuring Shared Internet BGP Instance on the Border Node 232
Configuring Custom VRF on Border Node 232
Configuring Custom VRF Context on the Border Node - 1 233
Configuring Custom VRF Instance in BGP on the Border Node 234
Example - Configuration Centralized VRF Route-Leaking - Shared Internet with Custom VRF 235
Centralized VRF Route-Leaking Brief - Shared Internet with VRF Default 237
Configuring Centralized VRF Route-Leaking - Shared Internet with VRF Default 238
Configuring VRF Default on Border Node 238
Configuring BGP Instance for VRF Default on the Border Node 239
Configuring Custom VRF on Border Node 239
Configuring Filter for Permitted Prefixes from VRF Default on the Border Node 240
Configuring Custom VRF Context on the Border Node - 2 240
Configuring Custom VRF Instance in BGP on the Border Node 242
Example - Configuration Centralized VRF Route-Leaking - VRF Default with Custom VRF 242

CHAPTER 10 **Configuring Seamless Integration of EVPN with L3VPN (MPLS LDP) 245**
Information About Configuring Seamless Integration of EVPN with L3VPN (MPLS LDP) 245
Guidelines and Limitations for Configuring Seamless Integration of EVPN with L3VPN (MPLS LDP) 245
Configuring Seamless Integration of EVPN with L3VPN (MPLS LDP) 246

CHAPTER 11 **Configuring Seamless Integration of EVPN with L3VPN (MPLS SR) 251**
Information About Configuring Seamless Integration of EVPN with L3VPN (MPLS SR) 251
Guidelines and Limitations for Configuring Seamless Integration of EVPN with L3VPN (MPLS SR) 254
Configuring Seamless Integration of EVPN with L3VPN (MPLS SR) 257
Example Configuration for Configuring Seamless Integration of EVPN with L3VPN (MPLS SR) 261
Configuring DSCP Based SR-TE Flow Steering 271

CHAPTER 12 **Configuring Seamless Integration of EVPN with L3VPN SRv6 273**
About Seamless Integration of EVPN with L3VPN SRv6 Handoff 273
Guidelines and Limitations for EVPN to L3VPN SRv6 Handoff 274

Importing L3VPN SRv6 Routes into EVPN VXLAN 275
Importing EVPN VXLAN Routes into L3VPN SRv6 277
Example Configuration for VXLAN EVPN to L3VPN SRv6 Handoff 278

CHAPTER 13 Configuring Seamless Integration of EVPN (TRM) with MVPN 281
About Seamless Integration of EVPN (TRM) with MVPN (Draft Rosen) 281
Supported RP Positions 282
Guidelines and Limitations for Seamless Integration of EVPN (TRM) with MVPN 282
Configuring the Handoff Node for Seamless Integration of EVPN (TRM) with MVPN 283
PIM/IGMP Configuration for the Handoff Node 283
BGP Configuration for the Handoff Node 284
VXLAN Configuration for the Handoff Node 285
MVPN Configuration for the Handoff Node 286
CoPP Configuration for the Handoff Node 287
Configuration Example for Seamless Integration of EVPN (TRM) with MVPN 288

CHAPTER 14 Configuring VXLAN EVPN Multi-Site 293
About VXLAN EVPN Multi-Site 293
About VXLAN EVPN Multi-Site with IPv6 Underlay 294
Dual RD Support for Multi-Site 295
RP Placement in the DCI Core 296
Interoperability with EVPN Multi-Homing Using ESI for Multi-Site Anycast BGW 296
Guidelines and Limitations for VXLAN EVPN Multi-Site 297
Guidelines and Limitations for VXLAN EVPN Multi-Site with IPv6 Underlay 302
Enabling VXLAN EVPN Multi-Site 302
Enabling VXLAN EVPN Multi-Site with IPv6 Multicast Underlay 304
Configuring Dual RD Support for Multi-Site 306
Configuring VNI Dual Mode 308
Configuring Fabric/DCI Link Tracking 309
Configuring Fabric External Neighbors 310
Configuring VXLAN EVPN Multi-Site Storm Control 311
Verifying VXLAN EVPN Multi-Site Storm Control 312
Multi-Site with vPC Support 312
About Multi-Site with vPC Support 312

Guidelines and Limitations for Multi-Site with vPC Support 312
Configuring Multi-Site with vPC Support 313
Verifying the Multi-Site with vPC Support Configuration 316
Configuration Example for Multi-Site with Asymmetric VNIs 318
Advertise Using PIP Towards Fabric 319
Guidelines and Limitations 319
Configuring BGW to Advertise using PIP Towards Fabric 320
Verifying the Configuration 320
TRM with Multi-Site 321
Information About Configuring TRM with Multi-Site 321
Information About Configuring TRM Multi-Site with IPv6 Underlay 323
Guidelines and Limitations for TRM with Multi-Site 325
Guidelines and Limitations for TRM Multi-Site with IPv6 Underlay 329
Configuring TRM with Multi-Site 329
Configuring TRM Multi-Site with IPv6 Underlay 331
Verifying TRM with Multi-Site Configuration 333
VXLAN EVPN Multi-Site with RFC 5549 Underlay 333
VXLAN EVPN Multi-Site with RFC 5549 Underlay Overview 333
Benefits of VXLAN EVPN Multi-Site with RFC 5549 Underlay 334
Supported Platform and Release for VXLAN EVPN Multi-Site with RFC 5549 Underlay 334
Functionalities of VXLAN EVPN Multi-Site with RFC 5549 Underlay 334
VXLAN BGP EVPN with RFC 5549 Underlay on Multi-Site 334
Guidelines for VXLAN EVPN Multi-Site with RFC 5549 Underlay 335
Supported and Unsupported Features of VXLAN EVPN Multi-Site with RFC 5549 Underlay 336
Configuring VXLAN EVPN Multi-Site with RFC 5549 Underlay on Multi-Site 337
Examples of VXLAN EVPN Multi-Site with RFC 5549 Underlay for Multi-Site 339

CHAPTER 15 **Configure VXLAN EVPN Traffic Engineering - Multi-Site Egress Load Balancing 341**
VXLAN EVPN TE - Multi-Site egress load balancing 341
Guidelines and Limitations for VXLAN EVPN TE - Multi-Site Egress Load-Balancing 342
VXLAN EVPN TE - Multi-Site egress load balancing configurations 344
Create an egress load balance filter policy for underlay 345
Creating an Egress Load-Balancing Auto-Multipath Policy for Underlay 346
Weight Derivation in Underlay 348

Load Share Weight Calculation 348
Explicit Load Share Weight Calculation 349
Enabling Egress Load-Balancing for Overlay 351
Enabling uECMP or wuECMP Load-Balancing for Overlay 352
Verifying VXLAN EVPN TE - Multi-Site Egress Load-Balancing Configuration 354
Configuration examples for VXLAN EVPN TE - Multi-Site egress load balancing 356
uECMP in Underlay, single EVPN next-hop for Overlay prefixes 356
Static wuECMP with Load-Share and Explicit load-share in Underlay, and single EVPN next-hop for Overlay Prefixes 360
Dynamic weight (wuECMP) in Overlay and Underlay, and with AIGP in Underlay 370

CHAPTER 16 **Configuring Tenant Routed Multicast (TRM) 381**
About Tenant Routed Multicast 382
About Tenant Routed Multicast Mixed Mode 383
About Tenant Routed Multicast with IPv6 Overlay 383
About Multicast Flow Path Visibility for TRM Flows 384
About Configuring VXLAN EVPN and TRM with IPv6 Underlay 384
Guidelines and Limitations for Tenant Routed Multicast 385
Guidelines and Limitations for Layer 3 Tenant Routed Multicast 387
Guidelines and Limitations for Layer 2/Layer 3 Tenant Routed Multicast (Mixed Mode) 390
Guidelines and Limitations for VXLAN EVPN and TRM with IPv6 in the Multicast Underlay 391
Rendezvous Point for Tenant Routed Multicast 392
Configuring a Rendezvous Point for Tenant Routed Multicast 393
Configuring a Rendezvous Point Inside the VXLAN Fabric 394
Configuring an External Rendezvous Point 395
Configuring RP Everywhere with PIM Anycast 397
Configuring a TRM Leaf Node for RP Everywhere with PIM Anycast 398
Configuring a TRM Border Leaf Node for RP Everywhere with PIM Anycast 399
Configuring an External Router for RP Everywhere with PIM Anycast 401
Configuring RP Everywhere with MSDP Peering 403
Configuring a TRM Leaf Node for RP Everywhere with MSDP Peering 404
Configuring a TRM Border Leaf Node for RP Everywhere with MSDP Peering 405
Configuring an External Router for RP Everywhere with MSDP Peering 408
Configuring Layer 3 Tenant Routed Multicast 410

Configuring TRM on the VXLAN EVPN Spine 415
Configuring Tenant Routed Multicast in Layer 2/Layer 3 Mixed Mode 417
Configuring VXLAN EVPN and TRM with IPv6 Multicast Underlay 422
Configuring L2-VNI Based Multicast Group in Underlay 422
Configuring L3-VNI Based Multicast Group in Underlay 423
Enabling PIMv6 for Underlay 424
Configuring Layer 2 Tenant Routed Multicast 425
Configuring TRM with vPC Support 426
Configuring TRM with vPC Support (Cisco Nexus 9504-R and 9508-R) 429
Flex Stats for TRM 432
Configuring Flex Stats for TRM 433
Configuring TRM Data MDT 434
About TRM Data MDT 434
Guidelines and Limitations for TRM Data MDT 434
Configuring TRM Data MDT 435
Verifying TRM Data MDT Configuration 436
Configuring IGMP Snooping 437
Overview of IGMP Snooping Over VXLAN 437
Guidelines and Limitations for IGMP Snooping Over VXLAN 437
Configuring IGMP Snooping Over VXLAN 437
Verifying VXLAN EVPN and TRM with IPv6 Multicast Underlay 438
Example Configuration for VXLAN EVPN and TRM with IPv6 Multicast Underlay 442

CHAPTER 17 **Optimized Layer 2 Overlay Multicast** 445
Multicast Layer 2 EVPN Type-6 routes 445
Functionality of Multicast Layer 2 EVPN Type-6 routes 446
Topology of Multicast Layer 2 EVPN Type-6 routes (Reference) 446
Guidelines and Limitations of Multicast Layer 2 EVPN Type-6 routes 450
Configure Multicast Layer 2 EVPN Route Type 6 452
Configure IPv4 IGMP Snooping 452
Configure IPv6 MLD Snooping 453
Verification of Multicast Layer 2 EVPN Route Types Configuration 454
Examples of Multicast Layer 2 EVPN Route Type 6 455

CHAPTER 18 VXLAN Path Validation and Verification 459
VXLAN OAM or VXLAN NGOAM 459
VXLAN OAM Tools 459
VXLAN OAM Payload 459
Guidelines and Limitations for VXLAN NGOAM 460
Configure VXLAN NGOAM 460
Configure NGOAM Profile 461
Fault Isolation and Verification Tools 462
Ping Message 463
Topology of Ping Functionality 463
Traceroute Message 464
Traceroute Functionality in Single site 464
Traceroute Functionality in Multi-Site 465
Pathtrace Message 467
Pathtrace Functionality in Single site 467
Pathtrace Functionality in Multi-Site 468
Comparison of Traceroute and Pathtrace Message 468
Guidelines for Fault Isolation and Verification Tools 469
Examples for Fault Isolation and Verification Tools 469
Ping Message Examples 469
Traceroute Message Examples 470
Pathtrace Message Examples 471
VXLAN EVPN Loop Detection and Mitigation 472
Causes and Impacts of Loop 472
VXLAN EVPN Loop Detection and Mitigation 472
Southbound Loop Detection on Layer-3 Interface 474
Functionalities of SLD on Layer-3 Interface 474
Topology Overview of SLD on Layer-3 Interface 475
L2 and L3 SLD Feature Functionality Comparison 476
Guidelines and Limitations for VXLAN EVPN Loop Detection and Mitigation 477
Supported Platform and Release for VXLAN EVPN Loop Detection and Mitigation 478
Guidelines and Limitations for SLD on L3 Interface 478
Supported Platform and Release for SLD on L3 Interface 478

Prerequisites of NGOAM Southbound Loop Detection 478
Configure NGOAM Southbound Loop Detection on Layer-2 Interfaces 479
Configure NGOAM Southbound Loop Detection on Layer-3 Interfaces 480
Detect Loops and Bring Up Ports On Demand 481
Configuration Examples for NGOAM Southbound Loop Detection and Mitigation 482

CHAPTER 19 **Configuring VXLAN QoS 483**
Information About VXLAN QoS 483
VXLAN QoS Terminology 484
VXLAN QoS Features 485
Trust Boundaries 485
Classification 485
Marking 486
Policing 486
Queuing and Scheduling 486
Traffic Shaping 486
Network QoS 487
VXLAN Priority Tunneling 487
MQC CLI 487
VXLAN QoS Topology and Roles 487
Ingress VTEP and Encapsulation in the VXLAN Tunnel 488
Transport Through the VXLAN Tunnel 488
Egress VTEP and Decapsulation of the VXLAN Tunnel 488
Classification at the Ingress VTEP, Spine, and Egress VTEP 489
IP to VXLAN 489
IP to VXLAN with Outer DSCP 490
Inside the VXLAN Tunnel 490
VXLAN to IP 491
Decapsulated Packet Priority Selection 491
CoS Preservation 492
Guidelines and Limitations for VXLAN QoS 493
Default Settings for VXLAN QoS 496
Configuring VXLAN QoS 497
Configuring Type QoS on the Egress VTEP 497

Setting Outer DSCP on the Ingress VTEP 499
Verifying the VXLAN QoS Configuration 500
VXLAN QoS Configuration Examples 500

CHAPTER 20 Configuring BGP EVPN Filtering 503
About BGP EVPN Filtering 503
Guidelines and Limitations for BGP EVPN Filtering 504
Configuring BGP EVPN Filtering 504
Configuring the Route Map with Match and Set Clauses 504
Matching Based on EVPN Route Type 505
Matching Based on MAC Address in the NLRI 505
Matching Based on RMAC Extended Community 506
Setting the RMAC Extended Community 507
Setting the EVPN Next-Hop IP Address 507
Setting the Gateway IP Address for Route Type-5 508
Applying the Route Map at the Inbound or Outbound Level 509
BGP EVPN Filtering Configuration Examples 510
Configuring a Table Map 518
Configuring a MAC List and a Route Map that Matches the MAC List 518
Applying the Table Map 519
Table Map Configuration Example 520
Verifying BGP EVPN Filtering 522

CHAPTER 21 Configuring VXLAN BGP-EVPN Null Route 525
About EVPN Null Route 525
Guidelines and Limitations for VXLAN BGP-EVPN Null Route 526
Configuring Static MAC 527
Configuring ARP/ND 527
Configuring Prefix-Null Route on Local VTEP 529
Configuring RPM Route-Map on Remote VTEP 531
Configuration Example for Null Route 533
Verifying EVPN Null Route Configuration 534

CHAPTER 22 Configuring Port VLAN Mapping 537

About Translating Incoming VLANs 537
Guidelines and Limitations for Port VLAN Mapping 538
Configuring Port VLAN Mapping on a Trunk Port 541
Configuring Inner VLAN and Outer VLAN Mapping on a Trunk Port 543
About Port Multi-VLAN Mapping 545
Guidelines and Limitations for Port Multi-VLAN Mapping 546
Configuring Port Multi-VLAN Mapping 547

CHAPTER 23 **Micro-segmentation for VXLAN Fabrics Using Group Policy Option (GPO) 553**
Overview 553
GPO 553
Terminology 554
Guidelines and Limitations 555
Configuring Micro-segmentation using GPO 558
Enabling GPO 558
Creating a Security Group 561
Creating Security Group using MAC 564
Creating a Security Class-Map 565
Creating a Security Policy-Map 566
Configuring Security contracts between Security Groups 567
Configuration Examples for GPO 568
Verifying GPO 569
VXLAN Multi-Site and GPO Interoperability 572

CHAPTER 24 **Configuring Layer 4 - Layer 7 Services 579**
About VXLAN Layer 4 - Layer 7 Services 579
Integrating Layer 3 Firewalls in VXLAN Fabrics 579
Single-Attached Firewall with Static Routing 580
Recursive Static Routes Distributed to the Rest of the Fabric 582
Redistribute Static Routes into BGP and Advertise to the Rest of the Fabric 582
Dual-Attached Firewall with Static Routing 582
Single-Attached Firewall with eBGP Routing 583
Dual-Attached Firewall with eBGP Routing 586
Per-VRF Peering via vPC Peer-Link 589

Dual-Attached Firewall with OSPF 589
Redistribute OSPF Routes into BGP and Advertise to the Rest of the Fabric 591
Firewall as Default Gateway 591
Transparent Firewall Insertion 592
Overview of EVPN with Transparent Firewall Insertion 592
EVPN with Transparent Firewall Insertion Example 594
Show Command Examples 597
Firewall Clustering with VXLAN BGP EVPN 598
Service Redirection in VXLAN EVPN Fabrics 601
Use of Policy-Based Redirect for Services Insertion 601
Guidelines and Limitations for Policy-Based Redirect 602
Enabling the Policy-Based Redirect Feature 602
Configuring a Route Policy 603
Verifying the Policy-Based Redirect Configuration 605
Configuration Example for Policy-Based Redirect 605
Enhanced-Policy Based Redirect (ePBR) 606

CHAPTER 25 **Configuring Proportional Multipath for VNF 609**
About proportional multipath for VNF 609
Proportional Multipath for VNF with Multi-Site 613
Prerequisites for Proportional Multipath for VNF 614
Guidelines and limitations for proportional multipath for VNF 614
Configuring the Spine or Route Reflector 616
Configuring the Leaf or ToR 617
Configuring the Border Leaf or Border Gateway 623
Configuring the BGP Legacy Peer 629
Configure next-hop filtering 630
Verify next-hop filtering 632
Configuring a User-Defined Profile for Maintenance Mode 634
Configuring a User-Defined Profile for Normal Mode 634
Configuring a Default Route Map 635
Applying a Route Map to a Route Reflector 635
Verifying Proportional Multipath for VNF 636
Configuration Example for Proportional Multipath for VNF with Multi-Site 640

CHAPTER 26 **EVPN Distributed NAT** 647

EVPN Distributed NAT 647

CHAPTER 27 **DHCP Relay in VXLAN BGP EVPN** 653

DHCP Relay in VXLAN BGP EVPN Overview 653

Guidelines and Limitations for DHCP Relay in VXLAN BGP EVPN 654

DHCP Relay in VXLAN BGP EVPN supported release and platform 654

DHCP Relay in VXLAN BGP EVPN Example 655

DHCP Relay on VTEPs 655

Client on Tenant VRF and Server on Layer 3 Default VRF 656

Client on Tenant VRF (SVI X) and Server on the Same Tenant VRF (SVI Y) 659

Client on Tenant VRF (VRF X) and Server on Different Tenant VRF (VRF Y) 663

Client on Tenant VRF and Server on Non-Default Non-VXLAN VRF 666

Configuring vPC Peers Example 668

vPC VTEP DHCP Relay Configuration Example 670

CHAPTER 28 **Configure VXLAN Cross Connect** 673

VXLAN cross connect 673

Topology of VXLAN Cross Connect 674

Guidelines and Limitations of VXLAN Cross Connect 674

Supported platform and release of VXLAN cross connect 676

Prerequisites 676

Configure the VXLAN Cross Connect 676

Verification of VXLAN Cross Connect Configuration 678

Remove a Cross Connect VNI 680

CHAPTER 29 **Configuring Q-in-VNI over VXLAN** 683

Q-in-VNI over VXLAN 683

Guidelines and Limitations for Q-in-VNI 683

Configuring Q-in-VNI 686

Configuring Selective Q-in-VNI 687

Configuring Q-in-VNI with Layer 2 Protocol Tunneling 689

Q-in-VNI with L2PT Overview 689

Guidelines and Limitations for Q-in-VNI with L2PT 690
Configuring Q-in-VNI with L2PT 690
Verifying Q-in-VNI with L2PT Configuration 692
Configuring Q-in-VNI with LACP Tunneling 693
Selective Q-in-VNI with Multiple Provider VLANs 696
About Selective Q-in-VNI with Multiple Provider VLANs 696
Guidelines and Limitations for Selective Q-in-VNI with Multiple Provider VLANs 696
Configuring Selective Q-in-VNI with Multiple Provider VLANs 697
Configuring QinQ-QinVNI 699
Overview for QinQ-QinVNI 699
Guidelines and Limitations for QinQ-QinVNI 700
Configuring QinQ-QinVNI 700

CHAPTER 30 **Configuring Bud Node** 703
VXLAN Bud Node Over vPC Overview 703
VXLAN Bud Node Over vPC Topology Example 704

PART I **Configuring VXLAN Security** 709

CHAPTER 31 **Configuring Secure VXLAN EVPN Multi-Site Using CloudSec** 711
About Secure VXLAN EVPN Multi-Site Using CloudSec 711
Key Lifetime and Hitless Key Rollover 712
Certificate Expiration and Replacement 712
Guidelines and Limitations for Secure VXLAN EVPN Multi-Site Using CloudSec 712
Configuring Secure VXLAN EVPN Multi-Site Using CloudSec 714
Enabling CloudSec VXLAN EVPN Tunnel Encryption 714
Configuring a CloudSec Keychain and Keys 717
Configuring CloudSec Certificate Based Authentication Using PKI 718
Attaching a Certificate to CloudSec 718
Separate Loopback 719
Configuring a CloudSec Policy 720
Configuring CloudSec Peers 721
Configuring CloudSec Peers 721
Enabling Secure VXLAN EVPN Multi-Site Using CloudSec on DCI Uplinks 722

Verifying the Secure VXLAN EVPN Multi-Site Using CloudSec 723
Displaying Statistics for Secure VXLAN EVPN Multi-Site Using CloudSec 729
Configuration Examples for Secure VXLAN EVPN Multi-Site Using CloudSec 730
Migrating from Multi-Site with VIP to Multi-Site with PIP 732
Migration of Existing vPC BGW 732
vPC Border Gateway Support for Cloudsec 733
Enhanced Convergence for vPC BGW CloudSec Deployments 734
Migration from PSK CloudSec Configuration to Certificate Based Authentication CloudSec Configuration 735

CHAPTER 32 **Configuring VXLAN ACL 737**
About Access Control Lists 737
Guidelines and Limitations for VXLAN ACLs 738
VXLAN Tunnel Encapsulation Switch 738
Port ACL on the Access Port on Ingress 738
VLAN ACL on the Server VLAN 740
Routed ACL on an SVI on Ingress 741
Routed ACL on the Uplink on Egress 743
VXLAN Tunnel Decapsulation Switch 743
Routed ACL on the Uplink on Ingress 743
Port ACL on the Access Port on Egress 744
VLAN ACL for the Layer 2 VNI Traffic 744
VLAN ACL for the Layer 3 VNI Traffic 745
Routed ACL on an SVI on Egress 747

CHAPTER 33 **Configuring PVLANs 749**
About Private VLANs over VXLAN 749
Guidelines and Limitations for Private VLANs over VXLAN 750
Configuration Example for Private VLANs 751

CHAPTER 34 **Configure First Hop Security 753**
DHCP Snooping in VXLAN BGP EVPN 753
How DHCP Snooping on VXLAN works 754
Guidelines and limitations for DHCP Snooping on VXLAN 756

Prerequisites for DHCP Snooping 758
Enable DHCP Snooping on VXLAN 758
Clear the duplicate host after permanent freeze 759
Verification of DHCP Snooping configuration 760

PART II **Configure Multi-Homing** 761

Cisco Nexus supported Multi-homing methods 761

CHAPTER 35 **Configuring vPC Multi-Homing** 763
Advertising Primary IP Address 763
Border PE Switches in a vPC Setup 764
DHCP Configuration in a vPC Setup 764
IP Prefix Advertisement in vPC Setup 764

CHAPTER 36 **Configuring vPC Fabric Peering** 765
Information About vPC Fabric Peering 765
Guidelines and Limitations for vPC Fabric Peering 766
Configuring vPC Fabric Peering 768
Migrating from vPC to vPC Fabric Peering 772
Verifying vPC Fabric Peering Configuration 775

CHAPTER 37 **Interoperability with EVPN Multi-Homing Using ESI** 777
Interoperability with EVPN Multi-Homing Using ESI 777
Guidelines and Limitations for Interoperability with EVPN Multi-Homing using ESI 778
Example of EVPN Multi-Homing Using ESI 779

CHAPTER 38 **EVPN Ethernet Segment Identifier Multi-Homing** 783
EVPN ESI multi-homing 783
EVPN ESI multi-homing solutions 783
EVPN multi-homing terms and definitions 784
EVPN multi-homing information 785
ESI multi-homing operation modes 786
Layer 2 fast reroute 792

Designated Forwarder elections 793
EVPN EAD-EVI routes 795
Type-2 route re-origination 796
Ethernet Segment delay-restore timers 796
How ESI multi-homing topologies work 796
Guidelines and limitations for ESI multi-homing 802
Enable EVPN ESI multi-homing 804
Configure ESI under a port channel interface 808
Verification commands for ESI multi-homing configuration 812
ESI multi-homing configuration examples 814
IGMP or MLD Snooping with ESI - EVPN Type-7 and Type-8 route 852
IGMP or MLD Snooping with ESI 853
Supported L2 multicast topologies and IGMP control flows 854
Guidelines and limitations for IGMP and MLD snooping with ESI 857
IGMP and MLD snooping ESI synchronization command outputs 858
ECMP reuse 860
ECMP reuse 860
Guidelines and limitations for ECMP reuse 860
Enable ECMP reuse 860
Layer 2 Gateway Spanning Tree Protocol 861
Layer 2 Gateway Spanning Tree Protocol 861
Best practice for moving to Layer 2 Gateway STP 862
Enable Layer 2 Gateway STP on a switch 863
Troubleshooting Layer 2 Gateway STP configurations 865



Preface

This preface includes the following sections:

- Audience, on page xxiii
- Document Conventions, on page xxiii
- Related Documentation for Cisco Nexus 9000 Series Switches, on page xxiv
- Documentation Feedback, on page xxiv
- Communications, services, and additional information, on page xxiv

Audience

This publication is for network administrators who install, configure, and maintain Cisco Nexus switches.

Document Conventions

Command descriptions use the following conventions:

Convention	Description
bold	Bold text indicates the commands and keywords that you enter literally as shown.
Italic	Italic text indicates arguments for which you supply the values.
[x]	Square brackets enclose an optional element (keyword or argument).
[x | y]	Square brackets enclosing keywords or arguments that are separated by a vertical bar indicate an optional choice.
{x | y}	Braces enclosing keywords or arguments that are separated by a vertical bar indicate a required choice.
[x {y | z}]	Nested set of square brackets or braces indicate optional or required choices within optional or required elements. Braces and a vertical bar within square brackets indicate a required choice within an optional element.

Convention	Description
`variable`	Indicates a variable for which you supply values, in context where italics cannot be used.
string	A nonquoted set of characters. Do not use quotation marks around the string or the string includes the quotation marks.

Examples use the following conventions:

Convention	Description
`screen font`	Terminal sessions and information the switch displays are in screen font.
`boldface screen font`	Information that you must enter is in boldface screen font.
italic screen font	Arguments for which you supply values are in italic screen font.
< >	Nonprinting characters, such as passwords, are in angle brackets.
[ ]	Default responses to system prompts are in square brackets.
!, #	An exclamation point (!) or a pound sign (#) at the beginning of a line of code indicates a comment line.

Related Documentation for Cisco Nexus 9000 Series Switches

The entire Cisco Nexus 9000 Series switch documentation set is available at the following URL:

https://www.cisco.com/en/US/products/ps13386/tsd_products_support_series_home.html

Documentation Feedback

To provide technical feedback on this document, or to report an error or omission, please send your comments to nexus9k-docfeedback@cisco.com. We appreciate your feedback.

Communications, services, and additional information

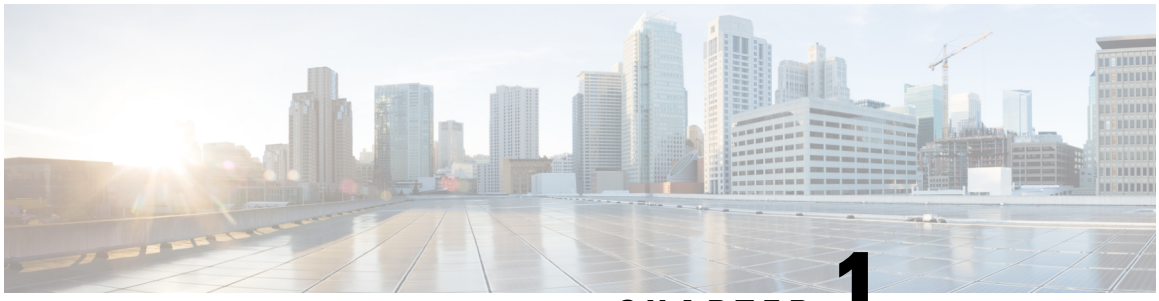
- To receive timely, relevant information from Cisco, sign up at Cisco Profile Manager.
- To get the business impact you're looking for with the technologies that matter, visit Cisco Services.
- To submit a service request, visit Cisco Support.
- To discover and browse secure, validated enterprise-class apps, products, solutions, and services, visit Cisco DevNet.
- To obtain general networking, training, and certification titles, visit Cisco Press.
- To find warranty information for a specific product or product family, access Cisco Warranty Finder.

Cisco Bug Search Tool

Cisco Bug Search Tool (BST) is a gateway to the Cisco bug-tracking system, which maintains a comprehensive list of defects and vulnerabilities in Cisco products and software. The BST provides you with detailed defect information about your products and software.

Documentation feedback

To provide feedback about Cisco technical documentation, use the feedback form available in the right pane of every online document.



CHAPTER 1

New and Changed Information

- New and Changed Information, on page 1

New and Changed Information

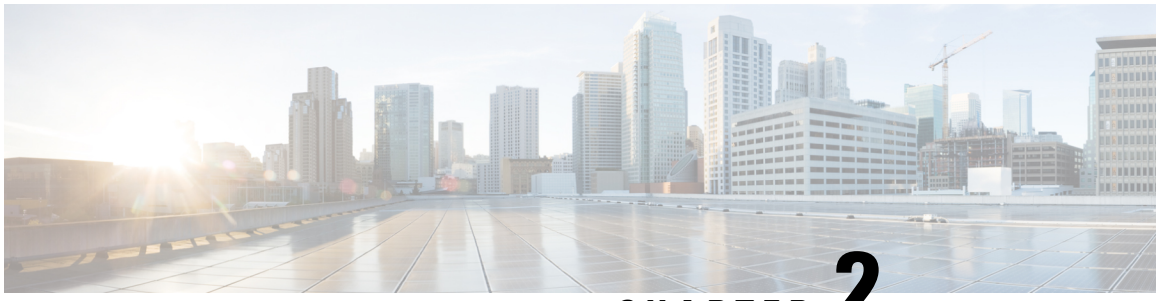
Table 1: New and Changed Features

Feature	Description	Changed in Release	Where Documented
EVPN ESI Multi-Homing	Added EVPN ESI multi-homing feature which is a standardized solution that uses BGP-EVPN based communication between ESI peers. This feature provides redundancy similar to vPC and vPC fabric peering. It supports 2-way, 3-way, and 4-way multi-homing on PIP or VIP modes.	10.6(1)F	Topology of Multicast Layer 2 EVPN Type-6 routes (Reference), on page 446 EVPN Ethernet Segment Identifier Multi-Homing, on page 783

Feature	Description	Changed in Release	Where Documented
VXLAN support on Cisco Nexus 9336C-SE1 switches	Added VXLAN support on Cisco Nexus 9336C-SE1 switches.	10.6(1)F	Guidelines and limitations for VXLAN support on Cisco Nexus 9336C-SE1 switches, on page 55 Guidelines and Limitations for VXLAN BGP EVPN, on page 115 Guidelines and limitations for VXLAN Multi-Site Anycast BGW support on Cisco Nexus 9336C-SE1 Switches, on page 301 Guidelines and Limitations for VXLAN EVPN with Downstream VNI, on page 123 Guidelines and Limitations for VXLAN NGOAM, on page 460 DHCP Relay in VXLAN BGP EVPN supported release and platform, on page 654 Guidelines and Limitations for vPC Fabric Peering , on page 766
TRMv4 support on Cisco Nexus 9336C-SE1 switches	Added TRMv4 support on Cisco Nexus 9336C-SE1 switches for • Standalone, vPC, vPC fabric peering inner and border leaf • Multi-site Anycast BGW.	10.6(1)F	Guidelines and Limitations for Tenant Routed Multicast, on page 385 Guidelines and Limitations for Layer 3 Tenant Routed Multicast, on page 387
L2VLAN segmentation with SVI and without SVI	Added support for MAC based L2 VLAN micro-segmentation with and without SVI.	10.6(1)F	Guidelines and Limitations, on page 555 Creating Security Group using MAC, on page 564
Micro-segmentaton Match Criteria: Port+VLAN	Added support for port along with VLAN based micro-segmentation.	10.6(1)F	Guidelines and Limitations, on page 555 Creating a Security Group, on page 561

Feature	Description	Changed in Release	Where Documented
PIM graceful-spt feature is now default	Added support to make PIM graceful-spt feature a default feature.	10.6(1)F	Guidelines and Limitations for Tenant Routed Multicast, on page 385
Match Criteria: Port+VLAN	Added support for port along with VLAN based micro-segmentation.	10.6(1)F	Guidelines and Limitations, on page 555 Creating a Security Group, on page 561
VXLAN Anycast BGW (L2VNI) and RFC5549	Added L2VNI and RFC5549 underlay support for VXLAN Anycast BGW on Cisco Nexus 9364E-SG2-Q and 9364E-SG2-O switches.	10.6(1)F	Guidelines and Limitations for VXLAN, on page 43 Guidelines for VXLAN BGP EVPN with RFC 5549 Underlay, on page 185 Guidelines for VXLAN EVPN Multi-Site with RFC 5549 Underlay, on page 335
SMET route Interoperability support	Added a new **format rfc6625** encoding option to the **advertise evpn multicast** command for EVPN route Type 6, 7 and 8 to address interoperability issues between Cisco NX-OS and other vendors devices due to different encoding methods for multicast routes.	10.6(1)F	Guidelines and Limitations of Multicast Layer 2 EVPN Type-6 routes, on page 450 Configure IPv4 IGMP Snooping, on page 452

New and Changed Information



CHAPTER 2

Overview

This chapter contains these sections:

- Licensing Requirements, on page 5
- Supported Platforms, on page 5
- VXLAN Overview, on page 5
- Cisco Nexus 9000 as Hardware-Based VXLAN Gateway, on page 6
- VXLAN Encapsulation and Packet Format, on page 6
- VXLAN Tunnel, on page 7
- VXLAN Tunnel Endpoint, on page 7
- Underlay Network, on page 7
- Overlay Network, on page 7
- Distributed Anycast Gateway, on page 7
- Control Plane, on page 8

Licensing Requirements

For a complete explanation of Cisco NX-OS licensing recommendations and how to obtain and apply licenses, see the *Cisco NX-OS Licensing Guide* and the *Cisco NX-OS Licensing Options Guide* .

Supported Platforms

Use the Nexus Switch Platform Support Matrix to know from which Cisco NX-OS releases various Cisco Nexus 9000 and 3000 switches support a selected feature.

VXLAN Overview

Virtual Extensible LAN (VXLAN) provides a way to extend Layer 2 networks across a Layer 3 infrastructure using MAC-in-UDP encapsulation and tunneling. This feature enables virtualized and multitenant data center fabric designs over a shared common physical infrastructure.

VXLAN has the following benefits:

- Flexible placement of workloads across the data center fabric.

It provides a way to extend Layer 2 segments over the underlying shared Layer 3 network infrastructure so that tenant workloads can be placed across physical pods in a single data center. Or even across several geographically divers data centers.

- Higher scalability to allow more Layer 2 segments.

  VXLAN uses a 24-bit segment ID, the VXLAN network identifier (VNID). This allows a maximum of 16 million VXLAN segments to coexist in the same administrative domain. In comparison, traditional VLANs use a 12-bit segment ID that can support a maximum of 4096 VLANs.

- Optimized utilization of available network paths in the underlying infrastructure.

  VXLAN packets are transferred through the underlying network based on their Layer 3 headers. They use equal-cost multipath (ECMP) routing and link aggregation protocols to use all available paths. In contrast, a Layer 2 network might block valid forwarding paths in order to avoid loops.

Cisco Nexus 9000 as Hardware-Based VXLAN Gateway

A Cisco Nexus 9000 Series switch can function as a hardware-based VXLAN gateway. It seamlessly connects VXLAN and VLAN segments as one forwarding domain across the Layer 3 boundary without sacrificing forwarding performance. The Cisco Nexus 9000 Series hardware-based VXLAN encapsulation and de-encapsulation provide line-rate performance for all frame sizes.

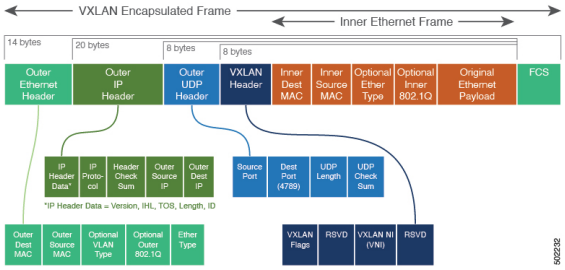
VXLAN Encapsulation and Packet Format

VXLAN is a Layer 2 overlay scheme over a Layer 3 network. It uses a MAC Address-in-User Datagram Protocol (MAC-in-UDP) encapsulation to provide a means to extend Layer 2 segments across the data center network. VXLAN is a solution to support a flexible, large-scale multitenant environment over a shared common physical infrastructure. The transport protocol over the physical data center network is IP plus UDP.

VXLAN defines a MAC-in-UDP encapsulation scheme where the original Layer 2 frame has a VXLAN header added and is then placed in a UDP-IP packet. With this MAC-in-UDP encapsulation, VXLAN tunnels Layer 2 network over Layer 3 network.

VXLAN uses an 8-byte VXLAN header that consists of a 24-bit VNID and a few reserved bits. The VXLAN header, together with the original Ethernet frame, go inside the UDP payload. The 24-bit VNID is used to identify Layer 2 segments and to maintain Layer 2 isolation between the segments. With all 24 bits in the VNID, VXLAN can support 16 million LAN segments.

Figure 1:



VXLAN Tunnel

A VXLAN encapsulated communication between two devices where they encapsulate and decapsulate an inner Ethernet frame, is called a VXLAN tunnel. VXLAN tunnels are stateless since they are UDP encapsulated.

VXLAN Tunnel Endpoint

VXLAN tunnel endpoints (VTEPs) are devices that terminate VXLAN tunnels. They perform VXLAN encapsulation and de-encapsulation. Each VTEP has two interfaces. One is a Layer 2 interface on the local LAN segment to support a local endpoint communication through bridging. The other is a Layer 3 interface on the IP transport network.

The IP interface has a unique address that identifies the VTEP device in the transport network. The VTEP device uses this IP address to encapsulate Ethernet frames and transmit the packets on the transport network. A VTEP discovers other VTEP devices that share the same VNIs it has locally connected. It advertises the locally connected MAC addresses to its peers. It also learns remote MAC Address-to-VTEP mappings through its IP interface.

Underlay Network

The VXLAN segments are independent of the underlying physical network topology. Conversely, the underlying IP network, often referred to as the underlay network, is independent of the VXLAN overlay. The underlay network forwards the VXLAN encapsulated packets based on the outer IP address header. The outer IP address header has the initiating VTEP's IP interface as the source IP address and the terminating VTEP's IP interface as the destination IP address.

The primary purpose of the underlay in the VXLAN fabric is to advertise the reachability of the Virtual Tunnel Endpoints (VTEPs). The underlay also provides a fast and reliable transport for the VXLAN traffic.

Overlay Network

In broadcast terms, an overlay is a virtual network that is built on top of an underlay network infrastructure. In a VXLAN fabric, the overlay network is built of a control plane and the VXLAN tunnels. The control plane is used to advertise MAC address reachability. The VXLAN tunnels transport the Ethernet frames between the VTEPs.

Distributed Anycast Gateway

Distributed Anycast Gateway refers to the use of default gateway addressing that uses the same IP and MAC address across all the leafs that are a part of a VNI. This ensures that every leaf can function as the default gateway for the workloads directly connected to it. The distributed Anycast Gateway functionality is used to facilitate flexible workload placement, and optimal traffic forwarding across the VXLAN fabric.

Control Plane

There are two widely adopted control planes that are used with VXLAN:

Flood and Learn Multicast-Based Learning Control Plane

Cisco Nexus 9000 Series switches support the flood and learn multicast-based control plane method.

- When configuring VXLAN with a multicast based control plane, every VTEP configured with a specific VXLAN VNI joins the same multicast group. Each VNI could have its own multicast group, or several VNIs can share the same group.
- The multicast group is used to forward broadcast, unknown unicast, and multicast (BUM) traffic for a VNI.
- The multicast configuration must support Any-Source Multicast (ASM) or PIM BiDir.
- Initially, the VTEPs only learn the MAC addresses of devices that are directly connected to them.
- Remote MAC address to VTEP mappings are learned via conversational learning.

VXLAN MPBGP EVPN Control Plane

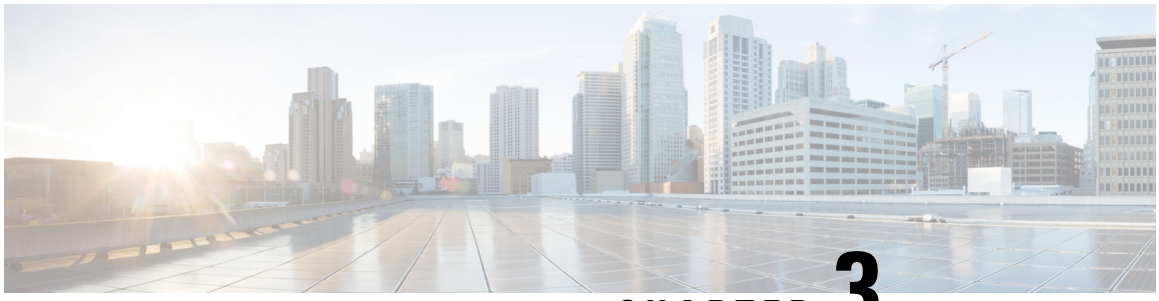
A Cisco Nexus 9000 Series switch can be configured to provide a Multiprotocol Border Gateway Protocol (MPBGP) ethernet VPN (EVPN) control plane. The control plane uses a distributed Anycast Gateway with Layer 2 and Layer 3 VXLAN overlay networks.

For a data center network, an MPBGP EVPN control plane provides:

- Flexible workload placement that is not restricted with physical topology of the data center network.
    - Place virtual machines anywhere in the data center fabric.
- Optimal East-West traffic between servers within and across data centers
    - East-West traffic between servers, or virtual machines, is achieved by most specific routing at the first hop router. First hop routing is done at the access layer. Host routes must be exchanged to ensure most specific routing to and from servers or hosts. Virtual machine (VM) mobility is supported by detecting new endpoint attachment when a new MAC address/IP address is seen directly connected to the local switch. When the local switch sees the new MAC/IP, it signals the new location to rest of the network.
- Eliminate or reduce flooding in the data center.
    - Flooding is reduced by distributing MAC reachability information via MP-BGP EVPN to optimize flooding relating to L2 unknown unicast traffic. Optimization of reducing broadcasts associated with ARP/IPv6 Neighbor solicitation is achieved by distributing the necessary information via MPBGP EVPN. The information is then cached at the access switches. Address solicitation requests can be responded locally without sending a broadcast to the rest of the fabric.
- A standards-based control plane that can be deployed independent of a specific fabric controller.
    - The MPBGP EVPN control plane approach provides:

- IP reachability information for the tunnel endpoints associated with a segment and the hosts behind a specific tunnel endpoint.
- Distribution of host MAC reachability to reduce/eliminate unknown unicast flooding.
- Distribution of host IP/MAC bindings to provide local ARP suppression.
- Host mobility.
- A single address family (MPBGP EVPN) to distribute both L2 and L3 route reachability information.

- Segmentation of Layer 2 and Layer 3 traffic
  - Traffic segmentation is achieved with using VXLAN encapsulation, where VNI acts as segment identifier.



CHAPTER 3

Configuring the Underlay

This chapter contains these sections:

- IP Fabric Underlay, on page 11

IP Fabric Underlay

Underlay Considerations

Unicast Underlay:

The primary purpose of the underlay in the VXLAN EVPN fabric is to advertise the reachability of Virtual Tunnel End Points (VTEPs) and BGP peering addresses. The primary criterion for choosing an underlay protocol is fast convergence in the event of node failures. Other criteria are:

- Simplicity of configuration.
- Ability to delay the introduction of a node into the network on boot up.

This document details the two primary protocols supported and tested by Cisco, IS-IS and OSPF. It will also illustrate the use of the eBGP protocol as an underlay for the VXLAN EVPN fabric.

For Configuring VXLAN with IPv6 in the Underlay (VXLANv6), see Configuring VXLAN with IPv6 in the Underlay (VXLANv6), on page 91.

From an underlay/overlay perspective, the packet flow from a server to another over the Virtual Extensible LAN (VXLAN) fabric as mentioned below:

1. The server sends traffic to the source VXLAN tunnel endpoint (VTEP). The VTEP performs Layer-2 or Layer-3 communication based on the destination MAC and derives the nexthop (destination VTEP).

Note When a packet is bridged, the target end host's MAC address is stamped in the DMAC field of the inner frame. When a packet is routed, the default gateway's MAC address is stamped in the DMAC field of the inner frame.

2. The VTEP encapsulates the traffic (frames) into VXLAN packets (overlay function – see Figure 1) and signals the underlay IP network.

3. Based on the underlay routing protocol, the packet is sent from the source VTEP to destination VTEP through the IP network (underlay function – see *Underlay Overview* figure).
4. The destination VTEP removes the VXLAN encapsulation (overlay function) and sends traffic to the intended server.

The VTEPs are a part of the underlay network as well since VTEPs need to be reachable to each other to send VXLAN encapsulated traffic across the IP underlay network.

The *Overlay Overview* and *Underlay Overview* images (below) depict the broad difference between an overlay and underlay. Since the focus is on the VTEPs, the spine switches are only depicted in the background. Note that, in real time, the packet flow from VTEP to VTEP traverses through the spine switches.

Figure 2: Overlay Overview

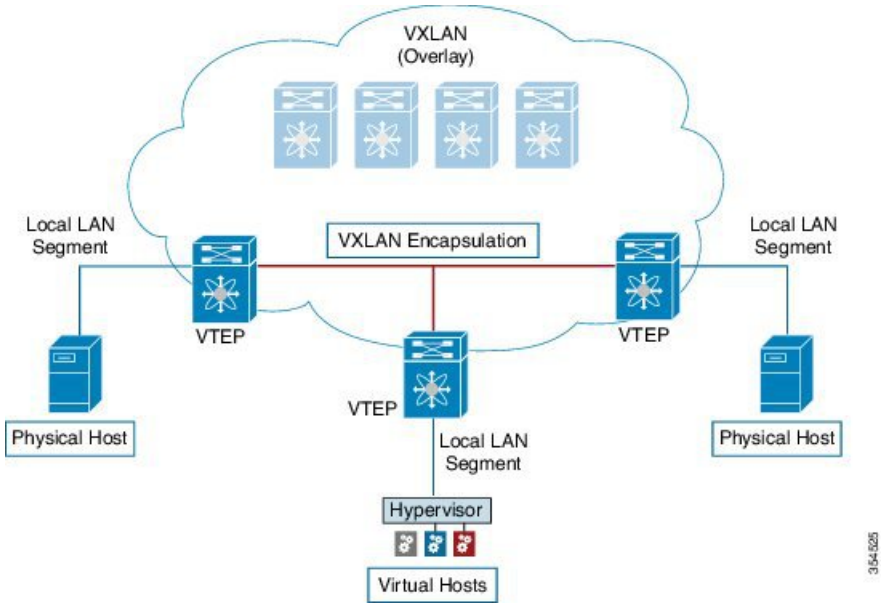
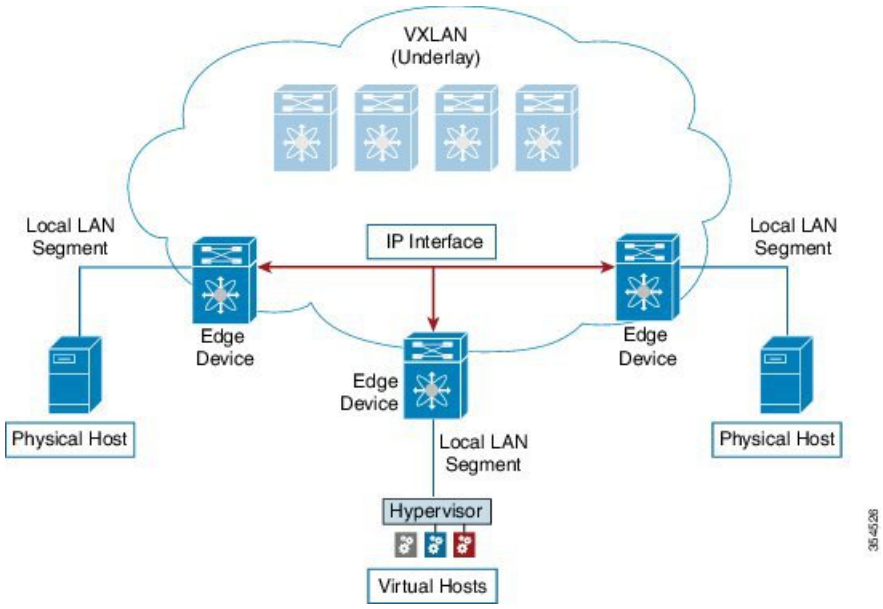


Figure 3: Underlay Overview



Deployment considerations for an underlay IP network in a VXLAN EVPN Programmable Fabric

The deployment considerations for an underlay IP network in a VXLAN EVPN Programmable Fabric are given below:

- Maximum transmission unit (MTU) – Due to VXLAN encapsulation, the MTU requirement is larger and we must avoid potential fragmentation.
  - An MTU of 9216 bytes on each interface on the path between the VTEPs accommodates maximum server MTU + VXLAN overhead. Most data center server NICs support up to 9000 bytes. So, no fragmentation is needed for VXLAN traffic.
  - The VXLAN IP fabric underlay supports the IPv4 address family.
- Unicast routing - Any unicast routing protocol can be used for the VXLAN IP underlay. You can implement OSPF, IS-IS, or eBGP to route between the VTEPs.

Note As a best practice, use a simple IGP (OSPF or IS-IS) for underlay reachability between VTEPs with iBGP for overlay information exchange.

- IP addressing – Point-to-point (P2P) or IP unnumbered links. For each point-to-point link, as example between the leaf switch nodes and spine switch nodes, typically a /30 IP mask should be assigned. Optionally a /31 mask or IP unnumbered links can be assigned. The IP unnumbered approach is leaner from an addressing perspective and consumes fewer IP addresses. The IP unnumbered option for the OSPF or IS-IS protocol underlay will minimize the use of IP addresses.

  /31 network - An OSPF or IS-IS point-to-point numbered network is only between two switch (interfaces), and there is no need for a broadcast or network address. So, a /31 network suffices for this network. Neighbors on this network establish adjacency and there is no designated router (DR) for the network.

Note IP Unnumbered for VXLAN underlay is supported starting with Cisco NX-OS Release 7.0(3)I7(2). Only a single unnumbered link between the same devices (for example, spine - leaf) is supported. If multiple physical links are connecting the same leaf and spine, you must use the single L3 port-channel with unnumbered link.

- Multicast protocol for multi-destination (BUM) traffic – Though VXLAN has the BGP EVPN control plane, the VXLAN fabric still requires a technology for Broadcast/Unknown unicast/Multicast (BUM) traffic to be forwarded.
- PIM Bidir is supported on Cisco Nexus 9300-FX/FX2 platform switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, PIM Bidir is also supported on Cisco Nexus 9300-FX3/GX/GX2/H2R/H1 platform switches, and 9500 switches with 9700-GX line cards.
- Beginning with Cisco NX-OS Release 10.3(5)M, PIM Bidir is also supported on Cisco Nexus 9300-FX3/GX/GX2/H2R/H1 platform switches, and 9500 switches with 9700-GX line cards.
- Beginning with Cisco NX-OS Release 10.5(2)F, VXLAN PIM Bidir is supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.
- vPC configuration — This is documented in **Configuring vPCs** of *Cisco Nexus 9000 Series NX-OS Interfaces Configuration Guide*.

Unicast routing and IP addressing options

Each unicast routing protocol option (OSPF, IS-IS, and eBGP) and sample configurations are given below. Use an option to suit your setup's requirements.

Important All routing configuration samples are from an IP underlay perspective and are not comprehensive. For complete configuration information including routing process, authentication, Bidirectional Forwarding Detection (BFD) information, and so on, see *Cisco Nexus 9000 Series NX-OS Unicast Routing Configuration Guide*.

OSPF Underlay IP Network

Some considerations are given below:

- For IP addressing, use P2P links. Since only two switches are directly connected, you can avoid a Designated Router/Backup Designated Router (DR/BDR) election.
- Use the *point-to-point* network type option. It is ideal for routed interfaces or ports, and is optimal from a Link State Advertisements (LSA) perspective.
- Do not use the broadcast type network. It is suboptimal from an LSA database perspective (LSA type 1 – Router LSA and LSA type 2 – Network LSA) and necessitates a DR/BDR election, thereby creating an additional election and database overhead.

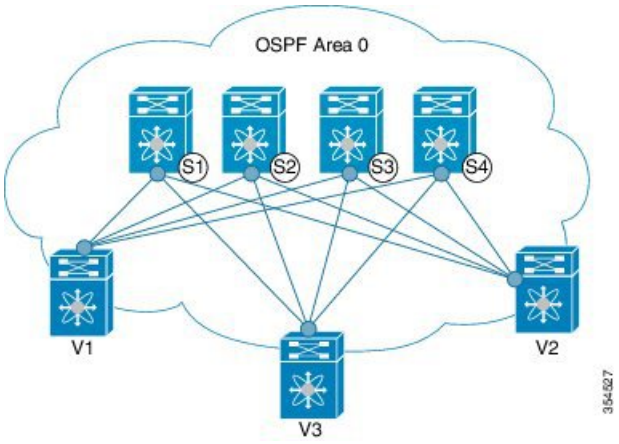
Note You can divide OSPF networks into areas when the size of the routing domain contains a high number of routers and/or IP prefixes.. The same general well known OSPF best practice rules in regards of scale and configuration are applicable for the VXLAN underlay too. For example, LSA type 1 and type 2 are never flooded outside of an area. With multiple areas, the size of the OSPF LSA databases can be reduced to optimize CPU and memory consumption.

Note

- For ease of use, the configuration mode from which you need to start configuring a task is mentioned at the beginning of each configuration.
- Configuration tasks and corresponding show command output are displayed for a part of the topology in the image. For example, if the sample configuration is shown for a leaf switch and connected spine switch, the show command output for the configuration displays corresponding configuration.

OSPF configuration sample – P2P and IP unnumbered network scenarios

Figure 4: OSPF as the underlay routing protocol



OSPF – P2P link scenario with /31 mask

In the above image, the leaf switches (V1, V2, and V3) are at the bottom of the image. They are connected to the 4 spine switches (S1, S2, S3, and S4) that are depicted at the top of the image. For P2P connections between a leaf switch (also having VTEP function) and each spine, leaf switches V1, V2, and V3 should each be connected to each spine switch.

For V1, we should configure a P2P interface to connect to each spine switch.

A sample P2P configuration between a leaf switch (V1) interface and a spine switch (S1) interface is given below:

OSPF global configuration on leaf switch V1

(config) #

```
feature ospf
router ospf UNDERLAY
 router-id 10.1.1.54
```

OSPF leaf switch V1 P2P interface configuration

(config) #

```
interface Ethernet 1/41
   description Link to Spine S1
   no switchport
   ip address 198.51.100.1/31
   mtu 9192
   ip router ospf UNDERLAY area 0.0.0.0
   ip ospf network point-to-point
```

The **ip ospf network point-to-point** command configures the OSPF network as a point-to-point network

The OSPF instance is tagged as UNDERLAY for better recall.

OSPF loopback interface configuration (leaf switch V1)

Configure a loopback interface so that it can be used as the OSPF router ID of leaf switch V1.

(config) #

```
interface loopback 0
   ip address 10.1.1.54/32
   ip router ospf UNDERLAY area 0.0.0.0
```

The interface will be associated with the OSPF instance UNDERLAY and OSPF area 0.0.0.0

OSPF global configuration on spine switch S1

(config) #

```
feature ospf
router ospf UNDERLAY
 router-id 10.1.1.53
```

(Corresponding) OSPF spine switch S1 P2P interface configuration

(config) #

```
interface Ethernet 1/41
   description Link to VTEP V1
   ip address 198.51.100.2/31
   mtu 9192
   ip router ospf UNDERLAY area 0.0.0.0
   ip ospf network point-to-point
   no shutdown
```



Note MTU size of both ends of the link should be configured identically.

OSPF loopback Interface Configuration (spine switch S1)

Configure a loopback interface so that it can be used as the OSPF router ID of spine switch S1.

(config) #

```
interface loopback 0
```

```
   ip address 10.1.1.53/32
   ip router ospf UNDERLAY area 0.0.0.0
```

The interface will be associated with the OSPF instance UNDERLAY and OSPF area 0.0.0.0

.

.

To complete OSPF topology configuration for the 'OSPF as the underlay routing protocol' image, configure the following

- *3 more V1 interfaces (or 3 more P2P links) to the remaining 3 spine switches.*
- *Repeat the procedure to connect P2P links between V2,V3 and V4 and the spine switches.*

OSPF - IP unnumbered scenario

A sample OSPF IP unnumbered configuration is given below:

OSPF leaf switch V1 configuration

OSPF global configuration on leaf switch V1

(config) #

```
feature ospf
router ospf UNDERLAY
  router-id 10.1.1.54
```

The OSPF instance is tagged as UNDERLAY for better recall.

OSPF leaf switch V1 P2P interface configuration

(config) #

```
interface Ethernet1/41
  description Link to Spine S1
  mtu 9192
  ip ospf network point-to-point
  ip unnumbered loopback0
  ip router ospf UNDERLAY area 0.0.0.0
```

The **ip ospf network point-to-point** command configures the OSPF network as a point-to-point network.

OSPF loopback interface configuration

Configure a loopback interface so that it can be used as the OSPF router ID of leaf switch V1.

(config) #

```
interface loopback0
  ip address 10.1.1.54/32
  ip router ospf UNDERLAY area 0.0.0.0
```

The interface will be associated with the OSPF instance UNDERLAY and OSPF area 0.0.0.0

OSPF spine switch S1 configuration:

OSPF global configuration on spine switch S1

(config) #

```
feature ospf
router ospf UNDERLAY
  router-id 10.1.1.53
```

(Corresponding) OSPF spine switch S1 P2P interface configuration

(config) #

```
interface Ethernet1/41
  description Link to VTEP V1
  mtu 9192
  ip ospf network point-to-point
  ip unnumbered loopback0
  ip router ospf UNDERLAY area 0.0.0.0
```

OSPF loopback interface configuration (spine switch S1)

Configure a loopback interface so that it can be used as the OSPF router ID of spine switch S1.

(config) #

```
interface loopback0
  ip address 10.1.1.53/32
  ip router ospf UNDERLAY area 0.0.0.0
```

The interface will be associated with the OSPF instance UNDERLAY and OSPF area 0.0.0.0

.

.

To complete OSPF topology configuration for the 'OSPF as the underlay routing protocol' image, configure the following:

- *3 more VTEP V1 interfaces (or 3 more IP unnumbered links) to the remaining 3 spine switches.*
- *Repeat the procedure to connect IP unnumbered links between VTEPs V2,V3 and V4 and the spine switches.*

OSPF Verification

Use the following commands for verifying OSPF configuration:

```
Leaf-Switch-V1# show ip ospf

 Routing Process UNDERLAY with ID 10.1.1.54 VRF default
 Routing Process Instance Number 1
 Stateful High Availability enabled
 Graceful-restart is configured
   Grace period: 60 state: Inactive
   Last graceful restart exit status: None
 Supports only single TOS(TOS0) routes
 Supports opaque LSA
 Administrative distance 110
 Reference Bandwidth is 40000 Mbps
 SPF throttling delay time of 200.000 msecs,
   SPF throttling hold time of 1000.000 msecs,
   SPF throttling maximum wait time of 5000.000 msecs
 LSA throttling start time of 0.000 msecs,
   LSA throttling hold interval of 5000.000 msecs,
   LSA throttling maximum wait time of 5000.000 msecs
 Minimum LSA arrival 1000.000 msec
```

```
 LSA group pacing timer 10 secs
 Maximum paths to destination 8
 Number of external LSAs 0, checksum sum 0
 Number of opaque AS LSAs 0, checksum sum 0
 Number of areas is 1, 1 normal, 0 stub, 0 nssa
 Number of active areas is 1, 1 normal, 0 stub, 0 nssa
 Install discard route for summarized external routes.
 Install discard route for summarized internal routes.
   Area BACKBONE(0.0.0.0)
        Area has existed for 03:12:54
        Interfaces in this area: 2 Active interfaces: 2
        Passive interfaces: 0  Loopback interfaces: 1
        No authentication available
        SPF calculation has run 5 times
         Last SPF ran for 0.000195s
        Area ranges are
        Number of LSAs: 3, checksum sum 0x196c2
```

```
Leaf-Switch-V1# show ip ospf interface

loopback0 is up, line protocol is up
    IP address 10.1.1.54/32
    Process ID UNDERLAY VRF default, area 0.0.0.0
    Enabled by interface configuration
    State LOOPBACK, Network type LOOPBACK, cost 1
    Index 1
 Ethernet1/41 is up, line protocol is up
    Unnumbered interface using IP address of loopback0 (10.1.1.54)
    Process ID UNDERLAY VRF default, area 0.0.0.0
    Enabled by interface configuration
    State P2P, Network type P2P, cost 4
    Index 2, Transmit delay 1 sec
    1 Neighbors, flooding to 1, adjacent with 1
    Timer intervals: Hello 10, Dead 40, Wait 40, Retransmit 5
      Hello timer due in 00:00:07
    No authentication
    Number of opaque link LSAs: 0, checksum sum 0
```

```
Leaf-Switch-V1# show ip ospf neighbors

OSPF Process ID UNDERLAY VRF default
 Total number of neighbors: 1
 Neighbor ID     Pri State            Up Time  Address         Interface
 10.1.1.53         1 FULL/ -          06:18:32 10.1.1.53        Eth1/41
```

For a detailed list of commands, refer to the Configuration and Command Reference guides.

IS-IS Underlay IP Network

Some considerations are given below:

- Because IS-IS uses Connectionless Network Service (CLNS) and is independent of the IP, full SPF calculation is avoided when a link changes.
- Net ID - Each IS-IS instance has an associated network entity title (NET) ID that uniquely identifies the IS-IS instance in the area. The NET ID is comprised of the IS-IS system ID, which uniquely identifies this IS-IS instance in the area, and the area ID. For example, if the NET ID is 49.0001.0010.0100.1074.00, the system ID is 0010.0100.1074 and the area ID is 49.0001.

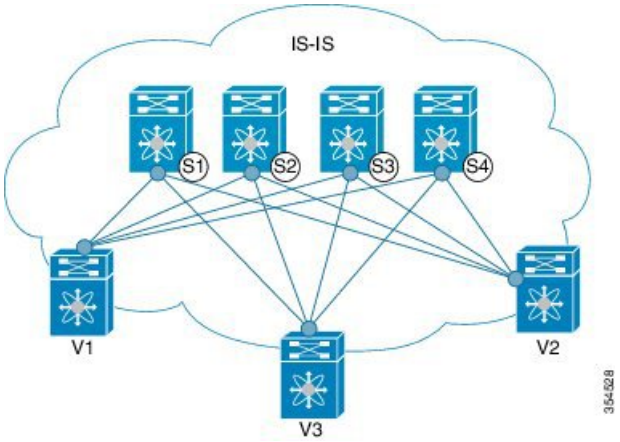
Important **Level 1 IS-IS in the Fabric**—Cisco has validated the use of IS-IS Level 1 only and IS-IS Level 2 only configuration on all nodes in the programmable fabric. The fabric is considered a stub network where every node needs an optimal path to every other node in the fabric. Cisco NX-OS IS-IS implementation scales well to support a number of nodes in a fabric. Hence, there is no anticipation of having to break up the fabric into multiple IS-IS domains.

Note

- For ease of use, the configuration mode from which you need to start configuring a task is mentioned at the beginning of each configuration.
- Configuration tasks and corresponding show command output are displayed for a part of the topology in the image. For example, if the sample configuration is shown for a leaf switch and connected spine switch, the show command output for the configuration displays corresponding configuration.

IS-IS configuration sample - P2P and IP unnumbered network scenarios

Figure 5: IS-IS as the underlay routing protocol



In the above image, the leaf switches (V1, V2, and V3, having the VTEP function) are at the bottom of the image. They are connected to the 4 spine switches (S1, S2, S3, and S4) that are depicted at the top of the image.

IS-IS – P2P link scenario with /31 mask

A sample P2P configuration between V1 and spine switch S1 is given below:

For P2P connections between a leaf switch and each spine switch, V1, V2, and V3 should each be connected to each spine switch.

For V1, we must configure a loopback interface and a P2P interface configuration to connect to S1. A sample P2P configuration between a leaf switch (V1) interface and a spine switch (S1) interface is given below:

IS-IS configuration on leaf switch V1

IS-IS global configuration

(config) #

```
feature isis
router isis UNDERLAY
  net 49.0001.0010.0100.1074.00
  is-type level-1
  set-overload-bit on-startup 60
```

Setting the overload bit - You can configure a Cisco Nexus switch to signal other devices not to use the switch as an intermediate hop in their shortest path first (SPF) calculations. You can optionally configure the overload bit temporarily on startup. In the above example, the **set-overload-bit** command is used to set the overload bit on startup to 60 seconds.

IS-IS P2P interface configuration (leaf switch V1)

(config) #

```
interface Ethernet 1/41
   description Link to Spine S1
   mtu 9192
   ip address 209.165.201.1/31
   ip router isis UNDERLAY
```

IS-IS loopback interface configuration (leaf switch V1)

Configure a loopback interface so that it can be used as the IS-IS router ID of leaf switch V1.

(config) #

```
interface loopback 0
   ip address 10.1.1.74/32
   ip router isis UNDERLAY
```

The IS-IS instance is tagged as UNDERLAY for better recall.

(Corresponding) IS-IS spine switch S1 configuration

IS-IS global configuration

(config) #

```
feature isis
router isis UNDERLAY
 net 49.0001.0010.0100.1053.00
 is-type level-1
 set-overload-bit on-startup 60
```

IS-IS P2P interface configuration (spine switch S1)

(config) #

```
interface Ethernet 1/1
   description Link to VTEP V1
   ip address 209.165.201.2/31
   mtu 9192
   ip router isis UNDERLAY
```

IS-IS loopback interface configuration (spine switch S1)

(config) #

```
interface loopback 0
   ip address 10.1.1.53/32
   ip router isis UNDERLAY
.
.
```

To complete IS-IS topology configuration for the above image, configure the following:

- 3 more leaf switch V1's interfaces (or 3 more P2P links) to the remaining 3 spine switches.
- Repeat the procedure to connect P2P links between leaf switches V2, V3 and V4 and the spine switches.

IS-IS - IP unnumbered scenario

IS-IS configuration on leaf switch V1

IS-IS global configuration

(config)#

```
feature isis
router isis UNDERLAY
  net 49.0001.0010.0100.1074.00
  is-type level-1
  set-overload-bit on-startup 60
```

IS-IS interface configuration (leaf switch V1)

(config) #

```
interface Ethernet1/41
  description Link to Spine S1
  mtu 9192
  medium p2p
  ip unnumbered loopback0
  ip router isis UNDERLAY
```

IS-IS loopback interface configuration (leaf switch V1)

(config)

```
interface loopback0
  ip address 10.1.1.74/32
  ip router isis UNDERLAY
```

IS-IS configuration on the spine switch S1

IS-IS global configuration

(config)#

```
feature isis
router isis UNDERLAY
  net 49.0001.0010.0100.1053.00
  is-type level-1
  set-overload-bit on-startup 60
```

IS-IS interface configuration (spine switch S1)

(config)#

```
interface Ethernet1/41
  description Link to V1
  mtu 9192
  medium p2p
  ip unnumbered loopback0
  ip router isis UNDERLAY
```

IS-IS loopback interface configuration (spine switch S1)

(config)#

```
interface loopback0
  ip address 10.1.1.53/32
  ip router isis UNDERLAY
```

IS-IS Verification

Use the following commands for verifying IS-IS configuration on leaf switch V1:

```
Leaf-Switch-V1# show isis

ISIS process : UNDERLAY
 Instance number :  1
 UUID: 1090519320
 Process ID 20258
VRF: default
  System ID : 0010.0100.1074  IS-Type : L1
  SAP : 412  Queue Handle : 15
  Maximum LSP MTU: 1492
  Stateful HA enabled
  Graceful Restart enabled. State: Inactive
  Last graceful restart status : none
  Start-Mode Complete
  BFD IPv4 is globally disabled for ISIS process: UNDERLAY
  BFD IPv6 is globally disabled for ISIS process: UNDERLAY
  Topology-mode is base
  Metric-style : advertise(wide), accept(narrow, wide)
  Area address(es) :
    49.0001
 Process is up and running
  VRF ID: 1
  Stale routes during non-graceful controlled restart
  Interfaces supported by IS-IS :
    loopback0
    loopback1
    Ethernet1/41
  Topology : 0
  Address family IPv4 unicast :
    Number of interface : 2
    Distance : 115
  Address family IPv6 unicast :
    Number of interface : 0
    Distance : 115
  Topology : 2
  Address family IPv4 unicast :
    Number of interface : 0
    Distance : 115
  Address family IPv6 unicast :
    Number of interface : 0
```

```
    Distance : 115
   Level1
  No auth type and keychain
  Auth check set
  Level2
  No auth type and keychain
  Auth check set
  L1 Next SPF: Inactive
  L2 Next SPF: Inactive
```

```
Leaf-Switch-V1# show isis interface

IS-IS process: UNDERLAY VRF: default
loopback0, Interface status: protocol-up/link-up/admin-up IP address: 10.1.1.74, IP subnet:
 10.1.1.74/32
IPv6 routing is disabled Level1
No auth type and keychain Auth check set
Level2
No auth type and keychain Auth check set
Index: 0x0001, Local Circuit ID: 0x01, Circuit Type: L1 BFD IPv4 is locally disabled for
Interface loopback0 BFD IPv6 is locally disabled for Interface loopback0 MTR is disabled
Level Metric 1  1
2 1
Topologies enabled:
    L  MT  Metric  MetricCfg  Fwdng IPV4-MT  IPV4Cfg  IPV6-MT  IPV6Cfg
    1  0        1       no   UP    UP       yes      DN       no
    2  0        1       no   DN    DN       no       DN       no

loopback1, Interface status: protocol-up/link-up/admin-up
  IP address: 10.1.2.74, IP subnet: 10.1.2.74/32
  IPv6 routing is disabled
  Level1
    No auth type and keychain
    Auth check set
  Level2
    No auth type and keychain
    Auth check set
  Index: 0x0002, Local Circuit ID: 0x01, Circuit Type: L1
  BFD IPv4 is locally disabled for Interface loopback1
  BFD IPv6 is locally disabled for Interface loopback1
  MTR is disabled
  Passive level: level-2
  Level      Metric
  1               1
  2               1
  Topologies enabled:
    L  MT  Metric  MetricCfg  Fwdng IPV4-MT  IPV4Cfg  IPV6-MT  IPV6Cfg
    1  0        1       no   UP    UP       yes      DN       no
    2  0        1       no   DN    DN       no       DN       no

Ethernet1/41, Interface status: protocol-up/link-up/admin-up
  IP unnumbered interface (loopback0)
  IPv6 routing is disabled
    No auth type and keychain
    Auth check set
  Index: 0x0002, Local Circuit ID: 0x01, Circuit Type: L1
  BFD IPv4 is locally disabled for Interface Ethernet1/41
  BFD IPv6 is locally disabled for Interface Ethernet1/41
  MTR is disabled
  Extended Local Circuit ID: 0x1A028000, P2P Circuit ID: 0000.0000.0000.00
  Retx interval: 5, Retx throttle interval: 66 ms
  LSP interval: 33 ms, MTU: 9192
  P2P Adjs: 1, AdjsUp: 1, Priority 64
  Hello Interval: 10, Multi: 3, Next IIH: 00:00:01
  MT    Adjs   AdjsUp  Metric   CSNP  Next CSNP  Last LSP ID
```

```
  1          1        1       4     60  00:00:35   ffff.ffff.ffff.ff-ff
  2          0        0       4     60  Inactive   ffff.ffff.ffff.ff-ff
  Topologies enabled:
    L  MT  Metric  MetricCfg  Fwdng IPV4-MT  IPV4Cfg  IPV6-MT  IPV6Cfg
    1  0        4       no   UP    UP       yes      DN       no
    2  0        4       no   UP    DN       no       DN       no
```

```
Leaf-Switch-V1# show isis adjacency

IS-IS process: UNDERLAY VRF: default
IS-IS adjacency database:
Legend: '!': No AF level connectivity in given topology
System ID          SNPA    Level  State Hold Time  Interface
Spine-Switch-S1    N/A     1      UP    00:00:23   Ethernet1/41
```

For a detailed list of commands, refer to the Configuration and Command Reference guides.

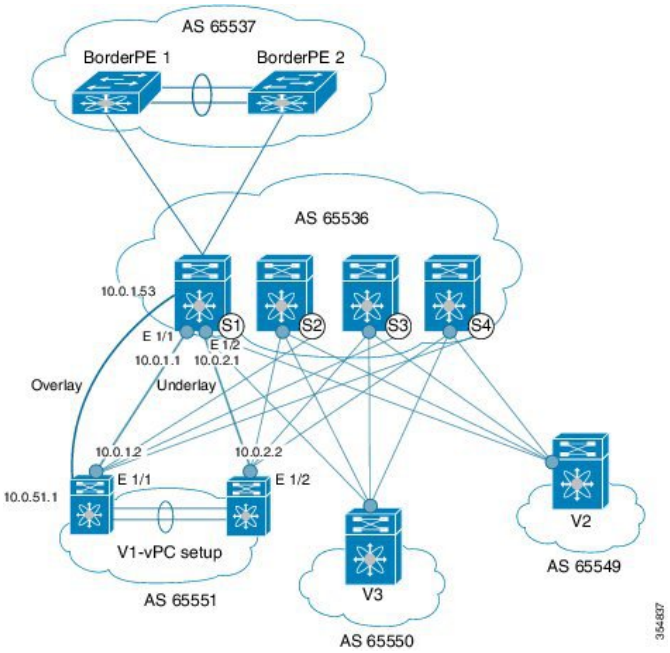
eBGP Underlay IP Network

Some customers would like to have the same protocol in the underlay and overlay in order to contain the number of protocols that need support in their network.

There are various ways to configure the eBGP based underlay. The configurations given in this section have been validated for function and convergence. The IP underlay based on eBGP can be built with these configurations detailed below. (For reference, see image below)

- The design below is following the multi AS model.
- eBGP underlay requires numbered interfaces between leaf and spine nodes. Numbered interfaces are used for the underlay BGP sessions as there is no other protocol to distribute peer reachability.
- The overlay sessions are configured on loopback addresses. This is to increase the resiliency in presence of link or node failures.
- BGP speakers on spine layer configure all leaf node eBGP neighbors individually. This is different from IBGP based peering which can be covered by dynamic BGP.
- Pointers for Multiple AS numbers in a fabric are given below:
  - All spine nodes configured as BGP speakers are in one AS.
  - All leaf nodes will have a unique AS number that is different than the BGP speakers in spine layer.
  - A pair of vPC leaf switch nodes, have the same AS number.
  - If a globally unique AS number is required to represent the fabric, then that can be configured on the border leaf or borderPE switches. All other nodes can use the private AS number range.
  - BGP Confederation has not been leveraged.

Figure 6: eBGP as underlay



eBGP configuration sample

Sample configurations for a spine switch and leaf switch are given below. The complete configuration is given for providing context, and the configurations added specifically for eBGP underlay are highlighted and further explained.

There is one BGP session per neighbor to set up the underlay. This is done within the global IPv4 address family. The session is used to distribute the loopback addresses for VTEP, Rendezvous Point (RP) and the eBGP peer address for the overlay eBGP session.

Spine switch S1 configuration—On the spine switch (S1 in this example), all leaf nodes are configured as eBGP neighbors.

(config) #

```
router bgp 65536
   router-id 10.1.1.53
   address-family ipv4 unicast
   redistribute direct route-map DIRECT-ROUTES-MAP
```

The **redistribute direct** command is used to advertise the loopback addresses for BGP and VTEP peering. It can be used to advertise any other direct routes in the global address space. The route map can filter the advertisement to include only eBGP peering and VTEP loopback addresses.

```
   maximum-paths 2
   address-family l2vpn evpn
     retain route-target all
```

Spine switch BGP speakers don't have any VRF configuration. Hence, the **retain route-target all** command is needed to retain the routes and send them to leaf switch VTEPs. The **maximum-paths** command is used for ECMP path in the underlay.

Underlay session towards leaf switch V1 (vPC set up) —As mentioned above, the underlay sessions are configured on the numbered interfaces between spine and leaf switch nodes.

(config) #

```
neighbor 10.0.1.2 remote-as 65551
   address-family ipv4 unicast
      disable-peer-as-check
      send-community both
```

The vPC pair of switches has the same AS number. The **disable-peer-as-check** command is added to allow route propagation between the vPC switches as they are configured with the same AS, for example, for route type 5 routes. If the vPC switches have different AS numbers, this command is not required.

Underlay session towards the border leaf switch—The underlay configurations towards leaf and border leaf switches are the same, barring the changes in IP address and AS values.

Overlay session on the spine switch S1 towards the leaf switch V1

(config) #

```
route-map UNCHANGED permit 10
   set ip next-hop unchanged
```



Note The route-map UNCHANGED is user defined whereas the keyword **unchanged** is an option within the **set ip next-hop** command. In eBGP, the next hop is changed to self when sending a route from one eBGP neighbor to another. The route map UNCHANGED is added to make sure that, for overlay routes, the originating leaf switch is set as next hop and not the spine switch. This ensures that VTEPs are next hops, and not spine switch nodes. The **unchanged** keyword ensures that the next-hop attribute in the BGP update to the eBGP peer is unmodified.

The overlay sessions are configured on loopback addresses.

(config) #

```
neighbor 10.0.51.1 remote-as 65551
   update-source loopback0
   ebgp-multihop 2
   address-family l2vpn evpn
     rewrite-evpn-rt-asn
       disable-peer-as-check
       send-community both
       route-map UNCHANGED out
```

The spine switch configuration concludes here. The *Route Target auto* feature configuration is given below for reference purposes:

(config) #

```
vrf context coke
    vni 50000
    rd auto
    address-family ipv4 unicast
      route-target both auto
      route-target both auto evpn
    address-family ipv6 unicast
      route-target both auto
      route-target both auto evpn
```

The **rewrite-evpn-rt-asn** command is required if the *Route Target auto* feature is being used to configure EVPN RTs.

Route target auto is derived from the Local AS number configured on the switch and the Layer-3 VNID of the VRF i.e. Local AS:VNID. In Multi-AS topology, as illustrated in this guide, each leaf node is represented as a different local AS, and the route target generated for the same VRF will be different on every switch. The command **rewrite-evpn-rt-asn** replaces the ASN portion of the route target in the BGP update message with the local AS number. For example, if VTEP V1 has a Local AS 65551, VTEP V2 has a Local AS 65549, and spine switch S1 has a Local AS 65536, then the route targets for V1, V2 and S1 are as follows:

- V1—65551:50000
- V2—65549:50000
- S1—65536:50000

In this scenario, V2 advertises the route with RT 65549:50000, the spine switch S1 replaces it with RT 65536:50000, and finally when V1 gets the update, it replaces the route target in the update with 65551:50000. This matches the locally configured RT on V1. This command requires that it be configured on all BGP speakers in the fabric.

If the *Route Target auto* feature is not being used, i.e., matching RTs are required to be manually configured on all switches, then this command is not necessary.

Leaf switch VTEP V1 configuration—In the sample configuration below, VTEP V1's interfaces are designated as BGP neighbors. All leaf switch VTEPs including border leaf switch nodes have the following configurations towards spine switch neighbor nodes:

(config) #

```
router bgp 65551
  router-id 10.1.1.54
  address-family ipv4 unicast
     maximum-paths 2
     address-family l2vpn evpn
```

The **maximum-paths** command is used for ECMP path in the underlay.

Underlay session on leaf switch VTEP V1 towards spine switch S1

(config) #

```
neighbor 10.0.1.1 remote-as 65536
   address-family ipv4 unicast
     allowas-in
```

```
     send-community both
```

The **allowas-in** command is needed if leaf switch nodes have the same AS. In particular, the Cisco validated topology had a vPC pair of switches share an AS number.

Overlay session towards spine switch S1

(config) #

```
neighbor 10.1.1.53 remote-as 65536
   update-source loopback0
   ebgp-multihop 2
   address-family l2vpn evpn
   rewrite-evpn-rt-asn
     allowas-in
     send-community both
```

The **ebgp-multihop 2** command is needed as the peering for the overlay is on the loopback address. NX-OS considers that as multi hop even if the neighbor is one hop away.

vPC backup session

(config) #

```
route-map SET-PEER-AS-NEXTHOP permit 10
  set ip next-hop peer-address

neighbor 192.168.0.1 remote-as 65551
    update-source Vlan3801
    address-family ipv4 unicast
      send-community both
      route-map SET-PEER-AS-NEXTHOP out
```



Note This session is configured on the backup SVI between the vPC leaf switch nodes.

To complete configurations for the above image, configure the following:

- *V1 as a BGP neighbor to other spine switches.*
- *Repeat the procedure for other leaf switches.*

BGP Verification

Use the following commands for verifying BGP configuration:

```
show bgp all
show bgp ipv4 unicast neighbors
show ip route bgp
```

For a detailed list of commands, refer to the Configuration and Command Reference guides.

Multicast Routing in the VXLAN Underlay

The VXLAN EVPN Programmable Fabric supports multicast routing for transporting BUM (broadcast, unknown unicast and multicast) traffic.

Refer the table below to know the multicast protocol(s) your Cisco Nexus switches support:

Cisco Nexus Series Switch(es) Combination	Multicast Routing Option
Cisco Nexus 7000/7700 Series switches with Cisco Nexus 9000 Series switches	PIM ASM (Sparse Mode)
Cisco Nexus 9000 Series	PIM ASM (Sparse Mode) *or* PIM BiDir **Note** PIM BiDir is supported on Cisco Nexus 9300-FX/FX2 platform switches. Beginning with Cisco NX-OS Release 10.4(3)F, PIM Bidir is also supported on Cisco Nexus 9300-FX3/GX/GX2/H2R/H1 switches, 9500 switches with 9700-GX line cards. Beginning with Cisco NX-OS Release 10.3(5)M, PIM Bidir is also supported on Cisco Nexus 9300-FX3/GX/GX2/H2R/H1 switches, 9500 switches with 9700-GX line cards. Beginning with Cisco NX-OS Release 10.5(2)F, PIM Bidir is also supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.

You can transport BUM traffic without multicast, through *ingress replication* . Ingress replication is currently available on Cisco Nexus 9000 Series switches.

PIM ASM and PIM Bidir Underlay IP Network

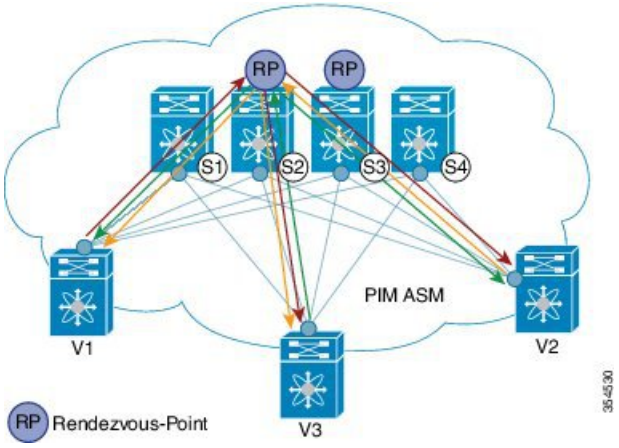
Some multicast topology design pointers are given below:

- Use spine/aggregation switches as Rendezvous-Point locations.
- Reserve a range of multicast groups (destination groups/DGroups) to service the overlay and optimize for diverse VNIs.
- In a spine-leaf topology with a lean spine,
  - Use multiple Rendezvous-Points across multiple spine switches.
  - Use redundant Rendezvous-Points.
  - Map different VNIs to different multicast groups, which are mapped to different Rendezvous-Points for load balancing.

Important The following configuration samples are from an IP underlay perspective and are not comprehensive. Functions such as PIM authentication, BFD for PIM, etc, are not shown here. Refer to the respective Cisco Nexus Series switch multicast configuration guide for complete information.

PIM Sparse-Mode (Any-Source Multicast [ASM])

Figure 7: PIM ASM as the IP multicast routing protocol



PIM ASM is supported on the Cisco Nexus 9000 series as the underlay multicast protocol.

In the above image, the leaf switches (V1, V2, and V3 having VTEP configuration) are at the bottom of the image. They are connected to the 4 spine switches (S1, S2, S3, and S4) that are depicted at the top of the image.

Two multicast Rendezvous-Points (S2 and S3) are configured. The second Rendezvous-Point is added for load sharing and redundancy purposes. *Anycast RP is represented in the PIM ASM topology image*. Anycast RP ensures redundancy and load sharing between the two Rendezvous-Points. To use Anycast RP, multiple spines serving as RPs will share the same IP address (the Anycast RP address). Meanwhile, each RP has its unique IP address added in the RP set for RPs to sync information with respect to sources between all spines which act as RPs.

The shared multicast tree is unidirectional, and uses the Rendezvous-Point for forwarding packets.

PIM ASM at a glance - 1 source tree per multicast group per leaf switch.

Programmable Fabric specific pointers are:

- All VTEPs that serve a VNI join a shared multicast tree. VTEPs V1, V2, and V3 have hosts attached from a single tenant (say x) and these VTEPs form a separate multicast (source, group) tree.
- A VTEP (say V1) might have hosts belonging to other tenants too. Each tenant may have different multicast groups associated with. A source tree is created for each tenant residing on the VTEP, if the tenants do not share a multicast group.

PIM ASM Configuration

Note For ease of use, the configuration mode from which you need to start configuring a task is mentioned at the beginning of each configuration.

Configuration tasks and corresponding show command output are displayed for a part of the topology in the image. For example, if the sample configuration is shown for a leaf switch and connected spine switch, the show command output for the configuration only displays corresponding configuration.

Leaf switch V1 Configuration — Configure RP reachability on the leaf switch.

PIM Anycast Rendezvous-Point association on leaf switch V1

(config) #

```
feature pim
ip pim rp-address 198.51.100.220 group-list 224.1.1.1
```

198.51.100.220 is the Anycast Rendezvous-Point IP address.

Loopback interface PIM configuration on leaf switch V1

(config) #

```
interface loopback 0
  ip address 209.165.201.20/32
  ip pim sparse-mode
```

Point-2-Point (P2P) interface PIM configuration for leaf switch V1 to spine switch S2 connectivity

(config) #

```
interface Ethernet 1/1
  no switchport
  ip address 209.165.201.14/31
  mtu 9216
  ip pim sparse-mode
.
.
```

Repeat the above configuration for a P2P link between V1 and the spine switch (S3) acting as the redundant Anycast Rendezvous-Point.

The VTEP also needs to be connected with spine switches (S1 and S4) that are not rendezvous points. A sample configuration is given below:

Point-2-Point (P2P) interface configuration for leaf switch V1 to non-rendezvous point spine switch (S1) connectivity

(config) #

```
interface Ethernet 2/2
  no switchport
  ip address 209.165.201.10/31
  mtu 9216
  ip pim sparse-mode
```

Repeat the above configuration for all P2P links between V1 and non- rendezvous point spine switches.

Repeat the complete procedure given above to configure all other leaf switches.

Rendezvous Point Configuration on the spine switches

PIM configuration on spine switch S2

(config) #

```
feature pim
```

Loopback Interface Configuration (RP)

(config) #

```
interface loopback 0
 ip address 10.10.100.100/32
 ip pim sparse-mode
```

Loopback interface configuration (Anycast RP)

(config) #

```
interface loopback 1
  ip address 198.51.100.220/32
  ip pim sparse-mode
```

Anycast-RP configuration on spine switch S2

Configure a spine switch as a Rendezvous Point and associate it with the loopback IP addresses of switches S2 and S3 for redundancy.

(config) #

```
feature pim
ip pim rp-address 198.51.100.220 group-list 224.1.1.1
ip pim anycast-rp 198.51.100.220 10.10.100.100
ip pim anycast-rp 198.51.100.220 10.10.20.100
.
.
```



Note The above configurations should also be implemented on the other spine switch (S3) performing the role of RP.

Non-RP Spine Switch Configuration

You also need to configure PIM ASM on spine switches that are not designated as rendezvous points, namely S1 and S4.

Earlier, leaf switch (VTEP) V1 has been configured for a P2P link to a non RP spine switch. A sample configuration on the non RP spine switch is given below.

PIM ASM global configuration on spine switch S1 (non RP)

(config) #

```
feature pim
ip pim rp-address 198.51.100.220 group-list 224.1.1.1
```

Loopback interface configuration (non RP)

(config) #

```
interface loopback 0
  ip address 10.10.100.103/32
  ip pim sparse-mode
```

Point-2-Point (P2P) interface configuration for spine switch S1 to leaf switch V1 connectivity

(config) #

```
interface Ethernet 2/2
  no switchport
  ip address 209.165.201.15/31
  mtu 9216
  ip pim sparse-mode
.
.
```

Repeat the above configuration for all P2P links between the non- rendezvous point spine switches and other leaf switches (VTEPs).

PIM ASM Verification

Use the following commands for verifying PIM ASM configuration:

```
Leaf-Switch-V1# show ip mroute 224.1.1.1

IP Multicast Routing Table for VRF "default"

(*, 224.1.1.1/32), uptime: 02:21:20, nve ip pim
  Incoming interface: Ethernet1/1, RPF nbr: 10.10.100.100
  Outgoing interface list: (count: 1)
    nve1, uptime: 02:21:20, nve

(10.1.1.54/32, 224.1.1.1/32), uptime: 00:08:33, ip mrib pim
  Incoming interface: Ethernet1/2, RPF nbr: 209.165.201.12
  Outgoing interface list: (count: 1)
    nve1, uptime: 00:08:33, mrib

(10.1.1.74/32, 224.1.1.1/32), uptime: 02:21:20, nve mrib ip pim
```

```
  Incoming interface: loopback0, RPF nbr: 10.1.1.74
  Outgoing interface list: (count: 1)
    Ethernet1/6, uptime: 00:29:19, pim

Leaf-Switch-V1# show ip pim rp

PIM RP Status Information for VRF "default"
BSR disabled
Auto-RP disabled
BSR RP Candidate policy: None
BSR RP policy: None
Auto-RP Announce policy: None
Auto-RP Discovery policy: None

RP: 198.51.100.220, (0), uptime: 03:17:43, expires: never,
  priority: 0, RP-source: (local), group ranges:
      224.0.0.0/9

Leaf-Switch-V1# show ip pim interface

PIM Interface Status for VRF "default"
Ethernet1/1, Interface status: protocol-up/link-up/admin-up
  IP address: 209.165.201.14, IP subnet: 209.165.201.14/31
  PIM DR: 209.165.201.12, DR's priority: 1
  PIM neighbor count: 1
  PIM hello interval: 30 secs, next hello sent in: 00:00:11
  PIM neighbor holdtime: 105 secs
  PIM configured DR priority: 1
  PIM configured DR delay: 3 secs
  PIM border interface: no
  PIM GenID sent in Hellos: 0x33d53dc1
  PIM Hello MD5-AH Authentication: disabled
  PIM Neighbor policy: none configured
  PIM Join-Prune inbound policy: none configured
  PIM Join-Prune outbound policy: none configured
  PIM Join-Prune interval: 1 minutes
  PIM Join-Prune next sending: 1 minutes
  PIM BFD enabled: no
  PIM passive interface: no
  PIM VPC SVI: no
  PIM Auto Enabled: no
  PIM Interface Statistics, last reset: never
    General (sent/received):
      Hellos: 423/425 (early: 0), JPs: 37/32, Asserts: 0/0
      Grafts: 0/0, Graft-Acks: 0/0
      DF-Offers: 4/6, DF-Winners: 0/197, DF-Backoffs: 0/0, DF-Passes: 0/0
    Errors:
      Checksum errors: 0, Invalid packet types/DF subtypes: 0/0
      Authentication failed: 0
      Packet length errors: 0, Bad version packets: 0, Packets from self: 0
      Packets from non-neighbors: 0
          Packets received on passiveinterface: 0
      JPs received on RPF-interface: 0
      (*,G) Joins received with no/wrong RP: 0
      (*,G)/(S,G) JPs received for SSM/Bidir groups: 0/0
      JPs filtered by inbound policy: 0
      JPs filtered by outbound policy: 0
loopback0, Interface status: protocol-up/link-up/admin-up
  IP address: 209.165.201.20, IP subnet: 209.165.201.20/32
  PIM DR: 209.165.201.20, DR's priority: 1
  PIM neighbor count: 0
  PIM hello interval: 30 secs, next hello sent in: 00:00:07
  PIM neighbor holdtime: 105 secs
  PIM configured DR priority: 1
  PIM configured DR delay: 3 secs
```

```
  PIM border interface: no
  PIM GenID sent in Hellos: 0x1be2bd41
  PIM Hello MD5-AH Authentication: disabled
  PIM Neighbor policy: none configured
  PIM Join-Prune inbound policy: none configured
  PIM Join-Prune outbound policy: none configured
  PIM Join-Prune interval: 1 minutes
  PIM Join-Prune next sending: 1 minutes
  PIM BFD enabled: no
  PIM passive interface: no
  PIM VPC SVI: no
  PIM Auto Enabled: no
  PIM Interface Statistics, last reset: never
    General (sent/received):
      Hellos: 419/0 (early: 0), JPs: 2/0, Asserts: 0/0
      Grafts: 0/0, Graft-Acks: 0/0
      DF-Offers: 3/0, DF-Winners: 0/0, DF-Backoffs: 0/0, DF-Passes: 0/0
    Errors:
      Checksum errors: 0, Invalid packet types/DF subtypes: 0/0
      Authentication failed: 0
      Packet length errors: 0, Bad version packets: 0, Packets from self: 0
     Packets from non-neighbors: 0
          Packets received on passiveinterface: 0
      JPs received on RPF-interface: 0
      (*,G) Joins received with no/wrong RP: 0/0
      (*,G)/(S,G) JPs received for SSM/Bidir groups: 0/0
      JPs filtered by inbound policy: 0
      JPs filtered by outbound policy: 0
```

```
Leaf-Switch-V1# show ip pim neighbor

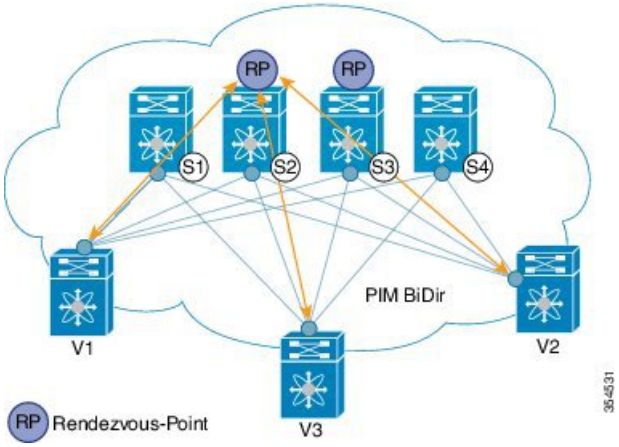
PIM Neighbor Status for VRF "default"

Neighbor        Interface      Uptime    Expires    DR          Bidir-        BFD
                                                    Priority    Capable       State
10.10.100.100   Ethernet1/1    1w1d      00:01:33   1           yes           n/a
```

For a detailed list of commands, refer to the Configuration and Command Reference guides.

PIM Bidirectional (BiDir)

Figure 8: PIM BiDir as the IP multicast routing protocol



VXLAN BiDir underlay is supported on Cisco Nexus 9300-FX/FX2/FX3 platform switches.

In the above image, the leaf switches (V1, V2, and V3) are at the bottom of the image. They are connected to the 4 spine switches (S1, S2, S3, and S4) that are depicted at the top of the image. The two PIM Rendezvous-Points using phantom RP mechanism are used for load sharing and redundancy purposes.

Note Load sharing happens only via different multicast groups, for the respective, different VNI.

With bidirectional PIM, one bidirectional, shared tree rooted at the RP is built for each multicast group. Source specific state are not maintained within the fabric which provides a more scalable solution.

Programmable Fabric specific pointers are:

- The 3 VTEPs share the same VNI and multicast group mapping to form a single multicast group tree.

PIM BiDir at a glance — *One shared tree per multicast group*.

PIM BiDir Configuration

The following is a configuration example of having two spine switches S2 and S3 serving as RPs using phantom RP for redundancy and loadsharing. Here S2 is the primary RP for group-list 227.2.2.0/26 and secondary for group-list 227.2.2.64/26. S3 is the primary RP for group-list 227.2.2.64/26 and secondary RP for group-list 227.2.2.0/26.

Note Phantom RP is used in a PIM BiDir environment where RP redundancy is designed using loopback networks with different mask lengths in the primary and secondary routers. These loopback interfaces are in the same subnet as the RP address, but with different IP addresses from the RP address. (Since the IP address advertised as RP address is not defined on any routers, the term phantom is used). The subnet of the loopback is advertised in the Interior Gateway Protocol (IGP). To maintain RP reachability, it is only necessary to ensure that a route to the RP exists.

Unicast routing longest match algorithms are used to pick the primary over the secondary router.

The primary router announces a longest match route (say, a /30 route for the RP address) and is preferred over the less specific route announced by the secondary router (a /29 route for the same RP address). The primary router advertises the /30 route of the RP, while the secondary router advertises the /29 route. The latter is only chosen when the primary router goes offline. We will be able to switch from the primary to the secondary RP at the speed of convergence of the routing protocol.

For ease of use, the configuration mode from which you need to start configuring a task is mentioned at the beginning of each configuration.

Configuration tasks and corresponding show command output are displayed for a part of the topology in the image. For example, if the sample configuration is shown for a leaf switch and connected spine switch, the show command output for the configuration only displays corresponding configuration.

Leaf switch V1 configuration

Phantom Rendezvous-Point association on leaf switch V1

(config) #

```
feature pim
```

```
ip pim rp-address 10.254.254.1 group-list 227.2.2.0/26 bidir
ip pim rp-address 10.254.254.65 group-list 227.2.2.64/26 bidir
```

Loopback interface PIM configuration on leaf switch V1

(config) #

```
interface loopback 0
  ip address 10.1.1.54/32
  ip pim sparse-mode
```

IP unnumbered P2P interface configuration on leaf switch V1

(config) #

```
interface Ethernet 1/1
  no switchport
  mtu 9192
  medium p2p
  ip unnumbered loopback 0
  ip pim sparse-mode

interface Ethernet 2/2
  no switchport
  mtu 9192
  medium p2p
  ip unnumbered loopback 0
  ip pim sparse-mode
```

Rendezvous Point configuration (on the two spine switches S2 and S3 acting as RPs)

Using phantom RP on spine switch S2

(config) #

```
feature pim
ip pim rp-address 10.254.254.1 group-list 227.2.2.0/26 bidir
ip pim rp-address 10.254.254.65 group-list 227.2.2.64/26 bidir
```

Loopback interface PIM configuration (RP) on spine switch S2/RP1

(config) #

```
interface loopback 0
  ip address 10.1.1.53/32
  ip pim sparse-mode
```

IP unnumbered P2P interface configuration on spine switch S2/RP1 to leaf switch V1

(config) #

```
interface Ethernet 1/1
  no switchport
```

```
  mtu 9192
  medium p2p
  ip unnumbered loopback 0
  ip pim sparse-mode
```

Loopback interface PIM configuration (for phantom RP) on spine switch S2/RP1

(config) #

```
interface loopback 1
  ip address 10.254.254.2/30
  ip pim sparse-mode
```

(config) #

```
interface loopback 2
  ip address 10.254.254.66/29
  ip pim sparse-mode
```

Using phantom RP on spine switch S3

(config) #

```
feature pim
ip pim rp-address 10.254.254.1 group-list 227.2.2.0/26 bidir
ip pim rp-address 10.254.254.65 group-list 227.2.2.64/26 bidir
```

Loopback interface PIM configuration (RP) on spine switch S3/RP2

(config) #

```
interface loopback 0
  ip address 10.10.50.100/32
  ip pim sparse-mode
```

IP unnumbered P2P interface configuration on spine switch S3/RP2 to leaf switch V1

(config) #

```
interface Ethernet 2/2
  no switchport
  mtu 9192
  medium p2p
  ip unnumbered loopback 0
  ip pim sparse-mode
```

Loopback interface PIM configuration (for phantom RP) on spine switch S3/RP2

(config) #

```
interface loopback 1
  ip address 10.254.254.66/30
  ip pim sparse-mode
```

```
interface loopback 2
  ip address 10.254.254.2/29
  ip pim sparse-mode
```

PIM BiDir Verification

Use the following commands for verifying PIM BiDir configuration:

```
Leaf-Switch-V1# show ip mroute

IP Multicast Routing Table for VRF "default"

(*, 227.2.2.0/26), bidir, uptime: 4d08h, pim ip
  Incoming interface: Ethernet1/1, RPF nbr: 10.1.1.53
  Outgoing interface list: (count: 1)
    Ethernet1/1, uptime: 4d08h, pim, (RPF)

(*, 227.2.2.0/32), bidir, uptime: 4d08h, nve ip pim
  Incoming interface: Ethernet1/1, RPF nbr: 10.1.1.53
  Outgoing interface list: (count: 2)
    Ethernet1/1, uptime: 4d08h, pim, (RPF)
    nve1, uptime: 4d08h, nve

(*, 227.2.2.64/26), bidir, uptime: 4d08h, pim ip
  Incoming interface: Ethernet1/5, RPF nbr: 10.10.50.100/32
  Outgoing interface list: (count: 1)
    Ethernet1/5, uptime: 4d08h, pim, (RPF)

(*, 232.0.0.0/8), uptime: 4d08h, pim ip
  Incoming interface: Null, RPF nbr: 0.0.0.0
  Outgoing interface list: (count: 0)


Leaf-Switch-V1# show ip pim rp

PIM RP Status Information for VRF "default"
BSR disabled
Auto-RP disabled
BSR RP Candidate policy: None
BSR RP policy: None
Auto-RP Announce policy: None
Auto-RP Discovery policy: None

RP: 10.254.254.1, (1),
 uptime: 4d08h   priority: 0,
 RP-source: (local),
 group ranges:
 227.2.2.0/26  (bidir)
RP: 10.254.254.65, (2),
 uptime: 4d08h   priority: 0,
 RP-source: (local),
 group ranges:
 227.2.2.64/26  (bidir)


Leaf-Switch-V1# show ip pim interface

PIM Interface Status for VRF "default"
loopback0, Interface status: protocol-up/link-up/admin-up
  IP address: 10.1.1.54, IP subnet: 10.1.1.54/32
  PIM DR: 10.1.1.54, DR's priority: 1
  PIM neighbor count: 0
  PIM hello interval: 30 secs, next hello sent in: 00:00:23
  PIM neighbor holdtime: 105 secs
```

```
  PIM configured DR priority: 1
  PIM configured DR delay: 3 secs
  PIM border interface: no
  PIM GenID sent in Hellos: 0x12650908
  PIM Hello MD5-AH Authentication: disabled
  PIM Neighbor policy: none configured
  PIM Join-Prune inbound policy: none configured
  PIM Join-Prune outbound policy: none configured
  PIM Join-Prune interval: 1 minutes
  PIM Join-Prune next sending: 1 minutes
  PIM BFD enabled: no
  PIM passive interface: no
  PIM VPC SVI: no
  PIM Auto Enabled: no
  PIM Interface Statistics, last reset: never
    General (sent/received):
      Hellos: 13158/0 (early: 0), JPs: 0/0, Asserts: 0/0
      Grafts: 0/0, Graft-Acks: 0/0
      DF-Offers: 0/0, DF-Winners: 0/0, DF-Backoffs: 0/0, DF-Passes: 0/0
    Errors:
      Checksum errors: 0, Invalid packet types/DF subtypes: 0/0
      Authentication failed: 0
      Packet length errors: 0, Bad version packets: 0, Packets from self: 0
      Packets from non-neighbors: 0
          Packets received on passiveinterface: 0
      JPs received on RPF-interface: 0
      (*,G) Joins received with no/wrong RP: 0/0
      (*,G)/(S,G) JPs received for SSM/Bidir groups: 0/0
      JPs filtered by inbound policy: 0
      JPs filtered by outbound policy: 0

Ethernet1/1, Interface status: protocol-up/link-up/admin-up
  IP unnumbered interface (loopback0)
  PIM DR: 10.1.1.54, DR's priority: 1
  PIM neighbor count: 1
  PIM hello interval: 30 secs, next hello sent in: 00:00:04
  PIM neighbor holdtime: 105 secs
  PIM configured DR priority: 1
  PIM configured DR delay: 3 secs
  PIM border interface: no
  PIM GenID sent in Hellos: 0x2534269b
  PIM Hello MD5-AH Authentication: disabled
  PIM Neighbor policy: none configured
  PIM Join-Prune inbound policy: none configured
  PIM Join-Prune outbound policy: none configured
  PIM Join-Prune interval: 1 minutes
  PIM Join-Prune next sending: 1 minutes
  PIM BFD enabled: no
  PIM passive interface: no
  PIM VPC SVI: no
  PIM Auto Enabled: no
  PIM Interface Statistics, last reset: never
    General (sent/received):
      Hellos: 13152/13162 (early: 0), JPs: 2/0, Asserts: 0/0
      Grafts: 0/0, Graft-Acks: 0/0
      DF-Offers: 9/5, DF-Winners: 6249/6254, DF-Backoffs: 0/1, DF-Passes: 0/1
    Errors:
      Checksum errors: 0, Invalid packet types/DF subtypes: 0/0
      Authentication failed: 0
      Packet length errors: 0, Bad version packets: 0, Packets from self: 0
      Packets from non-neighbors: 0
          Packets received on passiveinterface: 0
      JPs received on RPF-interface: 0
      (*,G) Joins received with no/wrong RP: 0/0
```

```
      (*,G)/(S,G) JPs received for SSM/Bidir groups: 0/0
      JPs filtered by inbound policy: 0
      JPs filtered by outbound policy: 0

Leaf-Switch-V1# show ip pim neighbor

PIM Neighbor Status for VRF "default"

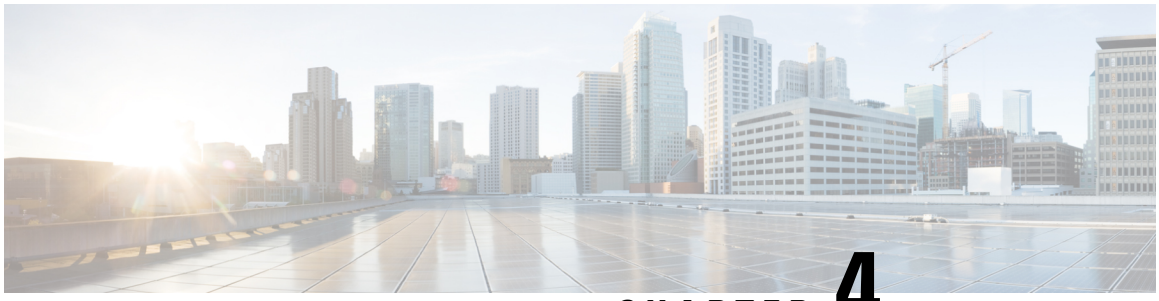
Neighbor     Interface      Uptime    Expires      DR          Bidir-     BFD
                                                  Priority    Capable    State
10.1.1.53    Ethernet1/1    1w1d      00:01:33     1             yes      n/a
10.10.50.100 Ethernet2/2    1w1d      00:01:33     1             yes      n/a
```

For a detailed list of commands, refer to the Configuration and Command Reference guides.

Underlay Deployment Without Multicast (Ingress Replication)

Ingress replication is supported on Cisco Nexus 9000 Series switches.

Beginning in NX-OS release 9.3(3), Ingress replication is supported on Cisco Nexus 9300-GX switches.



CHAPTER 4

Configuring VXLAN

This chapter contains these sections:

- Guidelines and Limitations for VXLAN, on page 43
- Considerations for VXLAN Deployment, on page 58
- vPC Considerations for VXLAN Deployment, on page 61
- Network Considerations for VXLAN Deployments, on page 65
- Considerations for the Transport Network, on page 66
- Considerations for Tunneling VXLAN, on page 67
- Configuring VXLAN, on page 69
- VXLAN and IP-in-IP Tunneling, on page 80
- Configuring VXLAN Static Tunnels, on page 83

Guidelines and Limitations for VXLAN

VXLAN has the following guidelines and limitations:

Switch or port restrictions

- FEX ports do not support IGMP snooping on VXLAN VLANs.
- The VXLAN UDP port number is used for VXLAN encapsulation. For Cisco Nexus NX-OS, the UDP port number is 4789. It complies with IETF standards and is not configurable.
- Cisco Nexus 9300 Series switches with 100G uplinks only support VXLAN switching/bridging.

  Cisco Nexus 9300-FX, and Cisco Nexus 9300-FX2 platform switches do not have this restriction.

Note For VXLAN routing support, a 40G uplink module is required.

- Beginning with Cisco NX-OS Release 10.2(3)F, VXLAN can coexist with the GRE tunnel feature or the MPLS (static or segment-routing) feature.
- Native VLANs are supported as transit traffic over a VXLAN fabric on Cisco Nexus 9300-FX/FX2/FX3/GX/GX2/H2R/H1 Series switches and 9800 Series switches.
- A FEX HIF (FEX host interface port) is supported for a VLAN that is extended with VXLAN.

- Bind NVE to a loopback address that is separate from other loopback addresses that are required by Layer 3 protocols. A best practice is to use a dedicated loopback address for VXLAN. This best practice should be applied not only for the vPC VXLAN deployment, but for all VXLAN deployments.
- Tenant VRF (VRF with VNI on it) cannot be used on an SVI that has no VNI binding into it (underlay infra VRF).
- For traceroute through a VXLAN fabric when using L3VNI, the following scenario is the expected behavior:

  If L3VNI is associated with a VRF and an SVI, the associated SVI does not have an L3 address that is configured but instead has the "ip forward" configuration command. Due to this interface setup it cannot respond back to the traceroute with its own SVI address. Instead, when a traceroute involving the L3VNI is run through the fabric, the IP address reported will be the lowest IP address of an SVI that belongs to the corresponding tenant VRF.
- In an ingress replication vPC setup, Layer 3 connectivity is needed between vPC peer devices.

VXLAN configuration restrictions

- **show** commands with the **internal** keyword are not supported.
- The **lacp vpc-convergence** command can be configured in VXLAN and non-VXLAN environments that have vPC port channels to hosts that support LACP.
- For scale environments, the VLAN IDs related to the VRF and Layer-3 VNI (L3VNI) must be reserved with the **system vlan nve-overlay id** command.
- The **load-share** keyword has been added to the Configuring a Route Policy procedure for the PBR over VXLAN feature.

  For information regarding the **load-share** keyword usage for PBR with VXLAN, see the Guidelines and Limitations for Policy-Based Routing section of the Cisco Nexus 9000 Series NX_OS Unicast Routing Configuration Guide, Release 9.x.
- The **lacp vpc-convergence** command is added for better convergence of Layer 2 EVPN VXLAN:

```
interface port-channel10
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 1001-1200
  spanning-tree port type edge trunk
  spanning-tree bpdufilter enable
  lacp vpc-convergence
  vpc 10

interface Ethernet1/34  <- The port-channel member-port is configured with LACP-active
 mode (for example, no changes are done at the member-port level.)
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 1001-1200
  channel-group 10 mode active
  no shutdown
```

- The VXLAN network identifier (VNID) 16777215 is reserved and should not be configured explicitly.
- To refresh the frozen duplicate host during fabric forwarding, use only **fabric forwarding dup-host-recovery-timer** command and do not use **fabric forwarding dup-host-unfreeze-timer** command, as it is deprecated.

ISSU restrictions

- VXLAN supports In-Service Software Upgrades (ISSUs). However, VXLAN ISSU is not supported for Cisco Nexus 9300-GX platform switches.
- To remove configurations from an NVE interface, we recommend manually removing each configuration rather than using the **default interface nve** command.
- Rollback is not supported on VXLAN VLANs that are configured with the port VLAN mapping feature.

Feature support and restrictions

- **ACL**
  - ACL Options for VXLAN Traffic on Cisco Nexus 9348GC-FXP Switches.

ACL Direction	ACL Type	VTEP Type	Port Type	Flow Direction	Traffic Type	Supported
Ingress	PACL	Ingress VTEP	L2 port	Access to Network [GROUP:encap direction]	Native L2 traffic [GROUP:inner]	YES
	VACL	Ingress VTEP	VLAN	Access to Network [GROUP:encap direction]	Native L2 traffic [GROUP:inner]	YES
Ingress	RACL	Ingress VTEP	Tenant L3 SVI	Access to Network [GROUP:encap direction]	Native L3 traffic [GROUP:inner]	YES
Egress	RACL	Ingress VTEP	Uplink L3/L3-PO/SVI	Access to Network [GROUP:encap direction]	VXLAN encap [GROUP:outer]	NO
Ingress	RACL	Egress VTEP	Uplink L3/L3-PO/SVI	Network to Access [GROUP:decap direction]	VXLAN encap [GROUP:outer]	NO
Egress	PACL	Egress VTEP	L2 port	Network to Access [GROUP:decap direction]	Native L2 traffic [GROUP:inner]	NO
	VACL	Egress VTEP	VLAN	Network to Access [GROUP:decap direction]	Native L2 traffic [GROUP:inner]	NO

ACL Direction	ACL Type	VTEP Type	Port Type	Flow Direction	Traffic Type	Supported
Egress	RACL	Egress VTEP	Tenant L3 SVI	Network to Access [GROUP:decap direction]	Post-decap L3 traffic [GROUP:inner]	YES

- Support added for MultiAuth Change of Authorization (CoA). For more information, see the Cisco Nexus 9000 Series NX-OS Security Configuration Guide, Release 9.3(x).

- **Multicast**
  - Beginning with Cisco NX-OS Release 10.3(1)F, the Non-blocking Multicast (NBM) feature and VXLAN can co-exist on the same box but in two different VRFs.



Note Make sure that the NBM is not enabled on the default VRF where underlay runs.

  - NLB in the unicast, multicast, and IGMP multicast modes is not supported on Cisco Nexus 9000 switch VXLAN VTEPs. The work-around is to move the NLB cluster behind the intermediary device (which supports NLB in the respective mode) and inject the cluster IP address as an external prefix into the VXLAN fabric.
  - On Cisco Nexus 9500 Series switches, if **feature nv overlay** is enabled, ensure that the interface NVE is configured and in UP state. Otherwise, multicast traffic may be silently dropped in Fabric Modules when it needs to be forwarded out of sub-interfaces.
  - If multiple VTEPs use the same multicast group address for underlay multicast but have different VNIs, the VTEPs should have at least one VNI in common. Doing so ensures that NVE peer discovery occurs and underlay multicast traffic is forwarded correctly. For example, leafs L1 and L4 could have VNI 10 and leafs L2 and L3 could have VNI 20, and both VNIs could share the same group address. When leaf L1 sends traffic to leaf L4, the traffic could pass through leaf L2 or L3. Because NVE peer L1 is not learned on leaf L2 or L3, the traffic is dropped. Therefore, VTEPs that share a group address need to have at least one VNI in common so that peer learning occurs and traffic is not dropped. This requirement applies to VXLAN bud-node topologies.

- **PIM BiDir**
  - PIM BiDir for VXLAN underlay with and without vPC is supported.

    The following features are not supported when PIM BiDir for VXLAN underlay is configured:
    - Flood and Learn VXLAN
    - Tenant Routed Multicast (TRM)
    - VXLAN EVPN Multi-Site
    - VXLAN EVPN Multihoming
    - vPC attached VTEPs

    For redundant RPs, use Phantom RP.

For transitioning from PIM ASM to PIM BiDir or from PIM BiDir to PIM ASM underlay, we recommend that you use the following example procedure:

```
no ip pim rp-address 192.0.2.100 group-list 230.1.1.0/8
clear ip mroute *
clear ip mroute date-created *
clear ip pim route *
clear ip igmp groups *
clear ip igmp snooping groups * vlan all
```

Wait for all tables to clean up.

```
ip pim rp-address 192.0.2.100 group-list 230.1.1.0/8 bidir
```

- When entering the **no feature pim** command, NVE ownership on the route is not removed so the route stays and traffic continues to flow. Aging is done by PIM. PIM does not age out entries having a VXLAN encap flag.

- **ARP suppression**
  - Beginning with Cisco NX-OS Release 9.3(3), ARP suppression is supported for Cisco Nexus 9300-GX platform switches.
  - Beginning with Cisco NX-OS Release 9.3(5), ARP suppression is supported with reflective relay for Cisco Nexus 9300-FX/FX2/FXP, and 9300-GX platform switches. For information on reflective relay, see the *Cisco Nexus 9000 Series NX-OS Layer 2 Switching Configuration Guide*.
  - ARP suppression is supported for a VNI only if the VTEP hosts the First-Hop Gateway (Distributed Anycast Gateway) for this VNI. The VTEP and SVI for this VLAN must be properly configured for the Distributed Anycast Gateway operation (for example, global anycast gateway MAC address configured and anycast gateway with the virtual IP address on the SVI).
  - ARP suppression is a per-L2VNI fabric-wide setting in the VXLAN fabric. Enable or disable this feature consistently across all VTEPs in the fabric. Inconsistent ARP suppression configuration across VTEPs is not supported.

- **FCoE/NPV**

  Fibre Channel over Ethernet (FCoE) N-port Virtualization (NPV) can coexist with VXLAN on different fabric uplinks but on the same or different front-panel ports on Cisco Nexus 93180YC-FX switches.

  Fibre Channel N-port Virtualization (NPV) can coexist with VXLAN on different fabric uplinks but on the same or different front-panel ports on Cisco Nexus 93180YC-FX switches. VXLAN can exist only on the Ethernet front-panel ports and not on the FC front-panel ports.

- **Subinterfaces**
  - Beginning with Cisco NX-OS Release 9.3(5), the subinterfaces on VXLAN uplinks has the ability to carry non-VXLAN L3 IP traffic for Cisco Nexus 9300-FX/FX2/FXP, and 9300-GX platform switches and Cisco Nexus 9500 platform switches with Nexus X97160YC-EX, 9700-FX line cards. This feature is supported for VXLAN flood and learn and VXLAN EVPN, VXLAN EVPN Multi-Site, and DCI.
  - Beginning with Cisco NX-OS Release 10.5(2)F, the subinterfaces on VXLAN uplinks has the ability to carry non-VXLAN L3 IP traffic for Cisco Nexus 9500 platform switches with N9K-X9736C-FX3. This feature is supported for VXLAN flood and learn and VXLAN EVPN, VXLAN EVPN Multi-Site, and DCI.

- Beginning with Cisco NX-OS Release 10.1(1), VXLAN-encapsulated traffic over Parent Interface that Carries Subinterfaces is supported on Cisco Nexus 9300-FX3 platform switches.
- Beginning with Cisco NX-OS Release 9.3(5), VTEPs support VXLAN-encapsulated traffic over parent interfaces if subinterfaces are configured. This feature is supported for VXLAN flood and learn, VXLAN EVPN, VXLAN EVPN Multi-Site, and DCI. As shown in the following configuration example, VXLAN traffic is forwarded on the parent interface (eth1/1) in the default VRF, and L3 IP (non-VXLAN) traffic is forwarded on subinterfaces (eth1/1.10) in the tenant VRF.

```
interface ethernet 1/1
 description VXLAN carrying interface
 no switchport
 ip address 10.1.1.1/30

interface ethernet 1/1.10
 description NO VXLAN
 no switchport
 vrf member Tenant10
 encapsulation dot1q 10
  ip address 10.10.1.1/30
```

Restrictions of Cisco Nexus 9504 and 9508 switches with -R line cards

- For the Cisco Nexus 9504 and 9508 switches with -R line cards, VXLAN Layer 2 Gateway is supported on the 9636C-RX line card. VXLAN and MPLS cannot be enabled on the Cisco Nexus 9508 switch at the same time.
- For the Cisco Nexus 9504 and 9508 switches with -R line cards, if VXLAN is enabled, the Layer 2 Gateway cannot be enabled when there is any line card other than the 9636C-RX.
- For the Cisco Nexus 9504 and 9508 switches with -R line cards, PIM/ASM is supported in the underlay ports. PIM/Bidir is not supported. For more information, see the *Cisco Nexus 9000 Series NX_OS Multicast Routing Configuration Guide, Release 9.3(x)*.
- For the Cisco Nexus 9504 and 9508 switches with -R line cards, IPv6 hosts routing in the overlay is supported.
- For the Cisco Nexus 9504 and 9508 switches with -R line cards, ARP suppression is supported.
- For the Cisco Nexus 9504 and 9508 switches with -R line cards, VXLAN with ingress replication is not supported.
- VXLAN does not support coexistence with MVR and MPLS for Cisco Nexus 9504 and 9508 with -R line cards.
- For Cisco Nexus 9504 and 9508 switches with -R line cards, the L3VNI's VLAN must be added on the vPC peer-link trunk's allowed VLAN list.

Not supported features

- VXLAN is not supported on the N9K-C92348GC-FX3 switches.
- MDP is not supported for VXLAN configurations.
- Consistency checkers are not supported for VXLAN tables.
- VTEP connected to FEX host interface ports is not supported.

- Resilient hashing (port-channel load-balancing resiliency) and VXLAN configurations are not compatible with VTEPs using ALE uplink ports.



Note Resilient hashing is disabled by default.

- Routing protocol adjacencies using Anycast Gateway SVIs is not supported.
- Beginning with Cisco NX-OS Release 10.4(3)F, DHCP snooping (Dynamic Host Configuration Protocol snooping) is not supported on VXLAN VLANs.
- RACLs are not supported on Layer 3 uplinks for VXLAN traffic. Egress VACLs support is not available for de-capsulated packets in the network to access direction on the inner payload.

  As a best practice, use PACLs/VACLs for the access to the network direction.
- The QoS buffer-boost feature is not applicable for VXLAN traffic.
- The following limitations apply to releases prior to Cisco NX-OS Release 9.3(5):
  - VTEPs do not support VXLAN-encapsulated traffic over subinterfaces, regardless of VRF participation or IEEE 802.1Q encapsulation.
  - VTEPs do not support VXLAN-encapsulated traffic over parent interfaces if subinterfaces are configured, regardless of VRF participation.
  - Mixing subinterfaces for VXLAN and non-VXLAN VLANs is not supported.
- Point-to-multipoint Layer 3 and SVI uplinks are not supported.
- SVI and subinterfaces as uplinks are not supported.

Supported Features of CloudScale switches

Table 2: Supported Features of CloudScale switches

Features	Release	Platforms	Limitations
VXLAN	7.0(3)I7(3)	Cisco Nexus 9348GC-FXP switch	—
	—	Cisco Nexus 9500 platform switches with 9500-R and X97160YC-EX, 9700-FX line cards	—
	10.2(3)F and later	Cisco Nexus 9500 platform switches with 9700-GX line cards	—
	10.4(1)F and later	Cisco Nexus 9348GC-FX3, 9348GC-FX3PH and 9332D-H2R switches	
	10.4(2)F and later	Cisco Nexus 93400LD-H1 switches	—
	10.4(3)F and later	Cisco Nexus 9364C-H1 switches	—
	10.5(2)F and later	Cisco Nexus 9500 Series switches with N9K-X9736C-FX3	• VXLAN with vPC • DHCP snooping, ACL, and QoS policies on VXLAN VLANs. • IGMP snooping on VXLAN enabled VLANs. • Nested VXLAN (Host Overlay over Network Overlay) • PBR with IPv4 underlay • NVE source-interface hold-down timer for non-VPC VTEPs
DHCP snooping	—	—	—

Features	Release	Platforms	Limitations
ARP suppression	9.3(3) and later	Cisco Nexus 9300-GX platform switches	—
	9.3(5) and later	Cisco Nexus 9300-FX/FX2/FXP, and 9300-GX platform switches	supported with reflective relay. For information on reflective relay, see the *Cisco Nexus 9000 Series NX-OS Layer 2 Switching Configuration Guide*.
ITD and ePBR over VXLAN	10.1(1) and later	N9K-X9716D-GX TOR and N9K-C93180YC-FX3S platform switches.	—
PBR over VXLAN	10.1(1) and later	N9K-C9316D-GX, N9K-C93600CD-GX, and N9K-C9364C-GX	—
VXLAN flood and learn mode	9.3(6) and later	Cisco Nexus 9300-GX platform switches	—
	10.1(1) and later	N9K-C9316D-GX, N9K-C93600CD-GX, and N9K-C9364C-GX TOR switches.	—
BFD multihop over VXLAN with L3VNI interfaces	10.4(1)F and later		—
Border Spine	10.4(3)F and later	Cisco Nexus 9800 switches	For more information on the supported and not supported features, see Guidelines and Limitations for VXLAN EVPN Multi-Site, on page 297 and Guidelines and Limitations for TRM with Multi-Site, on page 325.
Dynamic Load Balancing (DLB)	10.5(1)F and later	Cisco Nexus 9300-FX3, GX, GX2, H2R, and H1 Series switches	Feature can be enabled on the underlay for VXLAN tunnels, allowing for ECMP routing on Layer 3 interfaces.

Table 3: Supported and Unsupported Features of DLB with Limitations

Features	Supported/Unsupported	Limitations
VXLAN standalone or vPC VTEP	Supported	—
Fabric peering	Supported	DLB is not supported when fabric peering with the local link is down, and traffic is rerouted over the PIP tunnel.
VXLAN Anycast and vPC BGWs	Supported	-
Layer 3 uplinks	Supported	Port channel, sub interfaces or SVIs are not supported.
VXLAN Traffic Engineering	Supported	VXLAN Traffic Engineering can coexist with DLB. However, DLB is not utilized for Traffic Engineering ECMP.
IPv4 and IPv6 underlay	Supported	—
VXLAN PBR	Unsupported	—

Guidelines and Limitations for VXLAN Support on Cisco Nexus 9364E-SG2 Series Switches

- Beginning with Cisco NX-OS Release 10.5(3)F, the following VXLAN features are supported on Cisco Nexus 9364E-SG2-Q and 9364E-SG2-O switches.
- For the VXLAN scale limit information, see the Cisco Nexus 9000 Series NX-OS Verified Scalability Guide, Release 10.5(3)F.
- The following table lists the supported VXLAN features as a leaf :

Features	Limitations
VXLAN BGP EVPN	—
VXLAN L2VNI and L3VNI	—
Distributed Anycast Gateway	This feature is supported with **fabric forwarding anycast-mode** command. This configuration is supported with the following combination of configuration • On vxlan-vlan only • With SVI configured/up on such vxlan vlan • With global Fabric Anycast Mac config in system
ARP Suppression	—

Features	Limitations
DHCP Relay	—
IPv4 and IPv6 unicast overlay traffic	—
IPv4 unicast underlay	—
BUM traffic	Supported only with IPv4 Ingress Replication
VXLAN Uplinks	• Only Ethernet and Port-channel routed interfaces are supported as uplinks. • SVI or L3-subinterfaces as uplinks are not supported.
VXLAN Counters	• VNI based total packet and byte counters are supported • peer counters or per-peer-per-vni counters are not supported.
VXLAN NGOAM	VTEP and Host reachability are supported
Unconditional DLB (Dynamic Load Balancing) for VXLAN underlay. For more information, see Dynamic Load Balancing on Silicon One switches.	—

- Cisco Nexus 9364E-SG2-Q and 9364E-SG2-O switches support anycast BGW functions with VRF or L3VNI.

  Use the **redistribute** or **route-map** commands on leaf and BGW to distribute the host information.

- Beginning with Cisco NX-OS Release 10.6(1)F, Cisco Nexus 9364E-SG2-Q and 9364E-SG2-O switches support anycast BGW functions with both VLAN (L2VNI) and VRF (L3VNI) together.



Note Support for VRF (L3VNI) was introduced in Cisco NX-OS Release 10.5(3)F.

- VLAN based L3VNI configuration is not supported. Only **vni** *vni-id* **l3** command is supported.
- The following table lists the supported VXLAN features as anycast BGW:

Features	Limitations
VXLAN BGP EVPN	—
L3VNI and L2VNI	—
Fabric advertise PIP	—
IPv4 and IPv6 unicast overlay traffic	—
IPv4 unicast underlay	—

Features	Limitations
VXLAN Uplinks	• Only Ethernet and Port-channel routed interfaces are supported as uplinks. • SVI or L3-subinterfaces as uplinks are not supported
VXLAN Counters	• VNI based total packet or byte counters are supported • peer counters or per-peer-per-vni counters are not supported.
VXLAN NGOAM	VTEP and Host reachability are supported
Unconditional DLB (Dynamic Load Balancing) for VXLAN underlay. For more details, see Dynamic Load Balancing on Silicon One switches.	—
L3 QoS policies	—

- Cisco Nexus 9364E-SG2-Q and 9364E-SG2-O switches do not support the following features:
  - VXLAN Flood and Learn
  - IPv4/IPv6 multicast overlay
  - vPC based multihoming
  - VXLAN IPv6 underlay
  - Multicast underlay (PIMv4 SM or PIMv6 ASM, PIM-BiDir, BUD node)
  - IGMP or MLD snooping
  - TRMv4/v6, and TRM data MDT
  - ND suppression
  - Virtual Network Functions (VNF) Multipath
  - VXLAN Policy-Based Routing
  - VXLAN QoS policy
  - DSVNI and route leak
  - VXLAN Access Features
    - Private VLAN (PVLAN)
    - 802.1x
    - Multitag
    - Cross Connect
    - Port security

- Port VLAN translation
- QinVNI, Selective QinVNI
- L2PT

- VXLAN ESI multihoming
- Firewall clustering with VXLAN
- Southbound loop detection
- Micro-segmentation with Security Groups (SGs) and Security Group ACLs (SGACLs)
- VXLAN Traffic Engineering

To configure VXLAN features, see Configuring VXLAN BGP EVPN, on page 113.

Guidelines and limitations for VXLAN support on Cisco Nexus 9336C-SE1 switches

Beginning with Cisco NX-OS Release 10.6(1)F, the following VXLAN features are supported on Cisco Nexus 9336C-SE1 switches.

For the VXLAN scale limit information, see the Cisco Nexus 9000 Series NX-OS Verified Scalability Guide, Release 10.6(1)F.

VXLAN supported features

Features	Note	Release
VXLAN BGP EVPN	—	10.6(1)F
VXLAN L2VNI and L3VNI	VLAN based L3VNI configuration is deprecated. Only **vni** *vnid* **l3** command is supported	10.6(1)F
VXLAN BGP EVPN Multi-Site Anycast Border Gateway	—	10.6(1)F
VXLAN BGP EVPN Border Spine	—	10.6(1)F
Standalone, vPC, and vPC fabric peering	Supported as leaf or border leaf	10.6(1)F
Distributed Anycast Gateway	This feature is supported with **fabric forwarding anycast-mode** command. This configuration is supported with the following combination of configuration • On VXLAN-VLAN only • With SVI configured or up on such VXLAN-VLAN • With global Fabric Anycast MAC configuration in system	10.6(1)F

Features	Note	Release
DHCP Relay	—	10.6(1)F
IPv4 and IPv6 unicast overlay traffic	—	10.6(1)F
BUM traffic • IPv4 unicast underlay (IR) • IPv4 multicast underlay with PIM ASM	Fabric and DCI	10.6(1)F
VXLAN uplinks	• Only Ethernet and Port-channel routed interfaces are supported as uplinks • SVI or L3-subinterfaces as uplinks are not supported	10.6(1)F
VXLAN counters	• VXLAN peer-based total packet/byte counters are supported • VNI based total packet and byte counters are supported • Peer counters or per-peer-per-vni counters are not supported	10.6(1)F
Underlay ECMP and Overlay ECMP (L3)	—	10.6(1)F
VXLAN NGOAM	VTEP and Host reachability are supported.	10.6(1)F
Multicast underlay BUD node	-	10.6(1)F
DSVNI and route leak	For more information on this feature support, see Guidelines and Limitations for VXLAN EVPN with Downstream VNI, on page 123	10.6(1)F
IGMP snooping	Not supported for Anycast BGW	10.6(1)F
ARP Suppression	—	10.6(1)F
TRMv4 L3 Mode • IPv4 unicast underlay (IR) • IPv4 multicast underlay with PIM ASM	• **Fabric**: only multicast underlay is supported • **DCI**: IR and multicast underlay are supported	10.6(1)F

Features	Note	Release
advertise-pip	**advertise-pip** command. For more information on advertise PIP, see the Configuring vPC Multi-Homing, on page 763 section.	10.6(1)F
BGW advertise using PIP towards fabric and DCI	**fabric-advertise-pip l3** command. For more information on advertise PIP, see the Advertise Using PIP Towards Fabric, on page 319 section **dci-advertise-pip**	10.6(1)F

VXLAN unsupported features

- vPC multi-site BGW
- RFC 5549 underlay
- IPv6 unicast or IR underlay (fabric)
- IPv6 Multicast underlay and Multicast underlay with PIM BIDIR (fabric)
- IPv6 IR underlay (DCI)
- Proportional ECMP (Mixed path)
- VNF (gateway IP)
- VXLAN Traffic Engineering
- ESI-RX or VXLAN ESI multihoming
- Cloudsec
- Separate counters for broadcast, multicast, and unicast traffic.
- MLD snooping
- ND suppression
- ACL on VxLAN
- EVPN Multisite Storm Control
- VXLAN Policy-Based Routing
- VXLAN QoS policy
- TRMv6, TRM L2 mode and TRM mixed mode, and TRM data MDT
- VXLAN Flood and Learn
- VXLAN IPv6 underlay
- FHS
- Null route or Static remote MAC

- VXLAN Access Features
  - Private VLAN (PVLAN)
  - 802.1x
  - Multitag
  - Cross Connect
  - Port security
  - Port VLAN translation
  - QinVNI, Selective QinVNI
  - L2PT
- VXLAN to SR/MPLS handoff
- Southbound loop detection
- VXLAN distributed NAT support
- VXLAN Static tunnels
- FEX
- SGACL

Considerations for VXLAN Deployment

- For scale environments, the VLAN IDs related to the VRF and Layer-3 VNI (L3VNI) must be reserved with the **system vlan nve-overlay id** command.

  This is required to optimize the VXLAN resource allocation to scale the following platforms:
  - Cisco Nexus 9300 platform switches
  - Cisco Nexus 9500 platform switches with 9500 line cards

  The following example shows how to reserve the VLAN IDs related to the VRF and the Layer-3 VNI:

```
system vlan nve-overlay id 2000

        vlan 2000
          vn-segment 50000

        interface Vlan2000
          vrf member MYVRF_50000
          ip forward
          ipv6 forward

        vrf context MYVRF_50000
          vni 50000
```



Note The **system vlan nve-overlay id** command should be used for a VRF or a Layer-3 VNI (L3VNI) only. Do not use this command for regular VLANs or Layer-2 VNIs (L2VNI).

- When configuring VXLAN BGP EVPN, the "System Routing Mode: Default" is applicable for the following hardware platforms:
  - Cisco Nexus 9300-FX/FX2/FX3 platform switches
  - Cisco Nexus 9300-GX platform switches
  - Cisco Nexus 9500 platform switches with X9500 line cards
  - Cisco Nexus 9500 platform switches with Nexus X97160YC-EX and 9700-FX line cards
  - Cisco Nexus 9500 platform switches with X9700-FX3 line cards
- The “System Routing Mode: template-vxlan-scale” is not applicable.
- When using VXLAN BGP EVPN in combination with Cisco NX-OS Release 7.0(3)I4(x) or NX-OS Release 7.0(3)I5(1), the “System Routing Mode: template-vxlan-scale” is required on the following hardware platforms:
  - Cisco Nexus 9500 Switches with Nexus X97160YC-EX line cards
- Beginning with Cisco NX-OS Release 10.3(1)F, support for extended dual-stack-host-scale template is provided for ARP, ND, and MAC on the Cisco Nexus 9300-FX3/GX/GX2B ToR switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, support for extended dual-stack-host-scale template is provided for ARP, ND, and MAC on the Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, support for extended dual-stack-host-scale template is provided for ARP, ND, and MAC on the Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, support for extended dual-stack-host-scale template is provided for ARP, ND, and MAC on the Cisco Nexus 9364C-H1 switches.
- To scale ARP and ND, use **system routing template-dual-stack-host-scale** command. For scaling limit, refer to *Cisco Nexus 9000 Series NX-OS Verified Scalability Guide*.
- Changing the “System Routing Mode” requires a reload of the switch.
- A loopback address is required when using the **source-interface config** command. The loopback address represents the local VTEP IP.
- During boot-up of a switch, you can use the **source-interface hold-down-time** *hold-down-time* command to suppress advertisement of the NVE loopback address until the overlay has converged. The range for the *hold-down-time* is 0 - 2147483647 seconds. The default is 300 seconds.



Note Though the loopback is still down, the traffic is encapsulated and sent to fabric.

- To establish IP multicast routing in the core, IP multicast configuration, PIM configuration, and RP configuration is required.
- VTEP to VTEP unicast reachability can be configured through any IGP protocol.
- In VXLAN flood and learn mode, the default gateway for VXLAN VLAN is recommended to be a centralized gateway on a pair of vPC devices with FHRP (First Hop Redundancy Protocol) running between them.
- While running VXLAN EVPN, with
  - any SVI for a VLAN extended over VXLAN is configured with anycast gateway and
  - any other mode of operation is not supported.

  If one VTEP is configured with an L2VNI and associated (with anycast gateway enabled), then every other VTEP where that L2VNI is locally defined has the SVI with anycast gateway configured.
- For flood and learn mode, only a centralized Layer 3 gateway is supported. Anycast gateway is not supported. The recommended Layer 3 gateway design would be a pair of switches in vPC to be the Layer 3 centralized gateway with FHRP protocol running on the SVIs. The same SVI's cannot span across multiple VTEPs even with different IP addresses used in the same subnet.



Note When configuring SVI with flood and learn mode on the central gateway leaf, it is mandatory to configure **hardware access-list tcam region arp-ether** *size* **double-wide**. (You must decrease the size of an existing TCAM region before using this command.)

For example:

```
hardware access-list tcam region arp-ether 256 double-wide
```

- VXLAN tunnels cannot have more than one underlay next hop on a given underlay port. For example, on a given output underlay port, only one destination MAC address can be derived as the outer MAC on a given output port.

  This is a per-port limitation, not a per-tunnel limitation. This means that two tunnels that are reachable through the same underlay port cannot drive two different outer MAC addresses.
- When changing the IP address of a VTEP device, you must shut the NVE interface before changing the IP address.
- As a best practice, when migrating any sets of VTEP to a multisite BGW, NVE interface must be shut on all the VTEPs where this migration is being performed. NVE interface should be brought back up once the migration is complete and all necessary configurations for multisite are applied to the VTEPs.
- As a best practice, the RP for the multicast group should be configured only on the spine layer. Use the anycast RP for RP load balancing and redundancy.

  The following is an example of an anycast RP configuration on spines:

```
ip pim rp-address 1.1.1.10 group-list 224.0.0.0/4
ip pim anycast-rp 1.1.1.10 1.1.1.1
ip pim anycast-rp 1.1.1.10 1.1.1.2
```

Note

- 1.1.1.10 is the anycast RP IP address that is configured on all RPs participating in the anycast RP set.
- 1.1.1.1 is the local RP IP.
- 1.1.1.2 is the peer RP IP.

- Static ingress replication and BGP EVPN ingress replication do not require any IP Multicast routing in the underlay.

vPC Considerations for VXLAN Deployment

- As a best practice, when **feature vpc** is enabled or disabled on a VTEP, the NVE interfaces on both the vPC primary and the vPC secondary must be shut down before the change is made. Enabling **feature vpc** without the vPC domain being properly configured will result in the NVE loopback being held administratively down until the configuration is completed and the vPC peer-link is brought up.
- Bind NVE to a loopback address that is separate from other loopback addresses that are required by Layer 3 protocols. A best practice is to use a dedicated loopback address for VXLAN.
- On vPC VXLAN, it is recommended to increase the **delay restore interface-vlan** timer under the vPC configuration, if the number of SVIs are scaled up. For example, if there are 1000 VNIs with 1000 SVIs, we recommend to increase the **delay restore interface-vlan** timer to 45 seconds.
- If a ping is initiated to the attached hosts on VXLAN VLAN from a vPC VTEP node, the source IP address used by default is the anycast IP that is configured on the SVI. This ping can fail to get a response from the host in case the response is hashed to the vPC peer node. This issue can happen when a ping is initiated from a VXLAN vPC node to the attached hosts without using a unique source IP address. As a workaround for this situation, use VXLAN OAM or create a unique loopback on each vPC VTEP and route the unique address via a backdoor path.
- The loopback address used by NVE needs to be configured to have a primary IP address and a secondary IP address.

  The secondary IP address is used for all VXLAN traffic that includes multicast and unicast encapsulated traffic.
- vPC peers must have identical configurations.
  - Consistent VLAN to vn-segment mapping.
  - Consistent NVE1 binding to the same loopback interface
    - Using the same secondary IP address.
    - Using different primary IP addresses.
  - Consistent VNI to group mapping.
- For multicast, the vPC node that receives the (S, G) join from the RP (rendezvous point) becomes the DF (designated forwarder). On the DF node, encap routes are installed for multicast.

Decap routes are installed based on the election of a decapper from between the vPC primary node and the vPC secondary node. The winner of the decap election is the node with the least cost to the RP. However, if the cost to the RP is the same for both nodes, the vPC primary node is elected.

The winner of the decap election has the decap mroute installed. The other node does not have a decap route installed.

- On a vPC device, BUM traffic (broadcast, unknown-unicast, and multicast traffic) from hosts is replicated on the peer-link. A copy is made of every native packet and each native packet is sent across the peer-link to service orphan-ports connected to the peer vPC switch.

  To prevent traffic loops in VXLAN networks, native packets ingressing the peer-link cannot be sent to an uplink. However, if the peer switch is the encapper, the copied packet traverses the peer-link and is sent to the uplink.



Note Each copied packet is sent on a special internal VLAN (VLAN 4041 or VLAN 4046).

- When the peer-link is shut, the loopback interface used by NVE on the vPC secondary is brought down and the status is **Admin Shut**. This is done so that the route to the loopback is withdrawn on the upstream and that the upstream can divert all traffic to the vPC primary.



Note Orphans connected to the vPC secondary will experience loss of traffic for the period that the peer-link is shut. This is similar to Layer 2 orphans in a vPC secondary of a traditional vPC setup.

- When the vPC domain is shut, the loopback interface used by NVE on the VTEP with shutdown vPC domain is brought down and the status is Admin Shut. This is done so that the route to the loopback is withdrawn on the upstream and that the upstream can divert all traffic to the other vPC VTEP.
- When peer-link is no-shut, the NVE loopback address is brought up again and the route is advertised upstream, attracting traffic.
- For vPC, the loopback interface has two IP addresses: the primary IP address and the secondary IP address.

  The primary IP address is unique and is used by Layer 3 protocols.

  The secondary IP address on loopback is necessary because the interface NVE uses it for the VTEP IP address. The secondary IP address must be same on both vPC peers.

- The vPC peer-gateway feature must be enabled on both peers to facilitate NVE RMAC/VMAC programming on both peers. For peer-gateway functionality, at least one backup routing SVI is required to be enabled across peer-link and also configured with PIM. This provides a backup routing path in the case when VTEP loses complete connectivity to the spine. Remote peer reachability is re-routed over peer-link in his case. In BUD node topologies, the backup SVI needs to be added as a static OIF for each underlay multicast group.

```
switch# sh ru int vlan 2

interface Vlan2
 description backupl_svi_over_peer-link
```

```
no shutdown
ip address 30.2.1.1/30
ip router ospf 1 area 0.0.0.0
ip pim sparse-mode
ip igmp static-oif route-map match-mcast-groups

route-map match-mcast-groups permit 1
 match ip multicast group 225.1.1.1/32
```



Note In BUD node topologies, the backup SVI needs to be added as a static OIF for each underlay multicast group.

The SVI must be configured on both vPC peers and requires PIM to be enabled.

- When the NVE or loopback is shut in vPC configurations:
  - If the NVE or loopback is shut only on the primary vPC switch, the global VXLAN vPC consistency checker fails. Then the NVE, loopback, and vPCs are taken down on the secondary vPC switch.
  - If the NVE or loopback is shut only on the secondary vPC switch, the global VXLAN vPC consistency checker fails. Then, the NVE, loopback, and secondary vPC are brought down on the secondary. Traffic continues to flow through the primary vPC switch.
  - As a best practice, you should keep both the NVE and loopback up on both the primary and secondary vPC switches.
- Redundant anycast RPs configured in the network for multicast load-balancing and RP redundancy are supported on vPC VTEP topologies.
- As a best practice, when changing the secondary IP address of an anycast vPC VTEP, the NVE interfaces on both the vPC primary and the vPC secondary must be shut before the IP changes are made.
- When SVI is enabled on a VTEP (flood and learn, or EVPN) regardless of ARP suppression, make sure that ARP-ETHER TCAM is carved using the **hardware access-list tcam region arp-ether 256 double-wide** command. This requirement does not apply to Cisco Nexus 9300-FX/FX2/FX3 and 9300-GX/GX2/H2R/H1 platform switches and Cisco Nexus 9500 platform switches with Nexus X97160YC-EX line cards.
- The **show** commands with the **internal** keyword are not supported.
- DHCP snooping (Dynamic Host Configuration Protocol snooping) is not supported on VXLAN VLANs.
- RACLs are not supported on Layer 3 uplinks for VXLAN traffic. Egress VACLs support is not available for de-capsulated packets in the network to access direction on the inner payload.

  As a best practice, use PACLs/VACLs for the access to the network direction.

  See the Cisco Nexus 9000 Series NX-OS Security Configuration Guide, Release 9.3(x) for other guidelines and limitations for the VXLAN ACL feature.
- QoS classification is not supported for VXLAN traffic in the network to access direction on the Layer 3 uplink interface.

  See the Cisco Nexus 9000 Series NX-OS Quality of Service Configuration Guide, Release 9.3(x) for other guidelines and limitations for the VXLAN QoS feature.
- The QoS buffer-boost feature is not applicable for VXLAN traffic.

- Beginning with Cisco NX-OS Release 9.3(5), VTEPs support VXLAN-encapsulated traffic over parent interfaces if subinterfaces are configured.
- VTEPs do not support VXLAN encapsulated traffic over subinterfaces. This is regardless of VRF participation or IEEE802.1Q encapsulation.
- Mixing subinterfaces for VXLAN and non-VXLAN VLANs is not supported.
- Point-to-multipoint Layer 3 and SVI uplinks are not supported.
- Using the **ip forward** command enables the VTEP to forward the VXLAN de-capsulated packet destined to its router IP to the SUP/CPU.
- Before configuring it as an SVI, the backup VLAN needs to be configured on Cisco Nexus 9300-FX/FX2/FX3 and 9300-GX platform switches as an infra-VLAN with the **system nve infra-vlans** command.
- VXLAN is supported on Cisco Nexus 9500 platform switches with the following line cards:
  - N9K-X97160YC-EX
  - 9700-FX
  - 9700-GX
  - 9700-FX3
- When Cisco Nexus 9500 platform switches are used as VTEPs, 100G line cards are not supported on Cisco Nexus 9500 platform switches. This limitation does not apply to a Cisco Nexus 9500 switch with X97160YC-EX or -FX or -FX3 line cards.
- Cisco Nexus 9300 platform switches with 100G uplinks only support VXLAN switching/bridging. Cisco Nexus 9300-FX/FX2 platform switches do not have this restriction.



Note For VXLAN routing support, a 40 G uplink module is required.

- The VXLAN UDP port number is used for VXLAN encapsulation. For Cisco Nexus NX-OS, the UDP port number is 4789. It complies with IETF standards and is not configurable.
- For Cisco Nexus 9200 platform switches that have the Application Spine Engine (ASE2). There exists a Layer 3 VXLAN (SVI) throughput issue. There is a data loss for packets of sizes 99 - 122.
- The VXLAN network identifier (VNID) 16777215 is reserved and should not be configured explicitly.
- VXLAN supports In Service Software Upgrade (ISSU).
- VXLAN ISSU is not supported on the Cisco Nexus 9300-GX platform switches.
- VXLAN does not support coexistence with the GRE tunnel feature or the MPLS (static or segment routing) feature.
- VTEP connected to FEX host interface ports is not supported.
- Resilient hashing (port-channel load-balancing resiliency) and VXLAN configurations are not compatible with VTEPs using ALE uplink ports.

Note Resilient hashing is disabled by default.

- When ARP suppression is enabled or disabled in a vPC setup, a down time is required because the global VXLAN vPC consistency checker will fail and the VLANs will be suspended if ARP suppression is disabled or enabled on only one side.

Note For information about VXLAN BGP EVPN scalability, see the Cisco Nexus 9000 Series NX-OS Verified Scalability Guide, Release 9.3(x).

Network Considerations for VXLAN Deployments

- MTU Size in the Transport Network

  Due to the MAC-to-UDP encapsulation, VXLAN introduces 50-byte overhead to the original frames. Therefore, the maximum transmission unit (MTU) in the transport network needs to be increased by 50 bytes. If the overlays use a 1500-byte MTU, the transport network needs to be configured to accommodate 1550-byte packets at a minimum. Jumbo-frame support in the transport network is required if the overlay applications tend to use larger frame sizes than 1500 bytes.

- ECMP and LACP Hashing Algorithms in the Transport Network

  As described in a previous section, Cisco Nexus 9000 Series Switches introduce a level of entropy in the source UDP port for ECMP and LACP hashing in the transport network. As a way to augment this implementation, the transport network uses an ECMP or LACP hashing algorithm that takes the UDP source port as an input for hashing, which achieves the best load-sharing results for VXLAN encapsulated traffic.

- Multicast Group Scaling

  The VXLAN implementation on Cisco Nexus 9000 Series Switches uses multicast tunnels for broadcast, unknown unicast, and multicast traffic forwarding. Ideally, one VXLAN segment mapping to one IP multicast group is the way to provide the optimal multicast forwarding. It is possible, however, to have multiple VXLAN segments share a single IP multicast group in the core network. VXLAN can support up to 16 million logical Layer 2 segments, using the 24-bit VNID field in the header. With one-to-one mapping between VXLAN segments and IP multicast groups, an increase in the number of VXLAN segments causes a parallel increase in the required multicast address space and the amount of forwarding states on the core network devices. At some point, multicast scalability in the transport network can become a concern. In this case, mapping multiple VXLAN segments to a single multicast group can help conserve multicast control plane resources on the core devices and achieve the desired VXLAN scalability. However, this mapping comes at the cost of suboptimal multicast forwarding. Packets forwarded to the multicast group for one tenant are now sent to the VTEPs of other tenants that are sharing the same multicast group. This causes inefficient utilization of multicast data plane resources. Therefore, this solution is a trade-off between control plane scalability and data plane efficiency.

  Despite the suboptimal multicast replication and forwarding, having multiple-tenant VXLAN networks to share a multicast group does not bring any implications to the Layer 2 isolation between the tenant networks. After receiving an encapsulated packet from the multicast group, a VTEP checks and validates

the VNID in the VXLAN header of the packet. The VTEP discards the packet if the VNID is unknown to it. Only when the VNID matches one of the VTEP's local VXLAN VNIDs, does it forward the packet to that VXLAN segment. Other tenant networks will not receive the packet. Thus, the segregation between VXLAN segments is not compromised.

Considerations for the Transport Network

The following are considerations for the configuration of the transport network:

- On the VTEP device:
  - Enable and configure IP multicast.*
  - Create and configure a loopback interface with a /32 IP address.

    (For vPC VTEPs, you must configure primary and secondary /32 IP addresses.)
  - Enable IP multicast on the loopback interface.*
  - Advertise the loopback interface /32 addresses through the routing protocol (static route) that runs in the transport network.
  - Enable IP multicast on the uplink outgoing physical interface.*
- Throughout the transport network:
  - Enable and configure IP multicast.*

For Cisco Nexus 9300-FX/FX2/FX3 and 9300-GX platform switches, the use of the **system nve infra-vlans** command is required. Otherwise, VXLAN traffic (IP/UDP 4789) is actively treated by the switch. The following scenarios are a non-exhaustive list but most commonly seen, where the need for a **system nve infra-vlans** definition is required.

Every VLAN that is not associated with a VNI (vn-segment) is required to be configured as a **system nve infra-vlans** in the following cases:

In the case of VXLAN flood and learn as well as VXLAN EVPN, the presence of non-VXLAN VLANs could be related to:

- An SVI related to a non-VXLAN VLAN is used for backup underlay routing between vPC peers via a vPC peer-link (backup routing).
- An SVI related to a non-VXLAN VLAN is required for connecting downstream routers (external connectivity, dynamic routing over vPC).
- An SVI related to a non-VXLAN VLAN is required for per Tenant-VRF peering (L3 route sync and traffic between vPC VTEPs in a Tenant VRF).
- An SVI related to a non-VXLAN VLAN is used for first-hop routing toward endpoints (Bud-Node).

In the case of VXLAN flood and learn, the presence of non-VXLAN VLANs could be related to:

- An SVI related to a non-VXLAN VLAN is used for an underlay uplink toward the spine (Core port).

The rule of defining VLANs as **system nve infra-vlans** can be relaxed for special cases such as:

- An SVI related to a non-VXLAN VLAN that does not transport VXLAN traffic (IP/UDP 4789).
- Non-VXLAN VLANs that are not associated with an SVI or not transporting VXLAN traffic (IP/UDP 4789).



Note You must not configure certain combinations of infra-VLANs. For example, 2 and 514, 10 and 522, which are 512 apart. This is specifically but not exclusive to the “Core port” scenario that is described for VXLAN flood and learn.



Note * Not required for static ingress replication or BGP EVPN ingress replication.

Considerations for Tunneling VXLAN

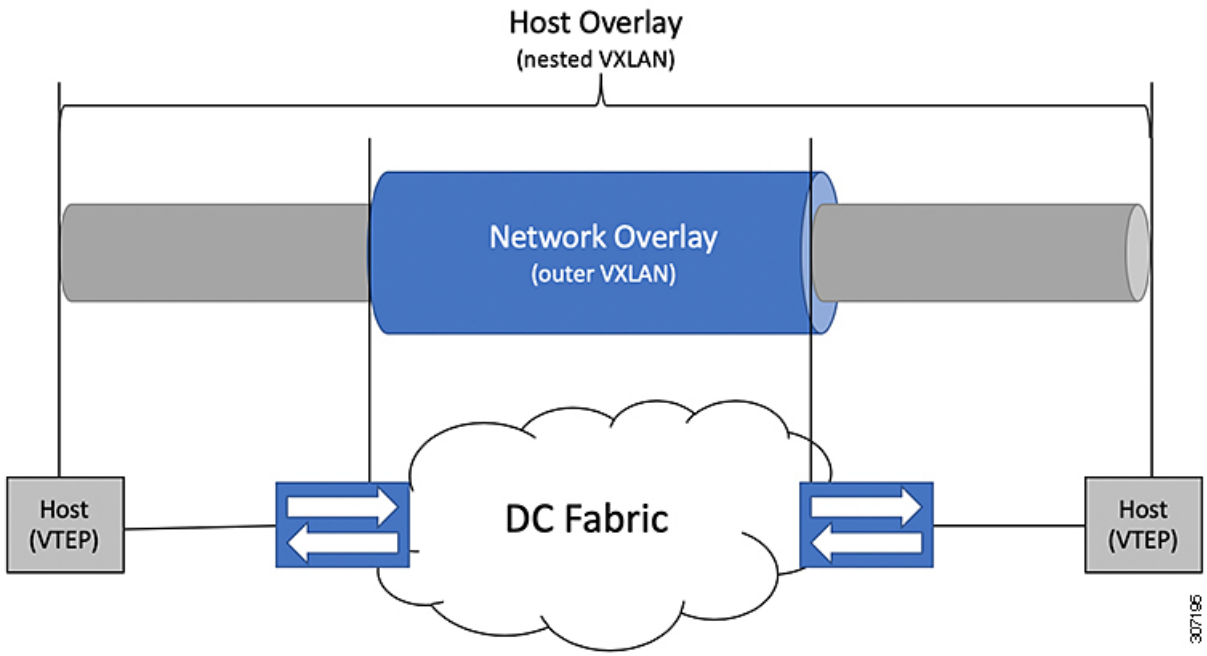
DC Fabrics with VXLAN BGP EVPN are becoming the transport infrastructure for overlays. These overlays, often originated on the server (Host Overlay), require integration or transport over the top of the existing transport infrastructure (Network Overlay).

Nested VXLAN (Host Overlay over Network Overlay) support has been added starting with Cisco NX-OS Release 7.0(3)I7(4) and Cisco NX-OS Release 9.2(2) on the Cisco Nexus 9300-FX, 9300-FX2, 9500-FX platform switches. It is also supported for Cisco Nexus 9300-FX3 platform switches starting with Cisco NX-OS Release 9.3(5).

Beginning with Cisco NX-OS Release 10.5(2)F, Nested VXLAN is supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.

Nested VXLAN is not supported on a Layer 3 interface or a Layer 3 port-channel interface in Cisco NX-OS Release 9.3(4) and prior releases. It is supported on a Layer 3 interface or a Layer 3 port-channel interface from Cisco NX-OS Release 9.3(5) onwards.

Figure 9: Host Overlay



To provide Nested VXLAN support, the switch hardware and software must differentiate between two different VXLAN profiles:

- VXLAN originated behind the Hardware VTEP for transport over VXLAN BGP EVPN (nested VXLAN)
- VXLAN originated behind the Hardware VTEP to integrated with VXLAN BGP EVPN (BUD Node)

The detection of the two different VXLAN profiles is automatic and no specific configuration is needed for nested VXLAN. As soon as VXLAN encapsulated traffic arrives in a VXLAN enabled VLAN, the traffic is transported over the VXLAN BGP EVPN enabled DC Fabric.

The following attachment modes are supported for Nested VXLAN:

- Untagged traffic (in native VLAN on a trunk port or on an access port)
- Tagged traffic Layer 2 ports (tagged VLAN on a IEEE 802.1Q trunk port)
- Untagged and tagged traffic that is attached to a vPC domain
- Untagged traffic on a Layer 3 interface or a Layer 3 port-channel interface
- Tagged traffic on Layer 3 interface or a Layer 3 port-channel interface

Configuring VXLAN

Enabling VXLANs

SUMMARY STEPS

1. **configure terminal**
2. [**no**] **feature nv overlay**
3. [**no**] **feature vn-segment-vlan-based**
4. (Optional) **copy running-config startup-config**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	[**no**] **feature nv overlay**	Enables the VXLAN feature.
Step 3	[**no**] **feature vn-segment-vlan-based**	Configures the global mode for all VXLAN bridge domains.
Step 4	(Optional) **copy running-config startup-config**	Saves the change persistently through reboots and restarts by copying the running configuration to the startup configuration.

Mapping VLAN to VXLAN VNI

SUMMARY STEPS

1. **configure terminal**
2. **vlan** *vlan-id*
3. **vn-segment** *vnid*
4. **exit**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**vlan** *vlan-id*	Specifies VLAN.

	Command or Action	Purpose
Step 3	**vn-segment** *vnid*	Specifies VXLAN VNID (Virtual Network Identifier)
Step 4	**exit**	Exit configuration mode.

Creating and Configuring an NVE Interface and Associate VNIs

An NVE interface is the overlay interface that terminates VXLAN tunnels.

You can create and configure an NVE (overlay) interface with the following:

SUMMARY STEPS

1. **configure terminal**
2. **interface nve** *x*
3. **source-interface** *src-if*
4. **member vni** *vni*
5. **mcast-group** *start-address* [*end-address*]

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**interface nve** *x*	Creates a VXLAN overlay interface that terminates VXLAN tunnels. **Note** Only 1 NVE interface is allowed on the switch.
Step 3	**source-interface** *src-if*	The source interface must be a loopback interface that is configured on the switch with a valid /32 IP address. This /32 IP address must be known by the transient devices in the transport network and the remote VTEPs. This is accomplished by advertising it through a dynamic routing protocol in the transport network.
Step 4	**member vni** *vni*	Associate VXLAN VNIs (Virtual Network Identifiers) with the NVE interface.
Step 5	**mcast-group** *start-address* [*end-address*]	Assign a multicast group to the VNIs. **Note** Used only for BUM traffic

Creating and Configuring an NVE Interface Loopback

Traditionally, a single loopback interface is configured as the NVE source interface, where both PIP and VIP of vPC complex are configured. You can configure a separate loop back for CloudSec enabled vPC BGW. Cisco recommends you to use separate loopback interfaces for source and anycast IP addresses under NVE for better convergence in MLAG deployments. The IP address that is configured on the source-interface is the PIP of the vPC node, and the IP address that is configured on the anycast interface is the VIP of that vPC complex. The secondary IP configured on the NVE source interface has no effect if the NVE anycast interface is also configured.

With separate loopbacks, the convergence for dual-attached EVPN Type-2 and Type-5 routes traffic that is destined for the DCI side will be improved.

From Cisco NX-OS Release 10.4(1)F, an orphan Type-2 host is advertised using PIP and a vPC Type-2 host is advertised using VIP. This is the default behavior for a Type-2 host for vMCT. PIP is up with the NVE interface before the hold down timer expires. Thus, all routes with PIP next-hop advertise before the hold-down timer expires. The routes include orphan Type-2 routes in vMCT and local Type-5 routes learned through redist HMM, direct or connected routes in vPC/vMCT.

The fabric-ready timer is added in vPC to indicate when orphan or locally attached routes can be advertised. The timer helps to enhance the convergence of the orphan or locally attached routes.



Note If you do not configure the fabric convergence timer in vPC node, default value of the timer is set to 75% of hold-down timer.

SUMMARY STEPS

1. **configure terminal**
2. **interface nve** *value*
3. **source-interface***loopback-interface-identifier*
4. (Optional) **source-interface** [**loopback-interface-identifier**]**anycast loopback**[**loopback-interface-identifier**]
5. **show nve interface nve1 detail**

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# configure terminal`
	`switch(config)#`
		Enters global configuration mode.

Step 2	**interface nve** *value* **Example:** `switch(config-if-nve)# interface nve 1`	Creates a VXLAN overlay interface that terminates VXLAN tunnels. **Note** Only 1 NVE interface is allowed on the switch.

	Command or Action	Purpose
Step 3	**source-interface***loopback-interface-identifier* **Example:** `switch(config-if-nve)# source-interface loopback 1`	Sets a loopback interface as the source interface for the VTEP.
Step 4	(Optional) **source-interface** [**loopback-interface-identifier**]**anycast loopback**[**loopback-interface-identifier**] **Example:** `switch(config-if-nve)# source-interface loopback 1 anycast loopback2`	Configures anycast loopback interface. **Note** This configuration exists for IPv6 underlay from earlier releases. From this release, configuration is added for IPv4 underlay.
Step 5	**show nve interface nve1 detail**	Shows information about the configured anycast loopback interface.

Example

The following configuration example shows configuring anycast loopback interface:

```
switch# configure terminal
switch(config)# interface nve 1
switch(config-if-nve)# source-interface loopback 1
switch (config-if-nve)# source-interface loopback 1 anycast loopback 4
```

The below example displays the show command for configured loopback interface on switch. This show command displays the details such as anycast loopback interface, IP associated with the anycast interface, state of the interface, and fabric convergence timer.



Note Fabric convergence timer default value is 135 seconds.

```
switch(config-if-nve)# show nve interface nve1 detail
Interface: nve1, State: Up, encapsulation: VXLAN
VPC Capability: VPC-VIP-Only [notified]
Local Router MAC: e41f.7b2e.977f
Host Learning Mode: Control-Plane
Source-Interface: loopback1 (primary: 20.1.0.15)
Anycast-Interface: loopback4 (secondary: 20.1.0.145)
Source Interface State: Up
Anycast Interface State: Up
Virtual RMAC Advertisement: Yes
NVE Flags:
Interface Handle: 0x49000001
Source Interface hold-down-time: 120
Source Interface hold-up-time: 30
Remaining hold-down time: 0 seconds
Virtual Router MAC: 0200.1401.0091
Interface state: nve-intf-add-complete
Fabric convergence time: 90 seconds
Fabric convergence time left: 0 seconds
```

Note You cannot downgrade the switch to a lower version, for which the split loopback feature is not supported. You can downgrade the switch to version which supports split loopback in MLAG deployment only if the downgrade is initiated from MLAG configuration.

Migration from Single NVE Source Loopback Interface to Separate Source Loopback

You can move existing vPC deployments with a single NVE source loopback interface to another source loopback for VIP and PIP. This migration has less impact on traffic loss and for help you to move existing to split loopback deployments.

Do the following procedure to migrate single NVE to split loopback deployment:

1. Isolate vPC secondary. This is to ensure the traffic flows through only primary.

   On vPC secondary, perform the following:

   a. ip pim isolate
   b. router bgp 2
   c. Isolate
   d. router ospf underlay
   e. Isolate
   f. sleep instance 2 20
   g. vPC domain 100
   h. shutdown

2. On vPC secondary
   a. Remove secondary IP on primary interface.
   b. Configure an anycast interface with the same IP address as the previous secondary. Due to this new behaviour, there is no failure of vPC CC and NVE will be up.

3. Connect vPC secondary. Allow the holddown timer to expire.
4. Change the vPC role.
5. Repeat the steps 1 to 3 for a new vPC secondary. This ensures that the configuration is changed and updated with new configuration for both new vPC secondary and vPC boxes.

Configuring a VXLAN VTEP in vPC

You can configure a VXLAN VTEP in a vPC.

SUMMARY STEPS

1. Enter global configuration mode.
2. Enable the vPC feature on the device.
3. Enable the interface VLAN feature on the device.
4. Enable the LACP feature on the device.
5. Enable the PIM feature on the device.
6. Enables the OSPF feature on the device.
7. Define a PIM RP address for the underlay multicast group range.
8. Define a non-VXLAN enabled VLAN as a backup routed path.
9. Create the VLAN to be used as an infra-VLAN.
10. Create the SVI used for the backup routed path over the vPC peer-link.
11. Create primary and secondary IP addresses.
12. Create a primary IP address for the data plane loopback interface.
13. Create a vPC domain.
14. Configure the IPv4 address for the remote end of the vPC peer-keepalive link.
15. Enable Peer-Gateway on the vPC domain.
16. Enable Peer-switch on the vPC domain.
17. Enable IP ARP synchronize under the vPC domain to facilitate faster ARP table population following device reload.
18. (Optional) Enable IPv6 nd synchronization under the vPC domain to facilitate faster nd table population following device reload.
19. Create the vPC peer-link port-channel interface and add two member interfaces.
20. Modify the STP hello-time, forward-time, and max-age time.
21. (Optional) Enable the delay restore timer for SVI's.

DETAILED STEPS

Procedure

Step 1 Enter global configuration mode.

```
switch# configure terminal
```

Step 2 Enable the vPC feature on the device.

```
switch(config)# feature vpc
```

Step 3 Enable the interface VLAN feature on the device.

```
switch(config)# feature interface-vlan
```

Step 4 Enable the LACP feature on the device.

```
switch(config)# feature lacp
```

Step 5 Enable the PIM feature on the device.

```
switch(config)# feature pim
```

Step 6 Enables the OSPF feature on the device.

```
switch(config)# feature ospf
```

Step 7 Define a PIM RP address for the underlay multicast group range.

```
switch(config)# ip pim rp-address 192.168.100.1 group-list 224.0.0/4
```

Step 8 Define a non-VXLAN enabled VLAN as a backup routed path.

```
switch(config)# system nve infra-vlans 10
```

Step 9 Create the VLAN to be used as an infra-VLAN.

```
switch(config)# vlan 10
```

Step 10 Create the SVI used for the backup routed path over the vPC peer-link.

```
switch(config)# interface vlan 10
switch(config-if)# ip address 10.10.10.1/30
switch(config-if)# ip router ospf UNDERLAY area 0
switch(config-if)# ip pim sparse-mode
switch(config-if)# no ip redirects
switch(config-if)# mtu 9216
(Optional)switch(config-if)# ip igmp static-oif route-map match-mcast-groups
switch(config-if)# no shutdown
(Optional)switch(config)# route-map match-mcast-gropus permit 10
(Optional)switch(config-route-map)# match ip multicast group 225.1.1.1/32
```

Step 11 Create primary and secondary IP addresses.

```
switch(config)# interface loopback 0
switch(config-if)# description Control_plane_Loopback
switch(config-if)# ip address x.x.x.x/32
switch(config-if)# ip router ospf process tag area area id
switch(config-if)# ip pim sparse-mode
switch(config-if)# no shutdown
```

Step 12 Create a primary IP address for the data plane loopback interface.

```
switch(config)# interface loopback 1
switch(config-if)# description Data_Plane_loopback
switch(config-if)# ip address z.z.z.z/32
switch(config-if)# ip address y.y.y.y/32 secondary
switch(config-if)# ip router ospf process tag area area id
switch(config-if)# ip pim sparse-mode
switch(config-if)# no shutdown
```

Step 13 Create a vPC domain.

```
switch(config)# vpc domain 5
```

Step 14 Configure the IPv4 address for the remote end of the vPC peer-keepalive link.

```
switch(config-vpc-domain)# peer-keepalive destination 172.28.230.85
```

Note
The system does not form the vPC peer link until you configure a vPC peer-keepalive link

The management ports and VRF are the defaults.

Note
We recommend that you configure a separate VRF and use a Layer 3 port from each vPC peer device in that VRF for the vPC peer-keepalive link. For more information about creating and configuring VRFs, see the Cisco Nexus 9000 Series NX-OS Unicast Routing Configuration Guide.

Step 15 Enable Peer-Gateway on the vPC domain.

```
switch(config-vpc-domain)# peer-gateway
```

Note
Disable IP redirects on all interface-vlans of this vPC domain for correct operation of this feature.

Step 16 Enable Peer-switch on the vPC domain.

```
switch(config-vpc-domain)# peer-switch
```

Note
Disable IP redirects on all interface-vlans of this vPC domain for correct operation of this feature.

Step 17 Enable IP ARP synchronize under the vPC domain to facilitate faster ARP table population following device reload.

```
switch(config-vpc-domain)# ip arp synchronize
```

Step 18 (Optional) Enable IPv6 nd synchronization under the vPC domain to facilitate faster nd table population following device reload.

```
switch(config-vpc-domain)# ipv6 nd synchronize
```

Step 19 Create the vPC peer-link port-channel interface and add two member interfaces.

```
switch(config)# interface port-channel 1
switch(config-if)# switchport
switch(config-if)# switchport mode trunk
switch(config-if)# switchport trunk allowed vlan 1,10,100-200
switch(config-if)# mtu 9216
switch(config-if)# vpc peer-link
switch(config-if)# no shutdown
switch(config-if)# interface Ethernet 1/1 , 1/21
switch(config-if)# switchport
switch(config-if)# mtu 9216
switch(config-if)# channel-group 1 mode active
switch(config-if)# no shutdown
```

Step 20 Modify the STP hello-time, forward-time, and max-age time.

As a best practice, we recommend changing the **hello-time** to four seconds to avoid unnecessary TCN generation when the vPC role change occurs. As a result of changing the **hello-time**, it is also recommended to change the **max-age** and **forward-time** accordingly.

```
switch(config)# spanning-tree vlan 1-3967 hello-time 4
switch(config)# spanning-tree vlan 1-3967 forward-time 30
switch(config)# spanning-tree vlan 1-3967 max-age 40
```

Step 21 (Optional) Enable the delay restore timer for SVI's.

We recommend that you tune this value when the SVI or VNI scale is high. For example, when the SVI count is 1000, we recommended setting the delay restore for interface-vlan to 45 seconds.

```
switch(config-vpc-domain)# delay restore interface-vlan 45
```

Configuring Static MAC for VXLAN VTEP

Static MAC for VXLAN VTEP is supported on Cisco Nexus 9300 Series switches with flood and learn. This feature enables the configuration of static MAC addresses behind a peer VTEP.

Note Static MAC cannot be configured for a control plane with a BGP EVPN-enabled VNI.

SUMMARY STEPS

1. **configure terminal**
2. **mac address-table static** *mac-address* **vni** *vni-id* **interface nve** *x* **peer-ip** *ip-address*
3. **exit**
4. (Optional) **copy running-config startup-config**
5. (Optional) **show mac address-table static interface nve** *x*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**mac address-table static** *mac-address* **vni** *vni-id* **interface nve** *x* **peer-ip** *ip-address*	Specifies the MAC address pointing to the remote VTEP.
Step 3	**exit**	Exits global configuration mode.
Step 4	(Optional) **copy running-config startup-config**	Saves the change persistently through reboots and restarts by copying the running configuration to the startup configuration.
Step 5	(Optional) **show mac address-table static interface nve** *x*	Displays the static MAC addresses pointing to the remote VTEP.

Example

The following example shows the output for a static MAC address configured for VXLAN VTEP:

```
switch# show mac address-table static interface nve 1

Legend:
        * - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC
        age - seconds since last seen,+ - primary entry using vPC Peer-Link,
        (T) - True, (F) - False
   VLAN     MAC Address      Type      age     Secure NTFY Ports
---------+-----------------+--------+---------+------+----+------------------
*  501     0047.1200.0000   static   -         F      F    nve1(33.1.1.3)
*  601     0049.1200.0000   static   -         F      F    nve1(33.1.1.4)
```

Disabling VXLANs

SUMMARY STEPS

1. **configure terminal**
2. **no feature vn-segment-vlan-based**
3. **no feature nv overlay**
4. (Optional) **copy running-config startup-config**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**no feature vn-segment-vlan-based**	Disables the global mode for all VXLAN bridge domains
Step 3	**no feature nv overlay**	Disables the VXLAN feature.
Step 4	(Optional) **copy running-config startup-config**	Saves the change persistently through reboots and restarts by copying the running configuration to the startup configuration.

Configuring BGP EVPN Ingress Replication

The following enables BGP EVPN with ingress replication for peers.

SUMMARY STEPS

1. **configure terminal**
2. **interface nve** *x*
3. **source-interface** *src-if*
4. **member vni** *vni*
5. **ingress-replication protocol bgp**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**interface nve** *x*	Creates a VXLAN overlay interface that terminates VXLAN tunnels. **Note**

	Command or Action	Purpose
		Only 1 NVE interface is allowed on the switch.
Step 3	**source-interface** *src-if*	The source interface must be a loopback interface that is configured on the switch with a valid /32 IP address. This /32 IP address must be known by the transient devices in the transport network and the remote VTEPs. This is accomplished by advertising it through a dynamic routing protocol in the transport network.
Step 4	**member vni** *vni*	Associate VXLAN VNIs (Virtual Network Identifiers) with the NVE interface.
Step 5	**ingress-replication protocol bgp**	Enables BGP EVPN with ingress replication for the VNI.

Configuring Static Ingress Replication

The following enables static ingress replication for peers.

SUMMARY STEPS

1. **configuration terminal**
2. **interface nve** *x*
3. **member vni** [*vni-id* | *vni-range*]
4. **ingress-replication protocol static**
5. **peer-ip** *n.n.n.n*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configuration terminal**	Enters global configuration mode.
Step 2	**interface nve** *x*	Creates a VXLAN overlay interface that terminates VXLAN tunnels. **Note** Only 1 NVE interface is allowed on the switch.
Step 3	**member vni** [*vni-id*	*vni-range*]	Maps VXLAN VNIs to the NVE interface.
Step 4	**ingress-replication protocol static**	Enables static ingress replication for the VNI.
Step 5	**peer-ip** *n.n.n.n*	Enables peer IP.

VXLAN and IP-in-IP Tunneling

Cisco NX-OS Release 9.3(6) and later releases support the coexistence of VXLAN and IP-in-IP tunneling.

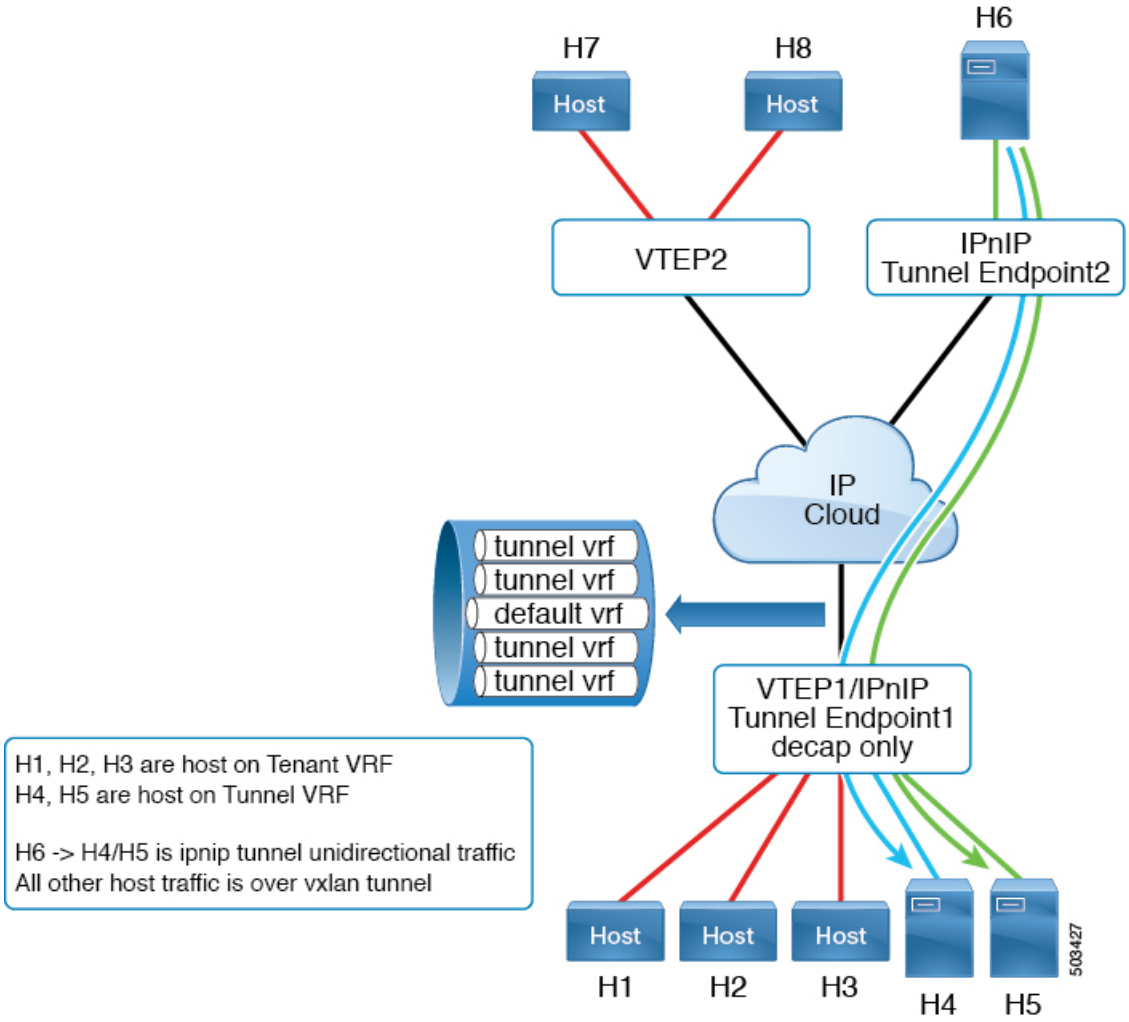
Coexistence of these features requires isolating IP-in-IP tunnels and VXLAN within their own VRFs. By isolating the VRFs, both VXLAN and the tunnels operate independently. VXLAN tunnel termination isn't reencapsulated as an IP-in-IP tunnel (or conversely) on the same or different VRFs.

By configuring subinterfaces under the interface to isolate VRFs, the same uplinks can be used to carry both VXLAN and IP-in-IP tunnel traffic. The parent port can be on the default VRF and subinterfaces on the non-default VRFs.

To terminate IP-in-IP encapsulated packets received on port-channel sub-interfaces, these sub-interfaces must be configured under the same non-default VRF as the tunnel interface, and can only be member of ***one*** non-default VRF.

Multiple port-channel sub interfaces from a different parent PC can still be configured under the same non-default VRF to terminate IP-in-IP encapsulation. The limitation only applies for sub-interfaces under one port-channel. This limitation is not applicable for L3 ports.

As the following example shows, VXLAN traffic is forwarded on the parent interface (eth1/1) in the default VRF, and IP-in-IP (non-VXLAN) traffic is forwarded on subinterfaces (eth1/1.10) in the tunnel VRF.



Cisco Nexus 9300-FX2 platform switches support the coexistence of VXLAN and IP-in-IP tunneling with the following limitations:

- VXLAN must be configured in the default VRF.
- Coexistence is supported on VXLAN with the EVPN control plane.
- IP-in-IP tunneling must be configured in the non-default VRF and is supported only in decapsulate-any mode.

Note If you try to enable VXLAN when a decapsulate-any tunnel is configured in the default VRF, an error message appears. It states that VXLAN and IP-in-IP tunneling can coexist only for a decapsulate-any tunnel in the non-default VRF and to remove the configuration.

- Point-to-point GRE tunnels are not supported. If you try to configure point-to-point tunnels, an error message appears indicating that VXLAN and IP-in-IP tunneling can coexist only for a decapsulate-any tunnel.

- Typically to configure a tunnel, you need to provide the two endpoints. However, decapsulate-any is a receive-only tunnel, so you need to provide only the source IP address or source interface name. The tunnel terminates on any IP interface in the same VRF.
- Tunnel statistics don't support egress counters.
- VXLAN and IP-in-IP tunnels can't share the same source loopback interface. Each tunnel must have its own source loopback interface.

The following example shows a sample configuration:

```
feature vn-segment-vlan-based
feature nv overlay
feature tunnel
nv overlay evpn

interface ethernet 1/1
    description VXLAN carrying interface
    no switchport
    ip address 10.1.1.1/30

interface ethernet 1/1.10
    description IPinIP carrying interface
    no switchport
    vrf member tunnel
    encapsulation dot1q 100
    ip address 10.10.1.1/30

interface loopback 0
    description VXLAN-loopback
    ip address 125.125.125.125/32

interface loopback 100
    description Tunnel_loopback
    vrf member tunnel
    ip address 5.5.5.5/32

interface Tunnel1
    vrf member tunnel
    ip address 55.55.55.1/24
    tunnel mode ipip decapsulate-any ip
    tunnel source loopback100
    tunnel use-vrf tunnel
    no shutdown

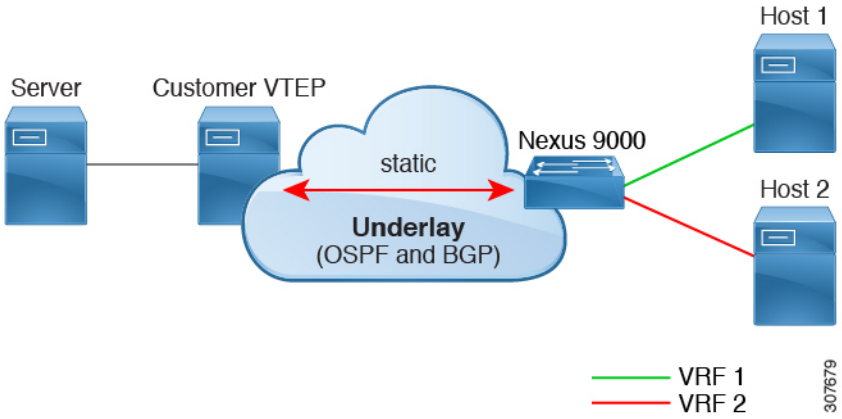
interface nve1
    host-reachability protocol bgp
    source-interface loopback0
    global mcast-group 224.1.1.1 L2
    global mcast-group 225.3.3.3 L3
    member vni 10000
    suppress-arp
    ingress-replication protocol bgp
    member vni 55500 associate-vrf
```

Configuring VXLAN Static Tunnels

About VXLAN Static Tunnels

Beginning with Cisco NX-OS Release 9.3(3), some Cisco Nexus switches can connect to a customer-provided software VTEP over static tunnels. Static tunnels are customer defined and support VXLAN-encapsulated traffic between hosts without requiring a control plane protocol such as BGP EVPN. You can configure static tunnels manually from the Nexus switch or programmatically, such as through a NETCONF client in the underlay.

Figure 10: VXLAN Static Tunnel Connecting Software VTEP



Static tunnels are supported per VRF. Each VRF can have a dedicated L3VNI to transport a packet with proper encapsulation and decapsulation on the switch and the software VTEP, the static peer. Typically, the static peer is a Cisco Nexus 1000V or bare-metal server with one or more VMs terminating one or more VNIs. However, a static peer can be any customer-developed device that complies with RFC 7348, *Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks*. Because the customer provides the static peer and a control plane protocol is not present, you must ensure that the static peer forwards the VXLAN-related configuration and routes to the correct hosts.

Beginning with Cisco NX-OS Release 9.3(5), this feature supports the handling of packets coming in and going out of the tunnel. Specifically, it allows the Nexus switch to send packets to the hosts or other switches over the tunnel. In Cisco NX-OS Releases 9.3(3) and 9.3(4), VXLAN static tunnels support communication only from the local host to the remote host.

Guidelines and Limitations for VXLAN Static Tunnels

The VXLAN static tunnels feature has the following guidelines and limitations:

- The Cisco Nexus 9300-FX/FX2/FX3, 9300-GX and 9300-FX3platform switches support VXLAN static tunnels.
- Beginning with Cisco NX-OS Release 10.1(1), VXLAN Static Tunnels are supported on Cisco Nexus 9300-FX3 platform switches.
- Beginning with Cisco NX-OS Release 10.2(3)F, the VXLAN Static Tunnels are supported on Cisco Nexus 9300-GX2 platform switches.

- Beginning with Cisco NX-OS Release 10.4(1)F, the VXLAN Static Tunnels are supported on Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, the VXLAN Static Tunnels are supported on Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, the VXLAN Static Tunnels are supported on Cisco Nexus 9364C-H1 switches.
- The following guidelines apply to software VTEPs:
  - The software VTEP must be configured as needed to determine how to forward traffic from the VNI.
  - The software VTEP must be compliant with RFC 7348.
- The underlay can be OSPFv2, BGP, IS-IS, or IPv4.
- The overlay can be IPv4 only.
- Additional VXLAN features (such as TRM, Multi-Site, OAM, Cross Connect, and VXLAN QoS), IGMP snooping, MPLS handoff, static MPLS, SR, and SRv6 are not supported.
- Pings across the overlay from local tenant VRF loopback to a host behind the software VTEP is not supported.
- Static tunnels do not support ECMP configuration.
- Static tunnels cannot be configured in the same fabric as traditional flood and learn or BGP EVPN fabrics.
- Local hosts are not supported for VNI-enabled VLANs. Therefore, you cannot have a host in the same VLAN where you configured the VNI.
- Fabric forwarding is supported with static tunnels. When fabric forwarding is enabled, be aware that it affects how SVIs and MAC addresses are used. Consider the following example configuration.

```
feature fabric forwarding
fabric forwarding anycast-gateway-mac 0000.0a0a.0a0a

interface Vlan802
no shutdown
vrf member vrfvxlan5201
ip address 103.33.1.1/16
fabric forwarding mode anycast-gateway
```

  When fabric forwarding is enabled:
  - all SVIs where **fabric forwarding mode anycast-gateway** is configured (for example, Vlan802) are used.
  - the MAC address configured with **fabric forwarding anycast-gateway-mac anycast-mac-address** (0000.0a0a.0a0a) is used.

Enabling VXLAN Static Tunnels

Enable the following features to enable VXLAN Static Tunnels.

SUMMARY STEPS

1. **config terminal**
2. **feature vn-segment**
3. **feature ofm**

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**config terminal**
	Example:
	`switch# configure terminal`
	`switch(config)#`
		Enter configuration mode.

Step 2	**feature vn-segment**
	Example:
	`switch(config)# feature vn-segment`
	`switch(config)#`
		Enable VLAN-based VXLAN.

Step 3	**feature ofm**
	Example:
	`switch(config)# feature ofm`
	`switch(config)#`
		Enable static VXLAN tunnels.

What to do next

Configure the VRF overlay VLAN for VXLAN routing over Static Tunnels.

Configuring VRF Overlay for Static Tunnels

A VRF overlay must be configured for the VXLAN Static Tunnels.

SUMMARY STEPS

1. **vlan** *number*
2. **vn-segment** *number*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**vlan** *number* **Example:**	Specify the VLAN.

	Command or Action	Purpose
	`switch(config)# vlan 2001` `switch(config-vlan)#`	

Step 2	**vn-segment** *number*
	Example:
	`switch(config-vlan)# vn-segment 20001`
	`switch(config-vlan)#`
		Specify the VN segment.

What to do next

Configure the VRF for VXLAN Routing over the Static Tunnel.

Configuring a VRF for VXLAN Routing

Configure the tenant VRF.

SUMMARY STEPS

1. **vrf context** *vrf-name*
2. **vni** *number*

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**vrf context** *vrf-name*
	Example:
	`switch(config-vlan)# vrf context cust1`
	`switch(config-vrf)#`
		Configure the tenant VRF.

Step 2	**vni** *number*
	Example:
	`switch(config-vrf)# vni 20001`
	`switch(config-vrf)#`
		Specify the VNI for the tenant VRF.

What to do next

Configure the L3 VNI for the host.

Configuring the L3 VNI for Static Tunnels

Configure the L3 VNI for the VTEPs.

Before you begin

The VLAN interface feature must be enabled. Use **feature interface-vlan** if needed.

SUMMARY STEPS

1. **vlan** *number*
2. **interface** *vlan-number*
3. **vrf member** *vrf-name*
4. **ip forward**
5. **no shutdown**

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**vlan** *number*
	Example:
	`switch(config-vrf)# vlan 2001`
	`switch(config-vlan)#`
		Specify the VLAN number

Step 2	**interface** *vlan-number*
	Example:
	`switch(config)# interface vlan2001`
	`switch(config-if)#`
		Specify the VLAN interface.

Step 3	**vrf member** *vrf-name*
	Example:
	`switch(config-if)# vrf member cust1`
	`Warning: Deleted all L3 config on interface Vlan2001`
	`switch(config-if)#`
		Assign the VLAN interface to the tenant VRF.

Step 4	**ip forward**
	Example:
	`switch(config-if)# ip forward`
	`switch(config-if)#`
		Enable IPv4 traffic on the interface.

Step 5	**no shutdown**
	Example:
	`switch(config-if)# no shutdown`
	`switch(config-if)#`
		Enables the interface.

What to do next

Configure the tunnel profile.

Configuring the Tunnel Profile

To configure static tunnels, you create a tunnel profile that specifies the interface on the Nexus switch, the MAC address of the static peer, and the interface on the static peer.

Before you begin

To configure VXLAN static tunnels, the underlay must be completely configured and operating correctly.

SUMMARY STEPS

1. **tunnel-profile** *profile-name*
2. **encapsulation {***VXLAN* | *VXLAN-GPE* | *SRv6***}**
3. **source-interface loopback** *virtual-interface-number*
4. **route vrf** *tenant-vrf destination-host-prefix destination-vtep-ip-address* **next-hop-vrf** *destination-vtep-vrf* **vni** *vni-number* **dest-vtep-mac** *destination-vtep-mac-address*

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**tunnel-profile** *profile-name*
	Example:
	`switch(config)# tunnel-profile test`
	`switch(config-tnl-profile)#`
		Create and name the tunnel profile.

Step 2	**encapsulation {***VXLAN* | *VXLAN-GPE* | *SRv6***}**
	Example:
	`switch(config-tnl-profile)# encapsulation vxlan`
	`switch(config-tnl-profile)#`
		Set the appropriate encapsulation type for the tunnel profile.
		Note
		In NX-OS release 9.3(3), only encapsulation type **vxlan** is supported.

Step 3	**source-interface loopback** *virtual-interface-number*
	Example:
	`switch(config-tnl-profile)# source-interface loopback 1`
	`switch(config-tnl-profile)#`
		Configure the loopback interface as the source interface for the tunnel profile, where the virtual interface number is from 0 to 1023.

Step 4	**route vrf** *tenant-vrf destination-host-prefix destination-vtep-ip-address* **next-hop-vrf** *destination-vtep-vrf* **vni** *vni-number* **dest-vtep-mac** *destination-vtep-mac-address*
	Example:
	`switch(tunnel-profile)# route vrf cust1 101.1.1.2/32 7.7.7.1 next-hop-vrf default vni 20001 dest-vtep-mac f80f.6f43.036c`
	`switch(tunnel-profile)#`
		Create the tunnel route by specifying the destination software VTEP and entering the route information for the VNI and destination VTEP MAC address.
		Note
		The **route vrf** command accepts one *destination-vtep-mac-address* per *destination-vtep-ip-address* across all the routes. If you configure additional routes, they are cached as errored routes and a error syslog is generated for each.

Verifying VXLAN Static Tunnels

VXLAN static tunnels remain configured if one end of the tunnel goes down. While one end of the tunnel is down, packets are dropped because that VTEP is unreachable. When the down VTEP comes back online, traffic can resume across the tunnel after the underlay relearns connectivity.

You can use **show** commands to check the state of the tunnel profile and tunnel route.

Before you begin

SUMMARY STEPS

1. **show tunnel-profile**
2. **show ip route** *tenant-vrf-name*
3. **show running-config ofm**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**show tunnel-profile**	Shows information about the tunnel profile for the software.
Step 2	**show ip route** *tenant-vrf-name*	Shows route information for the VRF connecting to the software VTEP. For example, you can use this command when a `route unreachable` error occurs to verify that a route exists for a VRF's tunnel.
Step 3	**show running-config ofm**	Shows the running config for the OFM feature and static tunnels. You can use this command when a `route unreachable` error occurs to check whether the route information for the destination VTEP is present.

What to do next

In addition to VXLAN verification, you can use SPAN to check the ports and source VLANs for packets traversing the switch.

Example Configurations for VXLAN Static Tunnels

The following configuration examples shows VXLAN static tunnel configurations through the supported methods.

NX-OS CLI

```
vlan 2001
vlan 2001
  vn-segment 20001

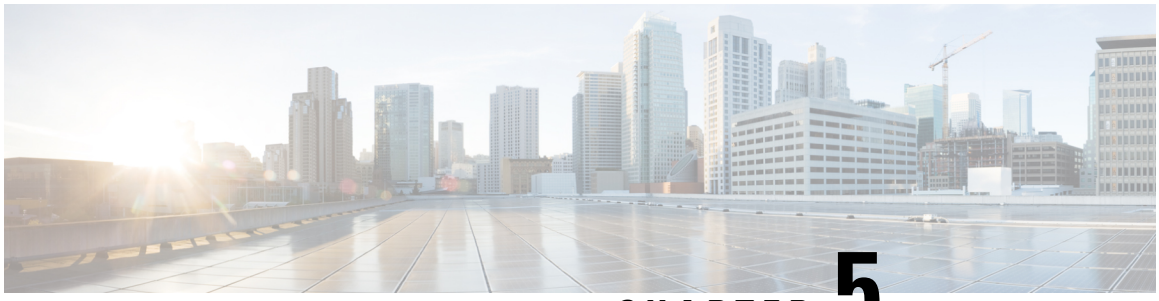
interface Vlan2001
  no shutdown
```

```
  vrf member cust1
  ip forward

vrf context cust1
  vni 20001


feature ofm

tunnel-profile test
  encapsulation vxlan
  source-interface loopback1
  route vrf cust1 101.1.1.2/32 7.7.7.1 next-hop-vrf default vni 20001 dest-vtep-mac
f80f.6f43.036c
```



CHAPTER 5

Configuring VXLAN with IPv6 in the Underlay (VXLANv6)

This chapter contains these sections:

- Information About Configuring VXLANv6 , on page 91
- Information About vPC and VXLAN with IPv6 in the Underlay (VXLANv6), on page 92
- Information About vPC Peer Keepalive and VXLAN with IPv6 in the Underlay (VXLANv6), on page 92
- Guidelines and Limitations for VXLAN with IPv6 in the Underlay (VXLANv6) , on page 93
- Configuring the VTEP IP Address, on page 96
- Configuring vPC for VXLAN with IPv6 in the Underlay (VXLANv6), on page 98
- Configuring VXLAN EVPN with IPv6 Ingress Replication, on page 99
- Configuring Multicast Underlay for BUM on L2-VNI, on page 101
- Example Configurations for VXLAN with IPv6 in the Underlay (VXLANv6), on page 102
- Verifying VXLAN with IPv6 in the Underlay (VXLANv6), on page 103

Information About Configuring VXLANv6

VXLAN BGP EVPN is deployed with IPv4 underlay and IPv4 VTEP. Hosts in the overlay can be IPv4 or IPv6. Support is added for VXLAN with IPv6 in the Underlay (VXLANv6) with an IPv6 VTEP. This requires IPv6 versions of the unicast routing protocols and utilizing ingress replication or multicast underlay for multi-destination traffic (BUM) in the underlay.

This solution is targeted for deployments where the VTEP is IPv6 only and the underlay is IPv6. The BGP sessions between the leaf and spine are also IPv6. The overlay hosts can be either IPv4 or IPv6.

VXLANv6 feature supports BGP unnumbered peering in the underlay.

The following protocols are supported in the underlay:

- IS-IS
- OSPFv3
- eBGP

Information About vPC and VXLAN with IPv6 in the Underlay (VXLANv6)

vPC VTEPs use vMAC (virtual MAC) with the VIP/PIP feature. vMAC is used with VIP and the system MAC is used with PIP.



Note Only unicast MAC address range is supported for vMAC address configuration.

In the IPv4 underlay, vMAC is derived from the IPv4 VIP address:

VMAC = 0x02 + 4 bytes IPv4 VIP address.

In the IPv6 underlay, VIP is IPv6 (128 bits) which cannot be used to generate a conflict free unique vMAC (48 bits). The default method is to autogenerate the vMAC by picking the last 32 bits from the IPv6 VIP:

Autogenerated vMAC = 0x06 + the last 4 bytes of the IPv6 VIP address.

If there are two vPC complexes which have different VIPs but the same last 4 bytes of IPv6 address in the VIP, both autogenerate the same vMAC. For a remote VTEP, it sees vMAC flapping between two different VIPs. This is not an issue for Cisco Nexus 9000 Series switches which support VXLAN IPv6.

For other vendor boxes, if this is an issue for interoperability reasons, the vMAC can be manually configured on Cisco Nexus 9000 Series switches to override the autogenerated vMAC. The default behavior for VXLAN with IPv6 in the Underlay (VXLANv6) is to autogenerate the VMAC. If a VMAC is configured manually, the manually configured VMAC takes precedence.

```
interface nve1
    virtual-rmac <48-bit mac address>
```

The VMAC must be managed by the administrator just like the VIP/PIP and must be unique in the fabric. All the preceding behavior is for VXLAN with IPv6 in the Underlay (VXLANv6) only and nothing changes about VMAC creation and advertisement for VXLAN IPv4 in the underlay.

The default behavior is that vMAC is autogenerated from the configured VIP and advertised. There is no need to use the **virtual-rmac** command as previously described except for interoperability cases. There is no need to use the existing **advertise virtual-rmac** command for VXLAN with IPv6 in the Underlay (VXLANv6).

Information About vPC Peer Keepalive and VXLAN with IPv6 in the Underlay (VXLANv6)

The modification for vPC is to allow IPv6 addresses to be used for the peer-keepalive link. The link can be on the management interface or any other interface. The keepalive link becomes operational only when both peers are configured correctly either with the IPv4 or IPv6 address and those addresses are reachable from each peer. Peer-keepalive can be configured on in-band and out-of-band interfaces.



Note peer-keepalive must be a global unicast address.

The configuration command for **peer-keepalive** accepts an IPv6 address

```
vpc domain 1
peer-keepalive destination 001:002::003:004 source 001:002::003:005 vrf management
```

Guidelines and Limitations for VXLAN with IPv6 in the Underlay (VXLANv6)

VXLAN with IPv6 in the Underlay (VXLANv6) has the following guidelines and limitations:

- Dual Stack (IPv4 and IPv6) is not supported for VXLAN underlay. It should either be IPv4 or IPv6, not both.
- NVE Source interface loopback for VTEP can either be IPv4 (VXLANv4) or IPv6 (VXLANv6), and not both.
- Next hop address in overlay (in bgp l2vpn evpn address family updates) should be resolved in underlay URIB to the same address family. For example, the use of VTEP (NVE source loopback) IPv4 addresses in fabric should only have BGP l2vpn evpn peering over IPv4 addresses.
- Usage of IPv6 LLA requires the TCAM Region for **ing-sup** to be re-carved from the default value of 512 to 768. This step requires a copy run start and reload

The following Cisco Nexus platforms are supported to provide the VTEP function (leaf and border). The BGP route reflector can be provided by any Cisco Nexus platform that supports the EVPN **address-family** command over an IPv6 MP-BGP peering.

- Cisco Nexus 9300-FX
- Cisco Nexus 9300-FX2
- Cisco Nexus 9300-FX3
- Cisco Nexus 9300-FXP
- Cisco Nexus 9300-GX
- Cisco Nexus 9300-GX2
- Cisco Nexus 9332D-H2R
- Cisco Nexus 93400LD-H1
- Cisco Nexus 9364C-H1

VXLAN with IPv6 in the Underlay (VXLANv6) supports the following features:

- Address Resolution Protocol (ARP) suppression in the overlay
- Access Control List (ACL) Quality of Service (QoS)
- Border Node with VRF-Lite
- Dynamic Host Configuration Protocol (DHCP)
- Guestshell support

- Internet Group Management Protocol (IGMP) Snooping in the overlay
- Virtual Extensible Local Area Network (VXLAN) Operation, Administration, and Maintenance (OAM)
- Storm Control for host ports (Access Side)
- Virtual Port Channel (vPC) with VIP and PIP support
- VXLAN Policy-Based Routing (PBR)
- vPC Fabric Peering
- Tenant Routed Multicast (TRM)
- VXLAN Multi-Site
- Downstream VNI
- Centralized Route Leak
- Overlay IGMP Snooping
- Protocol Independent Multicast (PIM) Any Source Multicast (ASM)
- VXLAN Access Features
  - Private VLAN (PVLAN)
  - 802.1x
  - Port security
  - Port VLAN translation
  - QinVNI
  - SelQinVNI
  - QinQ QinVNI

Beginning with Cisco NX-OS Release 9.3(1), VXLAN with IPv6 in the Underlay is supported on Cisco Nexus 9300-FX2, FX, and GX EOR switches and on the following modules:

- N9K-X97160YC-EX, N9K-X9736C-FX with N9K-C9508-FM-E2 fabric module

VXLAN with IPv6 in the Underlay (VXLANv6) does not support the following features:

- Bidirectional Forwarding Detection (BFD)
- Cisco Data Center Network Manager (DCNM) integration
- Cross Connect
- EVPN Multi-homing with Ethernet Segment (ES)
- Fabric Extender (FEX) attached to a VXLAN-enabled switch.
- VXLAN Flood and Learn
- Multiprotocol Label Switching (MPLS) and Locator/ID Separation Protocol (LISP) handoff
- Multicast underlay (PIM-BiDir, Snooping)

- NetFlow
- **peer vtep** command
- Sampled Flow (sFLOW)
- Static ingress replication (IR)
- Virtual Network Functions (VNF) Multipath

Beginning with Cisco NX-OS Release 10.1(1), IPv6 Underlay is supported for N9K-C9316D-GX, N9K-C93600CD-GX, and N9K-C9364C-GX TOR switches.

Beginning with Cisco NX-OS Release 10.2(3)F, IPv6 Underlay is supported on Cisco Nexus X97160YC-EX and 9700-FX/GX line cards.

Beginning with Cisco NX-OS Release 10.5(2)F, VXLAN with IPv6 in the Underlay is supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.

Beginning with Cisco NX-OS Release 10.3(2)F, vPC fabric peering with IPv6 underlay is supported on Cisco Nexus 9300-FX/FX2/FX3/GX/GX2 switches.

Beginning with Cisco NX-OS Release 10.4(1)F, vPC fabric peering with IPv6 underlay is supported on Cisco Nexus 9332D-H2R switches.

Beginning with Cisco NX-OS Release 10.4(2)F, vPC fabric peering with IPv6 underlay is supported on Cisco Nexus 93400LD-H1 switches.

Beginning with Cisco NX-OS Release 10.4(3)F, vPC fabric peering with IPv6 underlay is supported on Cisco Nexus 9364C-H1 switches.

Beginning with Cisco NX-OS Release 10.5(2)F, vPC fabric peering with IPv6 underlay is supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.

Beginning with Cisco NX-OS Release 10.2(3)F, the VTEP function (leaf and border) is supported on Cisco Nexus 9300-GX2 platform switches.

Beginning with Cisco NX-OS Release 10.4(1)F, the VTEP function (leaf and border) is supported on Cisco Nexus 9332D-H2R switches.

Beginning with Cisco NX-OS Release 10.4(2)F, the VTEP function (leaf and border) is supported on Cisco Nexus 93400LD-H1 switches.

Beginning with Cisco NX-OS Release 10.4(3)F, the VTEP function (leaf and border) is supported on Cisco Nexus 9364C-H1 switches.

Beginning with Cisco NX-OS Release 10.2(3)F, VXLAN PBR is supported with VXLAN v6 underlay on Cisco Nexus 9300-FX/FX2/FX3/GX/GX2 platforms ToR switches.

Beginning with Cisco NX-OS Release 10.4(1)F, VXLAN PBR is supported with VXLAN v6 underlay on Cisco Nexus 9332D-H2R switches.

Beginning with Cisco NX-OS Release 10.4(2)F, VXLAN PBR is supported with VXLAN v6 underlay on Cisco Nexus 93400LD-H1 switches.

Beginning with Cisco NX-OS Release 10.4(3)F, VXLAN PBR is supported with VXLAN v6 underlay on Cisco Nexus 9364C-H1 switches.

Beginning with Cisco NX-OS Release 10.2(3)F, IPv6 Underlay is supported on Cisco Nexus 9300-GX2 switches.

Beginning with Cisco NX-OS Release 10.4(1)F, IPv6 Underlay is supported on Cisco Nexus 9332D-H2R switches.

Beginning with Cisco NX-OS Release 10.4(2)F, IPv6 Underlay is supported on Cisco Nexus 93400LD-H1 switches.

Beginning with Cisco NX-OS Release 10.4(3)F, IPv6 Underlay is supported on Cisco Nexus 9364C-H1 switches.

The IPv6 Underlay is supported on the following features for VXLAN EVPN:

- Private VLAN (PVLAN) on Cisco Nexus 9300-FX/FX2/FX3/GX/GX2/H2R/H1, and Cisco Nexus 9500 switches with Nexus X97160YC-EX and 9700-FX/GX/FX3 line cards.
- 802.1x on Cisco Nexus 9300-FX/FX2/FX3/GX/GX2/H2R/H1, and Cisco Nexus 9500 switches with Nexus X97160YC-EX and 9700-FX/GX/FX3 line cards.
- Port security on Cisco Nexus 9300-FX/FX2/FX3/GX/GX2/H2R/H1, and Cisco Nexus 9500 switches with Nexus X97160YC-EX and 9700-FX/GX/FX3 line cards.
- Port VLAN translation on Cisco Nexus 9300-FX/FX2/FX3/GX/GX2/H2R/H1, and Cisco Nexus 9500 switches with Nexus X97160YC-EX and 9700-FX/GX/FX3 line cards.
- QinVNI on Cisco Nexus 9300-FX/FX2/FX3/GX/GX2/H2R/H1 platform switches.
- Selective QinVNI on Cisco Nexus 9300-FX/FX2/FX3/GX/GX2/H2R/H1 platform switches.
- QinQ-QinVNI on Cisco Nexus 9300-FX/FX2/FX3/GX/GX2/H2R/H1 platform switches.

Other guidelines and limitations:

- VXLAN/Fibre Channel co-existence

Configuring the VTEP IP Address

SUMMARY STEPS

1. **configure terminal**
2. **interface nve1**
3. **source-interface loopback** *src-if*
4. **exit**
5. **interface loopback** *loopback_number*
6. **ipv6 address** *ipv6_format*
7. **exit**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.
Step 2	**interface nve1** **Example:** `switch(config)# interface nve1`	Configure the NVE interface.
Step 3	**source-interface loopback** *src-if* **Example:** `switch(config-if-nve)# source interface loopback 1`	The source interface must be a loopback interface that is configured on the switch with a valid /128 IP address. This /128 IP address must be known by the intermediate devices in the transport network and the remote VTEPs. This is accomplished by advertising it through a dynamic routing protocol in the transport network. **Note** The IPv6 address on **loopback1** must be a /128 address. The VTEP IP address cannot be a link local IPv6 address.
Step 4	**exit** **Example:** `switch(config-if-nve)# exit`	Exit configuration mode.
Step 5	**interface loopback** *loopback_number* **Example:** `switch(config)# interface loopback 1`	Configure the loopback interface.
Step 6	**ipv6 address** *ipv6_format* **Example:** `switch(config-if)# ipv6 address 2001:db8:0:0:1:0:0:1/128`	Configure IPv6 address on the interface.
Step 7	**exit** **Example:** `switch(config-if)# exit`	Exit configuration mode.

Configuring vPC for VXLAN with IPv6 in the Underlay (VXLANv6)

VXLAN with IPv4 in the underlay leveraged the concept of a secondary IP address (VIP) used in vPC. IPv6 does not have the concept of secondary addresses as does IPv4. However, multiple IPv6 global addresses can be configured on an interface, which are treated equally in priority.

The CLI for the VIP configuration has been extended to specify the loopback interface that carries the VIP if there is a VXLAN with IPv6 in the Underlay (VXLANv6) vPC. The IPv6 primary IP address (PIP) and VIP are in two separate loopback interfaces.

Similar to IPv4, if there are multiple IPv6 addresses specified on either loopback, the lowest IP is selected for each.

The following steps outline the configuration of a VTEP IP (VIP/PIP) required on a vPC setup.



Note MVPN VRI ID must be configured for TRM in a vPC. This same VRI id must be configured on both vPC nodes that are part of the same vPC complex. However, each VRI ID must be unique within the network. This implies that two different vPC pairs must have distinct VRI ID configurations to ensure correct routing and avoid any conflicts.

The **anycast loopback** command is used only for VXLAN with IPv6 in the Underlay (VXLANv6).

SUMMARY STEPS

1. **configure terminal**
2. **interface nve1**
3. **source-interface loopback** *src-if* **anycast loopback** *any-if*
4. **exit**
5. **interface loopback** *loopback_number*
6. **ipv6 address** *ipv6_format*
7. **exit**
8. **interface loopback** *loopback_number*
9. **ipv6 address** *ipv6_format*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.
Step 2	**interface nve1** **Example:** `switch(config)# interface nve1`	Configure the NVE interface.

	Command or Action	Purpose
Step 3	**source-interface loopback** *src-if* **anycast loopback** *any-if* **Example:** `switch(config-if-nve)# source interface loopback 1 anycast loopback 2`	The source interface must be a loopback interface that is configured on the switch with a valid /128 IP address. This /128 IP address must be known by the transient devices in the transport network and the remote VTEPs. This is accomplished by advertising it through a dynamic routing protocol in the transport network. **Note** The IPv6 address on loopback1, the primary IP address (PIP), and loopback2, the secondary IP address ()VIP), must be a /128 address. The VTEP IP address cannot be a link local IPv6 address.
Step 4	**exit** **Example:** `switch(config-if-nve)# exit`	Exit configuration mode.
Step 5	**interface loopback** *loopback_number* **Example:** `switch(config)# interface loopback 1`	Configure the loopback interface.
Step 6	**ipv6 address** *ipv6_format* **Example:** `switch(config-if)# ipv6 address 2001:db8:0:0:1:0:0:1/128`	Configure IPv6 address on the interface.
Step 7	**exit** **Example:** `switch(config-if)# exit`	Exit configuration mode.
Step 8	**interface loopback** *loopback_number* **Example:** `switch(config)# interface loopback 2`	Configure the loopback interface.
Step 9	**ipv6 address** *ipv6_format* **Example:** `switch(config-inf)# ipv6 address 2001:db8:0:0:1:0:0:2/128`	Configure IPv6 address on the interface.

Configuring VXLAN EVPN with IPv6 Ingress Replication

Configuring IPv6 Ingress Replication in the VXLAN fabric involves the following configuration:

Configuring Ingress Replication for L2-VNI

Under NVE configuration on a leaf switch, perform the following steps to configure the IPv6 Ingress Replication (IR) for each L2-VNI (VLAN).

SUMMARY STEPS

1. **configure terminal**
2. **interface nve1**
3. **member vni** *vni*
4. **ingress-replication protocol bgp**
5. **global ingress-replication protocol bgp**
6. **exit**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**interface nve1** **Example:** `switch(config)# interface nve1`	Configures the NVE interface.
Step 3	**member vni** *vni* **Example:** `switch(config-if-nve)# member vni 10501`	Configures the Layer 2 virtual network identifier.
Step 4	**ingress-replication protocol bgp** **Example:** `switch(config-if-nve-vni)# ingress-replication protocol bgp`	Enables BGP EVPN with ingress replication for the VNI within a fabric.
Step 5	**global ingress-replication protocol bgp** **Example:** `switch(config-if-nve)# global ingress-replication protocol bgp`	Configures the global ingress replication for the Layer 2 VNI.
Step 6	**exit** **Example:** `switch(config-if-nve)# exit`	Exits configuration mode.

Configuring Multicast Underlay for BUM on L2-VNI

Under NVE configuration on a leaf switch, perform the following steps to configure the IPv6 Multicast Underlay for BUM on each L2-VNI (VLAN).

Before you begin

Ensure that the PIM feature is enabled on the device.

Ensure that the PIM RP is configured.

Ensure that the IP PIM sparse mode is configured on the interfaces (including loopback).

For more information on PIM configuration, see Configuring a VXLAN VTEP in vPC, on page 73.

SUMMARY STEPS

1. **configure terminal**
2. **interface nve1**
3. **member vni** *vni*
4. **mcast-group** *start-address [end-address]*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**interface nve1** **Example:** `switch(config)# interface nve1`	Configures the NVE interface.

Step 3	**member vni** *vni*
	Example:
	`switch(config-if-nve)# member vni 10701-10900`
	`switch(config-if-nve-vni)#`
		Configures the Layer 2 virtual network identifier for the specified range (10701-10900).

Step 4	**mcast-group** *start-address [end-address]* **Example:** `switch(config-if-nve-vni)# mcast-group ff03::101`	Assigns a multicast group to the VNIs.

Example Configurations for VXLAN with IPv6 in the Underlay (VXLANv6)

The following are configuration examples for VXLAN with IPv6 in the Underlay (VXLANv6):

With IPv6 address set/match in next-hop, BGP must set/match the IPv6 next-hop address in route type-2 (MAC-IP) and route type-5 (IP Prefix).

Under route-map:

```
set ipv6 next-hop <vtep address>
match ipv6 next-hop  <vtep address>
```

BGP Underlay



Note BGP IPv6 neighbor must support L2VPN EVPN address-family session.



Note The router ID in VXLAN with IPv6 in the Underlay (VXLANv6) must be an IPv4 address.

The BGP router ID is a 32-bit value that is often represented by an IPv4 address. By default, Cisco NX-OS sets the router ID to the IPv4 address of a loopback interface on the router. For VXLAN with IPv6 in the Underlay (VXLANv6), none of the loopbacks need to have an IPv4 address in which case the default selection of router ID does not happen correctly. You can configure the router ID manually to an IPv4 address.

BGP RD (Route distinguisher) which is 64 bits in length can be configured using the autonomous system number of the 4-byte IP address. For VXLAN with IPv6 in the Underlay (VXLANv6), when using an IP address for configuring RD, you must use IPv4 as in the case of VXLAN IPv4.

```
feature bgp
nv overlay evpn

router bgp 64496
   ! IPv4 router id
   router-id  35.35.35.35
   ! Redistribute the igp/bgp routes
   address-family ipv6 unicast
     redistribute direct route-map allow

   ! For IPv6 session, directly connected peer interface
   neighbor 2001:DB8:0:1::55
     remote-as 64496
     address-family ipv6 unicast
```

OSPFv3 Underlay

```
feature ospfv3
```

```
router ospfv3 201
router-id 290.0.2.1

interface ethernet 1/2
ipv6 address 2001:0DB8::1/48
ipv6 ospfv3 201 area 0.0.0.10
```

IS-IS Underlay

```
router isis Enterprise
is-type level-1
net 49.0001.0000.0000.0003.00

interface ethernet 2/1
ipv6 address 2001:0DB8::1/48
isis circuit-type level-1
ipv6 router isis Enterprise
```

Verifying VXLAN with IPv6 in the Underlay (VXLANv6)

To display the status for the VXLAN with IPv6 in the Underlay (VXLANv6) configuration, enter one of the following commands:

Table 4: VXLAN with IPv6 in the Underlay (VXLANv6) Verification Commands

Command	Purpose
show running-config interface nve 1	Displays interface NVE 1 running configuration information.
show nve interface 1 detail	Displays NVE interface detail.
show nve peers	Displays the peering time and VNI information for VTEP peers.
show nve vni ingress-replication	Displays NVE VNI ingress replication information.
show nve peers 2018:1015::abcd:1234:3 int nv1 counters	Displays NVE peers counter information.
show bgp l2vpn evpn 1012.0383.9600	Displays BGP L2VPN information for route type 2.
show bgp l2vpn evpn 303:304::1	Displays BGP L2VPN EVPN for route type 3.
show bgp l2vpn evpn 5.116.204.0	Displays BGP L2VPN EVPN for route type 5.
show l2route peerid	Displays L2route peerid.
show l2route topology detail	Displays L2route topology detail.
show l2route evpn imet all detail	Displays L2route EVPN imet detail.
show l2route fl all	Display L2route flood list detail.

Command	Purpose
show l2route mac all detail	Displays L2route MAC detail.
show l2route mac-ip all detail	Displays MAC address and host IP address.
show ip route 1.191.1.0 vrf vxlan-10101	Displays route table for VRF.
show forwarding ipv4 route 1.191.1.0 detail vrf vxlan-10101	Displays forwarding information.
show ipv6 route vrf vxlan-10101	Displays IPv6 routing table.
show bgp l2vpn evpn	Displays BGP's updated routes.
show bgp evi *evi-id*	Displays BGP EVI information.
show forwarding distribution peer-id	Displays forwarding information.
show forwarding nve l2 ingress-replication-peers	Displays forwarding information for ingress replication.
show forwarding nve l3 peers	Displays nv3 Layer 3 peers information.
show forwarding ecmp platform	Displays forwarding ECMP platform information.
show forwarding ecmp platform	Displays forwarding ECMP platform information.
show forwarding nve l3 ecmp	Displays forwarding NVE Layer 3 ECMP information.

Example of the **show running-config interface nve 1**

Command

```
switch# show running-config interface nve 1
interface nve1
  no shutdown
  source-interface loopback1 anycast loopback2
  host-reachability protocol bgp
  member vni 10011
    ingress-replication protocol bgp
  member vni 20011 associate-vrf
```

Example of the **show nve interface 1 detail**

Command

```
switch# show nve interface nve 1 detail
Interface: nve1, State: Up, encapsulation: VXLAN
 VPC Capability: VPC-VIP-Only [notified]
 Local Router MAC: a093.51cf.78f7
 Host Learning Mode: Control-Plane
 Source-Interface: loopback1 (primary: 30:3:1::2)
 Anycast-Interface: loopback2 (secondary: 303:304::1)
 Source Interface State: Up
 Anycast Interface State: Up
 Virtual RMAC Advertisement: Yes
 NVE Flags:
 Interface Handle: 0x49000001
```

```
 Source Interface hold-down-time: 745
 Source Interface hold-up-time: 30
 Remaining hold-down time: 0 seconds
 Virtual Router MAC: 0600.0000.0001
 Interface state: nve-intf-add-complete
```

Example of the **show nve peers** Command

```
switch# show nve peers
Interface Peer-IP          State LearnType Uptime   Router-Mac
--------- ---------------  ----- --------- -------- -----------------
nve1      1:1::1:1          Up    CP        00:44:09   5087.89d4.6bb7
```

Up

Example of the **show nve vni ingress-replication**

Command

```
switch# show nve vni ingress-replication
Interface VNI      Replication List  Source  Up Time
--------- -------- ----------------- ------- -------

nve1      10011    1:1::1:1           BGP-IMET   00:46:55
```

Example of the **show nve peers** *ipv6-address* **int nv1 counters** Command .

```
switch# show nve peers 2018:2015::abcd:1234:3 int nve 1 counters
Peer IP: 2018:1015::abcd:1234:3
TX
        0 unicast packets 0 unicast bytes
        0 multicast packets 0 multicast bytes
RX
        0 unicast packets 0 unicast bytes
        0 multicast packets 0 multicast bytes
```

Example of the **show bgp l2vpn evpn** Command for Route-Type 2.

```
switch# show bgp l2vpn evpn 1012.0383.9600
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 30.3.1.1:34067    (L2VNI 2001300)
BGP routing table entry for [2]:[0]:[0]:[48]:[1012.0383.9600]:[0]:[0.0.0.0]/216, version
1051240
Paths: (1 available, best #1)
Flags: (0x000102) (high32 00000000) on xmit-list, is not in l2rib/evpn
Multipath: iBGP

  Advertised path-id 1
  Path type: local, path is valid, is best path, no labeled nexthop
  AS-Path: NONE, path locally originated
    303:304::1 (metric 0) from 0:: (30.3.1.1)
      Origin IGP, MED not set, localpref 100, weight 32768
      Received label 2001300
      Extcommunity: RT:2:2001300 ENCAP:8

  Path-id 1 advertised to peers:
    2::21              2::66
BGP routing table entry for [2]:[0]:[0]:[48]:[1012.0383.9600]:[32]:[4.231.115.2]/272, version
 1053100
Paths: (1 available, best #1)
Flags: (0x000102) (high32 00000000) on xmit-list, is not in l2rib/evpn
Multipath: iBGP
```

```
  Advertised path-id 1
  Path type: local, path is valid, is best path, no labeled nexthop
  AS-Path: NONE, path locally originated
    303:304::1 (metric 0) from 0:: (30.3.1.1)
      Origin IGP, MED not set, localpref 100, weight 32768
      Received label 2001300 3003901
      Extcommunity: RT:2:2001300 RT:2:3003901 ENCAP:8 Router MAC:0600.0000.0001

  Path-id 1 advertised to peers:
    2::21              2::66
```

Example of the **show bgp l2vpn evpn** Command for Route-Type 3

```
switch# show bgp l2vpn evpn 303:304::1
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 30.3.1.1:32769    (L2VNI 2000002)
BGP routing table entry for [3]:[0]:[128]:[303:304::1]/184, version 1045060
Paths: (1 available, best #1)
Flags: (0x000002) (high32 00000000) on xmit-list, is not in l2rib/evpn
Multipath: iBGP

  Advertised path-id 1
  Path type: local, path is valid, is best path, no labeled nexthop
  AS-Path: NONE, path locally originated
    303:304::1 (metric 0) from 0:: (30.3.1.1)
      Origin IGP, MED not set, localpref 100, weight 32768
      Extcommunity: RT:2:2000002 ENCAP:8
      PMSI Tunnel Attribute:
        flags: 0x00, Tunnel type: Ingress Replication
        Label: 2000002, Tunnel Id: 303:304::1

  Path-id 1 advertised to peers:
    2::21              2::66
```

Example of the **show bgp l2vpn evpn** Command for Route-Type 5

```
switch# show bgp l2vpn evpn 5.116.204.0
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 2.0.0.52:302
BGP routing table entry for [5]:[0]:[0]:[24]:[5.116.204.0]/224, version 119983
Paths: (2 available, best #2)
Flags: (0x000002) (high32 00000000) on xmit-list, is not in l2rib/evpn, is not in HW
Multipath: iBGP

  Path type: internal, path is valid, not best reason: Neighbor Address, no labeled nexthop

  Gateway IP: 0.0.0.0
  AS-Path: 65001 5300 , path sourced external to AS
    3::52 (metric 200) from 2::66 (2.0.0.66)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 3003301
      Extcommunity: RT:2:3003301 ENCAP:8 Router MAC:f80b.cb53.4897
      Originator: 2.0.0.52 Cluster list: 2.0.0.66

  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop
             Imported to 2 destination(s)
             Imported paths list: evpn-tenant-0301 default
  Gateway IP: 0.0.0.0
  AS-Path: 65001 5300 , path sourced external to AS
    3::52 (metric 200) from 2::21 (2.0.0.21)
      Origin IGP, MED not set, localpref 100, weight 0
```

```
      Received label 3003301
      Extcommunity: RT:2:3003301 ENCAP:8 Router MAC:f80b.cb53.4897
      Originator: 2.0.0.52 Cluster list: 2.0.0.21

  Path-id 1 not advertised to any peer
```

Example of the **show l2route peerid** Command

```
switch# show l2route peerid
NVE Ifhdl    IP Address                                PeerID       Ifindex      Num of MAC's
 Num of NH's
------------ ---------------------------------------- ------------ ------------ ------------
 ------------
1224736769   4999:1::1:1:1                             4            1191182340   23377
  0
```

Example of the **show l2route topology detail** Command

```
switch# show l2route topology detail
Flags:(L2cp)=L2 Ctrl Plane; (Dp)=Data Plane; (Imet)=Data Plane BGP IMET; (L3cp)=L3 Ctrl
Plane; (Bfd)=BFD over Vxlan; (Bgp)=BGP EVPN; (Of)=Open Flow mode; (Mix)=Open Flow IR mixed
 mode; (Acst)=Anycast GW on spine;
Topology ID   Topology Name   Attributes
-----------   -------------   ----------
101           Vxlan-10101     VNI: 10101
                              Encap:1 IOD:0 IfHdl:1224736769
                              VTEP IP: 5001:1::1:1:7
                              Emulated IP: ::
                              Emulated RO IP: 0.0.0.0
                              TX-ID: 2004 (Rcvd Ack: 0)
                              RMAC: 00fe.c83e.84a7, VRFID: 3
                              VMAC: 00fe.c83e.84a7
                              VMAC RO: 0000.0000.0000
                              Flags: L3cp, Sub_Flags: --, Prev_Flags: -
```

Example of the **show l2route evpn imet all detail** Command

```
switch# show l2route evpn imet all detail
Flags- (F): Originated From Fabric, (W): Originated from WAN

Topology ID  VNI   Prod  IP Addr        Eth Tag PMSI-Flags Flags   Type Label(VNI)  Tunnel
 ID     NFN Bitmap
-----------  ----- ----- -------------- ------- ---------- ------- ---- ------------
------------- -----------
901          10901 BGP   4999:1::1:1:1  0       0          -       6    10901
4999:1::1:1:1
```

Example of the **show l2route fl all** Command

```
switch# show l2route fl all
Topology ID Peer-id     Flood List                              Service Node
----------- ----------- --------------------------------------- ------------
901         4           4999:1::1:1:1                           no
```

Example of the **show l2route mac all detail** Command

Beginning with Cisco NX-OS Release 10.5(3)F, the labeled nexthop and an asymmetric VNI flag is added as shown. For symmetric VNI, labels and flags will not be shown as part of nexthops for EAD and PL.

```
switch# show l2route mac all detail
```

```
Flags -(Rmac):Router MAC (Stt):Static (L):Local (R):Remote (V):vPC link
(Dup):Duplicate (Spl):Split (Rcv):Recv (AD):Auto-Delete (D):Del Pending
(S):Stale (C):Clear, (Ps):Peer Sync (O):Re-Originated (Nho):NH-Override
(Pf):Permanently-Frozen, (Orp): Orphan

Topology    Mac Address    Prod   Flags         Seq No     Next-Hops

----------- -------------- ------ ------------- ----------
--------------------------------------
100         abcd.ef12.3456 BGP    Rcv              0         5.5.5.5 (Label: 20000)(Flags:
 Asy)
                                                            6.6.6.6 (Label: 10000)(Flags:
 Asy)
            Route Resolution Type: ESI
            Forwarding State: Resolved (PL)
            Resultant PL: 5.5.5.5 (Label: 20000)(Flags: Asy)
                          6.6.6.6 (Label: 10000)(Flags: Asy)
            Sent To: L2FM
            ESI : aaaa.aaaa.aaaa.aaaa.99aa
```

Example of the **show l2route mac-ip all detail** Command

Beginning with Cisco NX-OS Release 10.5(3)F, the labeled nexthop and an asymmetric VNI flag is added as shown. For symmetric VNI, labels and flags will not be shown as part of nexthops for EAD and PL.

```
switch# show l2route mac-ip all detail
Flags -(Rmac):Router MAC (Stt):Static (L):Local (R):Remote
(Dup):Duplicate (Spl):Split (Rcv):Recv(D):Del Pending (S):Stale (C):Clear
(Ps):Peer Sync (Ro):Re-Originated (Orp):Orphan (Asy):Asymmetric (Gw):Gateway
(Bh):Blackhole
(Piporp): Directly connected Orphan to PIP based vPC BGW
(Pipporp): Orphan connected to peer of PIP based vPC BGW
Topology    Mac Address    Host IP                                     Prod   Flags
   Seq No     Next-Hops
----------- -------------- ------------------------------------------ ------ -----------------
 ---------- ------------------------------------
100         abcd.ef12.3455 5.5.5.5                                     BGP    PS,
   2         1.1.1.10 (Label: 10000)(Flags: Asy)
            ESI : 0300.0000.0000.6400.0319
```

Example of the **show ip route 1.191.1.0 vrf vxlan-10101** Command

```
switch# show ip route 1.191.1.0 vrf vxlan-10101
IP Route Table for VRF "vxlan-10101"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

1.191.1.0/29, ubest/mbest: 6/0
    *via fe80::2fe:c8ff:fe09:8fff%default, Po1001, [200/0], 00:56:21, bgp-4002, internal,
tag 4007 (evpn)
segid: 10101 VTEP:(5001:1::1:1:1, underlay_vrf: 1) encap: VXLAN

*via fe80::2fe:c8ff:fe09:8fff%default, Po1002, [200/0], 00:56:21, bgp-4002, internal, tag
4007 (evpn)
segid: 10101 VTEP:(5001:1::1:1:1, underlay_vrf: 1) encap: VXLAN

    *via fe80::2fe:c8ff:fe09:8fff%default, Po1001, [200/0], 00:56:32, bgp-4002, internal,
tag 4007 (evpn)
segid: 10101 VTEP:(5001:1::1:1:2, underlay_vrf: 1) encap: VXLAN

    *via fe80::2fe:c8ff:fe09:8fff%default, Po1002, [200/0], 00:56:32, bgp-4002, internal,
tag 4007 (evpn)
```

```
segid: 10101 VTEP:(5001:1::1:1:2, underlay_vrf: 1) encap: VXLAN
```

Example of the **show forwarding ipv4 route 1.191.1.0 detail vrf vxlan-10101** Command

```
switch# show forwarding ipv4 route 1.191.1.0 detail vrf vxlan-10101

slot  1
=======
Prefix 1.191.1.0/29, No of paths: 2, Update time: Mon Apr 15 15:38:17 2019

   5001:1::1:1:1      nve1
   5001:1::1:1:2      nve1
```

Example of the **show ipv6 route vrf vxlan-10101** Command

```
switch# show ipv6 route vrf vxlan-10101
IPv6 Routing Table for VRF "vxlan-10101"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]

2:2:2::101/128, ubest/mbest: 1/0
    *via 5001:1::1:1:1/128%default, [200/0], 00:55:31, bgp-4002, internal, tag 4002 (evpn)
 segid 10101
VTEP:(5001:1::1:1:1, underlay_vrf: 1) encap: VXLAN
```

Example of the **show forwarding distribution peer-id**

Command

```
switch# show forwarding distribution peer-id
UFDM Peer-id allocations: App id 0
App: VXLAN   Vlan: 1     Id: 4999:1::1:1:1 0x49030001 Peer-id: 0x6
App: VXLAN   Vlan: 1     Id: 5001:1::1:1:1 0x49030001 Peer-id: 0x2
App: VXLAN   Vlan: 1     Id: 5001:1::1:1:2 0x49030001 Peer-id: 0x1
App: VXLAN   Vlan: 1     Id: 5001:1::1:1:7 0x49030001 Peer-id: 0x7
App: VXLAN   Vlan: 1     Id: 5001:1::1:2:101 0x49030001 Peer-id: 0x8
App: VXLAN   Vlan: 1     Id: 5001:1::1:2:102 0x49030001 Peer-id: 0x5
App: VXLAN   Vlan: 1     Id: 5001:1::1:2:103 0x49030001 Peer-id: 0x9
App: VXLAN   Vlan: 1     Id: 5001:1::1:2:104 0x49030001 Peer-id: 0xa
App: VXLAN   Vlan: 1     Id: 5001:1::1:2:105 0x49030001 Peer-id: 0xb
App: VXLAN   Vlan: 1     Id: 5001:1::1:2:106 0x49030001 Peer-id: 0xc
App: VXLAN   Vlan: 1     Id: 5001:1::1:2:107 0x49030001 Peer-id: 0xd
```

Example of the **show forwarding nve l2 ingress-replication-peers**

Command

```
switch# show forwarding nve l2 ingress-replication-peers
slot  1
=======

Total count of VLANS with ingr-repl peers: 1950
VLAN 1024 VNI 0 Vtep Ifindex 0x0 plt_space : 0x1ca75e14
         peer :  6002:1::1:1:1
         peer :  5001:1::1:1:7
         peer :  4999:1::1:1:1

PSS VLAN:1024, VNI:0, vtep:0x0x0, peer_cnt:3
         peer :  6002:1::1:1:1  marked : 0
         peer :  5001:1::1:1:7  marked : 0
         peer :  4999:1::1:1:1  marked : 0
```

```
 VLAN 1280 VNI 0 Vtep Ifindex 0x0 plt_space : 0x1ca75e14
         peer :  6002:1::1:1:1
         peer :  5001:1::1:1:7
         peer :  4999:1::1:1:1

PSS VLAN:1280, VNI:0, vtep:0x0x0, peer_cnt:3
         peer :  6002:1::1:1:1  marked : 0
         peer :  5001:1::1:1:7  marked : 0
         peer :  4999:1::1:1:1  marked : 0
```

Example of the **show forwarding nve l3 peers**

Command

```
switch# show forwarding nve l3 peers
slot  1
=======

EVPN configuration state: disabled, PeerVni Adj enabled
NVE cleanup transaction-id 0
tunnel_id    Peer_id  Peer_address   Interface       rmac         origin state del count
-------------------------------------------------------------------------------------
0x0        1225261062 4999:1::1:1:1        nve1      0600.0001.0001 URIB       merge-done
no    100
0x0        1225261058 5001:1::1:1:1        nve1      2cd0.2d51.9f1b NVE        merge-done
no    100
0x0        1225261057 5001:1::1:1:2        nve1      00a6.cab6.bbbb NVE        merge-done
no    100
0x0        1225261063 5001:1::1:1:7        nve1      00fe.c83e.84a7 URIB       merge-done
no    100
0x0        1225261064 5001:1::1:2:101      nve1      0000.5500.0001 URIB       merge-done
no    100
0x0        1225261061 5001:1::1:2:102      nve1      0000.5500.0002 URIB       merge-done
no    100
0x0        1225261065 5001:1::1:2:103      nve1      0000.5500.0003 URIB       merge-done
no    100
0x0        1225261066 5001:1::1:2:104      nve1      0000.5500.0004 URIB       merge-done
no    100
0x0        1225261067 5001:1::1:2:105      nve1      0000.5500.0005 URIB       merge-done
no    100
```

Example of the **show forwarding ecmp platform**

Command

```
switch# show forwarding ecmp platform
slot  1
=======



ECMP Hash: 0x198b8aae, Num Paths: 2, Hw index: 0x17532
Partial Install: No
Hw ecmp-index: unit-0:1073741827 unit-1:0 unit-2:0, cmn-index: 95538
Hw NVE ecmp-index: unit-0:0 unit-1:0 unit-2:0, cmn-index: 95538
Refcount: 134, Holder: 0x0,   Intf: Ethernet1/101, Nex-Hop: fe80:7::1:2
    Hw adj: unit-0:851977 unit-1:0 unit-2:0, cmn-index: 500010 LIF:4211
  Intf: Ethernet1/108, Nex-Hop: fe80:8::1:2
    Hw adj: unit-0:851978 unit-1:0 unit-2:0, cmn-index: 500012 LIF:4218
   VOBJ count: 0, VxLAN VOBJ count: 0, VxLAN: 0


ECMP Hash: 0x2bb2905e, Num Paths: 3, Hw index: 0x17533
```

```
Partial Install: No
Hw ecmp-index: unit-0:1073741828 unit-1:0 unit-2:0, cmn-index: 95539
Hw NVE ecmp-index: unit-0:0 unit-1:0 unit-2:0, cmn-index: 95539
Refcount: 16, Holder: 0x0,   Intf: Ethernet1/101, Nex-Hop: fe80:7::1:2
    Hw adj: unit-0:851977 unit-1:0 unit-2:0, cmn-index: 500010 LIF:4211
  Intf: Ethernet1/108, Nex-Hop: fe80:8::1:2
    Hw adj: unit-0:851978 unit-1:0 unit-2:0, cmn-index: 500012 LIF:4218
  Intf: port-channel1003, Nex-Hop: fe80:9::1:2
    Hw adj: unit-0:851976 unit-1:0 unit-2:0, cmn-index: 500011 LIF:4106
   VOBJ count: 0, VxLAN VOBJ count: 0, VxLAN: 0
```

Example of the **show forwarding ecmp recursive**

Command

```
switch# show forwarding ecmp recursive
slot  1
=======

Virtual Object 17 (vxlan):
    Hw vobj-index (0): unit-0:851976 unit-1:0 unit-2:0, cmn-index: 99016
    Hw NVE vobj-index (0): unit-0:0 unit-1:0 unit-2:0, cmn-index: 99016
    Hw vobj-index (1): unit-0:0 unit-1:0 unit-2:0, cmn-index: 0
    Hw NVE vobj-index (1): unit-0:0 unit-1:0 unit-2:0 cmn-index: 0
    Num prefixes : 1
Partial Install: No
    Active paths:
        Recursive NH 5001:1::1:2:10a/128 , table 0x80000001
    CNHs:
        fe80:9::1:2, port-channel1003
        Hw adj: unit-0:851976 unit-1:0 unit-2:0, cmn-index: 500011, LIF:4106
        Hw NVE adj: unit-0:0 unit-1:0 unit-2:0, cmn-index: 500011, LIF:4106
    Hw instance new : (0x182c8, 99016) ls count new 1
    FEC: fec_type 0
        VOBJ Refcount : 1
Virtual Object 167 (vxlan): ECMP-idx1:0x17536(95542), ECMP-idx2:0x0(0),
    Hw vobj-index (0): unit-0:1073741832 unit-1:0 unit-2:0, cmn-index: 99166
    Hw NVE vobj-index (0): unit-0:3 unit-1:0 unit-2:0, cmn-index: 99166
    Hw vobj-index (1): unit-0:0 unit-1:0 unit-2:0, cmn-index: 0
    Hw NVE vobj-index (1): unit-0:0 unit-1:0 unit-2:0 cmn-index: 0
    Num prefixes : 1
Partial Install: No
    Active paths:
        Recursive NH 5001:1::1:3:125/128 , table 0x80000001
    CNHs:
        fe80:7::1:2, Ethernet1/101
        Hw adj: unit-0:851977 unit-1:0 unit-2:0, cmn-index: 500010, LIF:4211
        Hw NVE adj: unit-0:0 unit-1:0 unit-2:0, cmn-index: 500010, LIF:4211
        fe80:8::1:2, Ethernet1/108
        Hw adj: unit-0:851978 unit-1:0 unit-2:0, cmn-index: 500012, LIF:4218
        Hw NVE adj: unit-0:0 unit-1:0 unit-2:0, cmn-index: 500012, LIF:4218
    Hw instance new : (0x1835e, 99166) ls count new 2
    FEC: fec_type 0
        VOBJ Refcount : 1
```

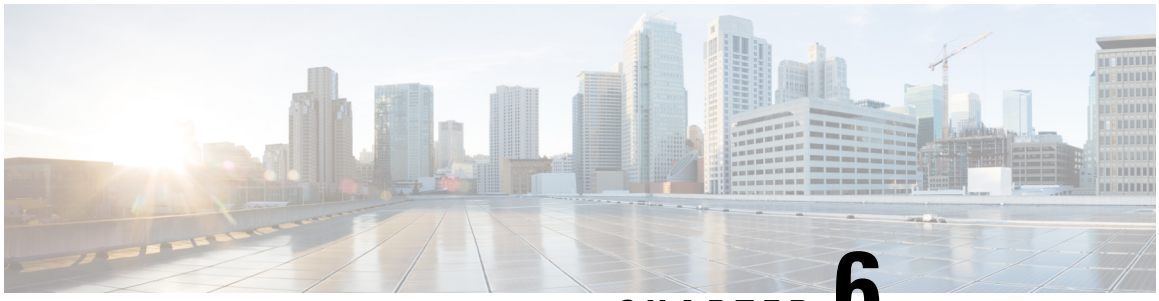
Example of the **show forwarding nve l3 ecmp**

Command

```
switch# show forwarding nve l3 ecmp
slot  1
=======
```

```
ECMP Hash: 0x70a50e4, Num Paths: 2, Hw Index: 0x17534
table_id: 403, flags: 0x0, adj_flags: 0x0, Ref-ct: 101
  tunnel_id: 5001:1::1:1:1, segment_id: 10101
  tunnel_id: 5001:1::1:1:2, segment_id: 10101
Hw ecmp-index: unit0: 1073741830 unit1: 0 unit2: 0

ECMP Hash: 0x1189f35e, Num Paths: 2, Hw Index: 0x17535
table_id: -2147483245, flags: 0x0, adj_flags: 0x0, Ref-ct: 50
  tunnel_id: 5001:1::1:1:1, segment_id: 10101
  tunnel_id: 5001:1::1:1:2, segment_id: 10101
Hw ecmp-index: unit0: 1073741831 unit1: 0 unit2: 0
```



CHAPTER 6

Configuring VXLAN BGP EVPN

This chapter contains these sections:

- About VXLAN BGP EVPN, on page 113
- Guidelines and Limitations for VXLAN BGP EVPN, on page 115
- About VXLAN EVPN with Downstream VNI, on page 121
- Guidelines and Limitations for VXLAN EVPN with Downstream VNI, on page 123
- Configuring VXLAN BGP EVPN, on page 126
- Configuring ND Suppression, on page 177
- VXLAN BGP EVPN with RFC 5549 Underlay, on page 183

About VXLAN BGP EVPN

About RD Auto

The auto-derived Route Distinguisher (rd auto) is based on the Type 1 encoding format as described in IETF RFC 4364 section 4.2 https://tools.ietf.org/html/rfc4364#section-4.2. The Type 1 encoding allows a 4-byte administrative field and a 2-byte numbering field. Within Cisco NX-OS, the auto derived RD is constructed with the IP address of the BGP Router ID as the 4-byte administrative field (RID) and the internal VRF identifier for the 2-byte numbering field (VRF ID).

The 2-byte numbering field is always derived from the VRF, but results in a different numbering scheme depending on its use for the IP-VRF or the MAC-VRF:

- The 2-byte numbering field for the IP-VRF uses the internal VRF ID starting at 1 and increments. VRF IDs 1 and 2 are reserved for the default VRF and the management VRF respectively. The first custom defined IP VRF uses VRF ID 3.
- The 2-byte numbering field for the MAC-VRF uses the VLAN ID + 32767, which results in 32768 for VLAN ID 1 and incrementing.

Example auto-derived Route Distinguisher (RD)

- IP-VRF with BGP Router ID 192.0.2.1 and VRF ID 6 - RD 192.0.2.1:6
- MAC-VRF with BGP Router ID 192.0.2.1 and VLAN 20 - RD 192.0.2.1:32787

About Route-Target Auto

The auto-derived Route-Target (route-target import/export/both auto) is based on the Type 0 encoding format as described in IETF RFC 4364 section 4.2 (https://tools.ietf.org/html/rfc4364#section-4.2). IETF RFC 4364 section 4.2 describes the Route Distinguisher format and IETF RFC 4364 section 4.3.1 refers that it is desirable to use a similar format for the Route-Targets. The Type 0 encoding allows a 2-byte administrative field and a 4-byte numbering field. Within Cisco NX-OS, the auto derived Route-Target is constructed with the Autonomous System Number (ASN) as the 2-byte administrative field and the Service Identifier (VNI) for the 4-byte numbering field.

2-byte ASN

The Type 0 encoding allows a 2-byte administrative field and a 4-byte numbering field. Within Cisco NX-OS, the auto-derived Route-Target is constructed with the Autonomous System Number (ASN) as the 2-byte administrative filed and the Service Identifier (VNI) for the 4-byte numbering field.

Examples of an auto derived Route-Target (RT):

- IP-VRF within ASN 65001 and L3VNI 50001 - Route-Target 65001:50001
- MAC-VRF within ASN 65001 and L2VNI 30001 - Route-Target 65001:30001

For Multi-AS environments, the Route-Targets must either be statically defined or rewritten to match the ASN portion of the Route-Targets.

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus9000/sw/7-x/command_references/configuration_commands/b_N9K_Config_Commands_703i7x/b_N9K_Config_Commands_703i7x_chapter_010010.html#wp4498893710

4-byte ASN

The Type 0 encoding allows a 2-byte administrative field and a 4-byte numbering field. Within Cisco NX-OS, the auto-derived Route-Target is constructed with the Autonomous System Number (ASN) as the 2-byte administrative filed and the Service Identifier (VNI) for the 4-byte numbering field. With the ASN demand of 4-byte length and the VNI requiring 24-bit (3-bytes), the Sub-Field length within the Extended Community is exhausted (2-byte Type and 6-byte Sub-Field). As a result of the length and format constraint and the importance of the Service Identifiers (VNI) uniqueness, the 4-byte ASN is represented in a 2-byte ASN named AS_TRANS, as described in IETF RFC 6793 section 9 (https://tools.ietf.org/html/rfc6793#section-9). The 2-byte ASN 23456 is registered by the IANA (https://www.iana.org/assignments/iana-as-numbers-special-registry/iana-as-numbers-special-registry.xhtml) as AS_TRANS, a special purpose AS number that aliases 4-byte ASNs.

Example auto derived Route-Target (RT) with 4-byte ASN (AS_TRANS):

- IP-VRF within ASN 65656 and L3VNI 50001 - Route-Target 23456:50001
- MAC-VRF within ASN 65656 and L2VNI 30001 - Route-Target 23456:30001



Note Beginning with Cisco NX-OS Release 9.2(1), auto derived Route-Target for 4-byte ASN is supported.

Guidelines and Limitations for VXLAN BGP EVPN

VXLAN BGP EVPN has the following guidelines and limitations:

- The following guidelines and limitations apply to VXLAN/VTEP using BGP EVPN:
  - SPAN source or destination is supported on any port.

  For more information, see the Cisco Nexus 9000 Series NX-OS System Management Configuration Guide, Release 9.3(x).
- When SVI is enabled on a VTEP (flood and learn, or EVPN) regardless of ARP suppression, make sure that ARP-ETHER TCAM is carved using the **hardware access-list tcam region arp-ether 256 double-wide** command. This requirement does not apply to Cisco Nexus 9300-FX/FX2/FX3 and 9300-GX platform switches and Cisco Nexus 9500 platform switches with X97160YC-EX and 9700-FX/FX3 line cards.

Configuration recommendations

- In a VXLAN EVPN setup, border nodes must be configured with unique route distinguishers, preferably using the **auto rd** command. Not using unique route distinguishers across all border nodes is not supported. The use of unique route distinguishers is strongly recommended for all VTEPs of a fabric.
- Mobility Sequence number of a locally originated type-2 route (MAC/MAC-IP) can be mismatched between vPC peers, with one vTEP having a sequence number K while other vTEP in the same complex can have the same route with sequence number 0. This does not cause any functional impact and the traffic is not impacted even after the host moves.
- For SVI-related triggers (such as shut/unshut or PIM enable/disable), a 30-second delay was added, allowing the Multicast FIB (MFIB) Distribution module (MFDM) to clear the hardware table before toggling between L2 and L3 modes or vice versa.
- The VXLAN network identifier (VNID) 16777215 is reserved and should explicitly not be configured.
- Non-Disruptive In Service Software Upgrade (ND-ISSU) is supported on Nexus 9300 with VXLAN enabled. For more information, see Cisco Nexus 9000 and 3000 Upgrade and ISSU Matrix.
- Beginning with Cisco NX-OS Release 10.5(2)F, Ingress Replication and Multicast Underlay are supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.
- You can configure EVPN over segment routing or MPLS. See the Cisco Nexus 9000 Series NX-OS Label Switching Configuration Guide, Release 9.3(x) for more information.
- You can use MPLS tunnel encapsulation using the new CLI encapsulation mpls command. You can configure the label allocation mode for the EVPN address family. See the Cisco Nexus 9000 Series NX-OS Label Switching Configuration Guide, Release 9.3(x) for more information.
- It is recommended to use the **vpc orphan-ports suspend** command for single attached and/or routed devices on a Cisco Nexus 9000 platform switch acting as vPC VTEP.
- Cisco Nexus supports Type-6 EVPN routes (for IPv4) based on earlier version of **draft-ietf-bess-evpn-igmp-mld-proxy** draft, where SMET flag field is set as optional.
- Routing protocol adjacencies using Anycast Gateway SVIs is not supported.

- When running VXLAN EVPN, any SVI for a VLAN extended over VXLAN must be configured with Anycast Gateway. Any other mode of operation is not supported.

Supported platform and releases for VXLAN BGP EVPN

- VXLAN BGP EVPN is supported on the following Cisco Nexus 9300-FX/GX/FX2/FX3 platform switches.
- VXLAN BGP EVPN is supported on Cisco Nexus 9500 platform switches with X97160YC-EX and 9700-FX/GX/FX3 line cards.
- For the Cisco Nexus 9504 and 9508 with R-series line cards, VXLAN EVPN (Layer 2 and Layer 3) is only supported with the 9636C-RX and 96136YC-R line cards.
- Beginning with NX-OS version 9.3(3), the Cisco Nexus 9300-GX switch supports VXLAN BGP EVPN for Layer-2 and Layer-3 Services with both Ingress Replication and Multicast in the underlay.
- Beginning with Cisco NX-OS Release 10.2(1q)F, VXLAN EVPN is supported on Cisco Nexus N9KC9332D-GX2B platform switches.
- Beginning with Cisco NX-OS Release 10.2(3)F, VXLAN EVPN is supported on Cisco Nexus 9364D-GX2A, and 9348D-GX2A platform switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, VXLAN EVPN is supported on Cisco Nexus 9348GC-FX3, 9348GC-FX3PH, and 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, VXLAN EVPN is supported on Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, VXLAN EVPN is supported on Cisco Nexus 9364C-H1 switches.
- Beginning with Cisco NX-OS Release 10.5(2)F, VXLAN EVPN (Layer 2 and Layer 3) is supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.
- Beginning with Cisco NX-OS Release 10.5(3)F, Cisco Nexus 9364E-SG2 and 9364E-SG2-Q switches support VXLAN BGP EVPN based VTEP endpoint information distribution.
- Beginning with Cisco NX-OS Release 10.6(1)F, VXLAN BGP EVPN (Layer 2 and Layer 3) is supported on Cisco Nexus 9336C-SE1 switches.

Unsupported features

- DHCP snooping (Dynamic Host Configuration Protocol snooping) is not supported on VXLAN VLANs.
- RACLs are not supported on VXLAN uplink interfaces. VACLs are not supported on VXLAN de-capsulated traffic in egress direction; this applies for the inner traffic coming from network (VXLAN) towards the access (Ethernet).

  As a best practice, always use PACLs/VACLs for the access (Ethernet) to the network (VXLAN) direction. See the Cisco Nexus 9000 Series NX-OS Security Configuration Guide, Release 9.3(x) for other guidelines and limitations for the VXLAN ACL feature.
- The Cisco Nexus 9000 QoS buffer-boost feature is not applicable for VXLAN traffic.
-

Scale

Note For information about VXLAN BGP EVPN scalability, see the Cisco Nexus 9000 Series NX-OS Verified Scalability Guide.

- In a VXLAN EVPN setup that has 2K VNI scale and 3900 new VNI scaleconfiguration, the control plane down time may take more than 200 seconds. To avoid potential BGP flap, extend the graceful restart time to 300 seconds.
- Beginning with Cisco NX-OS Release 10.2(2)F, the following scale limits are enhanced — Layer 2 VNIs, Extended Layer 2 VNIs, Layer 3 VNIs, SVI with Distributed Anycast Gateway, IPv4 and IPv6 host routes in internet-peering mode and the ECMP paths. For the VXLAN scale limit information, see the Cisco Nexus 9000 Series NX-OS Verified Scalability Guide, Release 10.2(2)F.

ARP suppression

- The command **clear ip arp** *interface* **vrf** *vrf-name* **force-delete** on specific interface normally deletes entries from ARP belonging to that interface and will relearn on traffic. However, when ARP for same IP is resolved on all ECMP paths, force-deleting ARP entry belonging to one of the ECMP interface will result in automatic relearning of that entry unless that link is down.
- IP unnumbered in EVPN underlay supports ECMP. Multiple IP unnumbered links are connected back to back between same switches. ARP will be resolved on all connected interfaces, thus providing ECMP.
- ARP suppression is only supported for a VNI if the VTEP hosts the First-Hop Gateway (Distributed Anycast Gateway) for this VNI. The VTEP and the SVI for this VLAN have to be properly configured for the distributed Anycast Gateway operation, for example, global Anycast Gateway MAC address configured and Anycast Gateway feature with the virtual IP address on the SVI.
- The ARP suppression setting must match across the entire fabric. For a specific VNID, all VTEPs must be either configured or not configured.
- Beginning with Cisco NX-OS Release 10.5(2)F, ARP suppression is supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.
- Beginning with Cisco NX-OS Release 10.6(1)F, ARP suppression is supported on Cisco Nexus 9336C-SE1 switches.

VXLAN BGP EVPN fabrics with eBGP

- For VXLAN BGP EVPN fabrics with eBGP, the following recommendations are applicable:
  - It is recommended to use loopbacks for the eBGP EVPN peering sessions (overlay control-plane).
  - It is a best practice to use the physical interfaces for eBGP IPv4/IPv6 peering sessions (underlay).
- Only eBGP peering between a VTEP and external nodes (Edge Router, Core Router or VNF) is supported.
  - eBGP peering from the VTEP to the external node using a physical interface or subinterfaces is recommended and it is a best practice (external connectivity).
  - The eBGP peering from the VTEP to the external node can be in the default VRF or in a tenant VRF (external connectivity).

- The eBGP peering from the VTEP to a external node over VXLAN must be in a tenant VRF and must use the update-source of a loopback interface (peering over VXLAN).
- Using an SVI for eBGP peering from the VTEP to the External Node requires the VLAN to be local (not VXLAN extended).

NVE interface

- Bind the NVE source-interface to a dedicated loopback interface and do not share this loopback with any function or peerings of Layer-3 protocols. A best practice is to use a dedicated loopback address for the VXLAN VTEP function.
- You must bind NVE to a loopback address that is separate from other loopback addresses that are required by Layer 3 protocols. NVE and other Layer 3 protocols using the same loopback is not supported.
- The NVE source-interface loopback is required to be present in the default VRF.
- During the vPC Border Gateway boot up process the NVE source loopback interface undergoes the hold down timer twice instead of just once. This is a day-1 and expected behavior.
- The value of the delay timer on NVE interface must be configured to a value that is less than the multi-site delay-restore timer.

Supported Templates

- When configuring VXLAN BGP EVPN, only the "System Routing Mode: Default" is applicable for the following hardware platforms:
  - Cisco Nexus 9300 platform switches
  - Cisco Nexus 9300-FX/FX2/FX3 platform switches
  - Cisco Nexus 9300-GX/GX2/H2R/H1 platform switches
  - Cisco Nexus 9500 platform switches with X97160YC-EX andX9700-FX/GX/FX3 line cards
- Changing the “System Routing Mode” requires a reload of the switch.

VXLAN uplinks

- Starting from Cisco NX-OS Release 9.3(5), new VXLAN uplink capabilities are introduced:
  - A physical interface in default VRF is supported as VXLAN uplink.
  - A parent interface in default VRF, carrying subinterfaces with VRF and dot1q tags, is supported as VXLAN uplink.
  - A subinterface in any VRF and/or with dot1q tag remains not supported as VXLAN uplink.
  - An SVI in any VRF remains not supported as VXLAN uplink.
  - In vPC with physical peer-link, a SVI can be leveraged as backup underlay, default VRF only between the vPC members (infra-VLAN, system nve infra-vlans).

    - On a vPC pair, shutting down NVE or NVE loopback on one of the vPC nodes is not a supported configuration. This means that traffic failover on one-side NVE shut or one-side loopback shut is not supported.
    - FEX host interfaces remain not supported as VXLAN uplink and cannot have VTEPs connected (BUD node).

- You need to configure the VXLAN uplink with **ip unreachables** in order to enable Path maximum transmission unit (MTU) discovery (PMTUD) in a VXLAN set up. PMTUD prevents fragmentation in the path between two endpoints by dynamically determining the lowest MTU along the path from the packet's source to its destination.
- Cisco Nexus 9500 platform switches with 9700-FX or -GX or -FX3 line cards support 1G, 10G, 25G, 40G, 100G and 400G for VXLAN uplinks.
- Cisco Nexus 9300-FX/FX2/FX3 and -GX support 1G, 10G, 25G, 40G, 100G and 400G for VXLAN uplinks.
- Beginning with Cisco NX-OS Release 10.2(3)F, Cisco Nexus 9300-GX2 platform switches support 10G, 25G, 40G, 100G and 400G for VXLAN uplinks.
- Beginning with Cisco NX-OS Release 10.4(1)F, Cisco Nexus 9332D-H2R switches support 10G, 25G, 40G, 100G and 400G for VXLAN uplinks.
- Beginning with Cisco NX-OS Release 10.4(2)F, Cisco Nexus 93400LD-H1 switches support 10G, 25G, 40G, 100G and 400G for VXLAN uplinks.
- Beginning with Cisco NX-OS Release 10.4(3)F, Cisco Nexus 9364C-H1 switches support 10G, 25G, 40G, 100G and 400G for VXLAN uplinks.
- Beginning with Cisco NX-OS Release 10.6(1)F, Cisco Nexus 9336C-SE1 switches supports these uplinks:
    - Only Ethernet Port-channel routed interfaces are supported as uplinks.
    - SVI or L3-subinterfaces as uplinks are not supported

UDP port

- The Cisco Nexus 9000 platform switches use standards conforming UDP port number 4789 for VXLAN encapsulation. This value is not configurable.

Gateway functionality

- Gateway functionality for VXLAN to MPLS (LDP), VXLAN to MPLS-SR (Segment Routing) and VXLAN to SRv6 can be operated on the same Cisco Nexus 9000 Series platform.
    - VXLAN to MPLS (LDP) Gateway is supported on the Cisco Nexus 3600-R and the Cisco Nexus 9500 with R-Series line cards.
    - VXLAN to MPLS-SR Gateway is supported on the Cisco Nexus 9300-FX2/FX3/GX and Cisco Nexus 9500 with R-Series line cards.
    - Beginning with Cisco NX-OS Release 10.2(3)F, VXLAN to MPLS-SR Gateway is supported on the Cisco Nexus 9300-GX2 platform switches.

  - Beginning with Cisco NX-OS Release 10.4(1)F, VXLAN to MPLS-SR Gateway is supported on the Cisco Nexus 9332D-H2R switches.
  - Beginning with Cisco NX-OS Release 10.4(2)F, VXLAN to MPLS-SR Gateway is supported on the Cisco Nexus 93400LD-H1 switches.
  - Beginning with Cisco NX-OS Release 10.4(3)F, VXLAN to MPLS-SR Gateway is supported on the Cisco Nexus 9364C-H1 switches.

- VXLAN to SRv6 is supported on the Cisco Nexus 9300-GX platform.
- Beginning with Cisco NX-OS Release 10.2(3)F, VXLAN to SRv6 is supported on the Cisco Nexus 9300-GX2 platform switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, VXLAN to SRv6 is supported on the Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, VXLAN to SRv6 is supported on the Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, VXLAN to SRv6 is supported on the Cisco Nexus 9364C-H1 switches.

VXLAN and GRE co-existence

- Beginning with Cisco NX-OS Release 10.2(3)F, VXLAN and GRE co-existence is supported on Cisco Nexus 9300-FX/FX2/FX3/GX/GX2 switches, and N9K-C93108TC-FX3P, N9K-C93180YC-FX3, N9K-X9716D-GX switches. Only GRE RX path (decapsulation) is supported. GRE TX path (encapsulation) is not supported.
- Beginning with Cisco NX-OS Release 10.4(1)F, VXLAN and GRE co-existence is supported on Cisco Nexus 9332D-H2R switches. Only GRE RX path (decapsulation) is supported. GRE TX path (encapsulation) is not supported.
- Beginning with Cisco NX-OS Release 10.4(2)F, VXLAN and GRE co-existence is supported on Cisco Nexus 93400LD-H1 switches. Only GRE RX path (decapsulation) is supported. GRE TX path (encapsulation) is not supported.
- Beginning with Cisco NX-OS Release 10.4(3)F, VXLAN and GRE co-existence is supported on Cisco Nexus 9364C-H1 switches. Only GRE RX path (decapsulation) is supported. GRE TX path (encapsulation) is not supported.
- Beginning with Cisco NX-OS Release 10.5(2)F, VXLAN and GRE co-existence is supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.

ECMP resilient hashing

- Resilient hashing is supported on the following switch platform with a VXLAN VTEP configured:
  - Cisco Nexus 9300-FX/FX2/FX3/GX support ECMP resilient hashing.
  - Cisco Nexus 9300 with ALE uplink ports does not support resilient hashing.

Note Resilient hashing is disabled by default.

- Beginning with Cisco NX-OS Release 10.2(3)F, the ECMP resilient hashing is supported on the Cisco Nexus 9300-GX2 platform switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, the ECMP resilient hashing is supported on the Cisco Nexus 9300-H2R platform switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, the ECMP resilient hashing is supported on the Cisco Nexus 93400LD-H1 platform switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, the ECMP resilient hashing is supported on the Cisco Nexus 9364C-H1 switches.

ECMP

Static MAC support on BGP EVPN

- Beginning with Cisco NX-OS Release 10.3(2)F, Static MAC for BGP EVPN is supported on Cisco Nexus 9300-FX/FXP/FX2/FX3/GX/GX2 Series switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, Static MAC for BGP EVPN is supported on Cisco Nexus 9300-H2R series switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, Static MAC for BGP EVPN is supported on Cisco Nexus 93400LD-H1 series switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, Static MAC for BGP EVPN is supported on Cisco Nexus 9364C-H1 switches.
- The **mac address-table static mac-address vlan vlan-id {[drop | interface {type slot/port} | port-channel number]}** command is supported on BGP EVPN.

About VXLAN EVPN with Downstream VNI

Cisco NX-OS Release 9.3(5) introduces VXLAN EVPN with downstream VNI. In earlier releases, the VNI configuration must be consistent across all nodes in the VXLAN EVPN network in order to enable communication between them.

VXLAN EVPN with downstream VNI provides the following solutions:

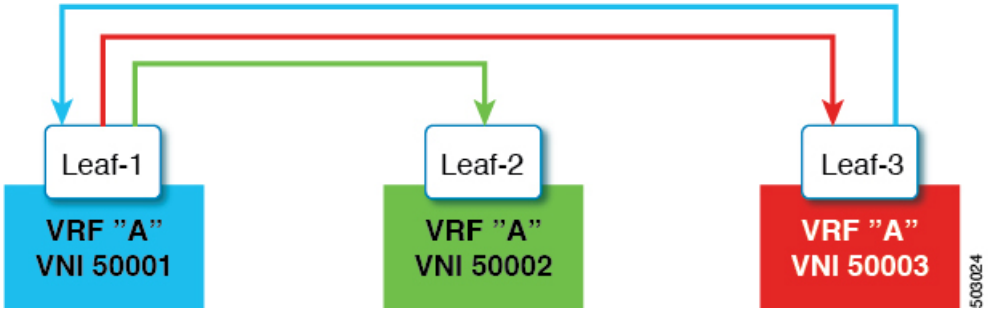
- Enables asymmetric VNI communication across nodes in a VXLAN EVPN network
- Provides customers access to a common shared service outside of their domain (tenant VRF)
- Supports communication between isolated VXLAN EVPN sites that have different sets of VNIs

Asymmetric VNIs

VXLAN EVPN with downstream VNI supports asymmetric VNI allocation.

The following figure shows an example of asymmetric VNIs. All three VTEPs have different VNIs configured for the same IP VRF or MAC VRF.

Figure 11: Asymmetric VNIs



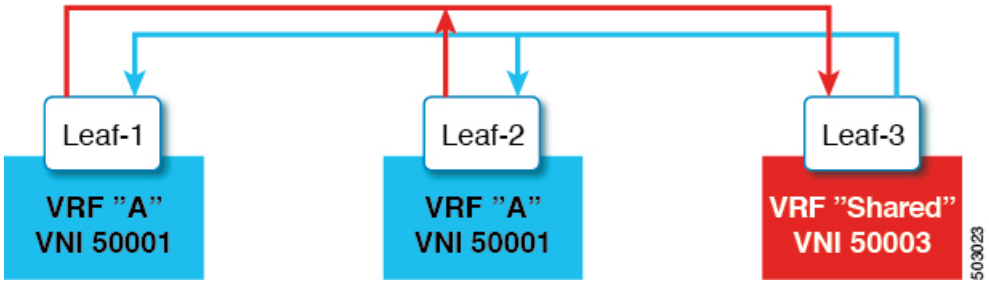
Shared Services VRFs

VXLAN EVPN with downstream VNI supports shared services VRFs. It does so by importing multiple L3VRFs into a single local L3VRF and supporting disparate values of downstream L3VNIs on a per-peer basis.

For example, a DNS server needs to serve multiple hosts in a data center regardless of the tenant VRFs on which the hosts sit. The DNS server is attached to a shared services VRF, which is attached to an L3VNI. To access this server from any of the tenant VRFs, the switches must import the routes from the shared services VRF to the tenant VRF, even though the L3VNI associated to the shared services VRF is different from the L3VNI associated to the tenant VRF.

In the following figure, Tenant VRF A in Leaf-1 can communicate with Tenant VRF A in Leaf-2. However, Tenant VRF A requires access to a shared service sitting behind Leaf-3.

Figure 12: Shared Services VRFs

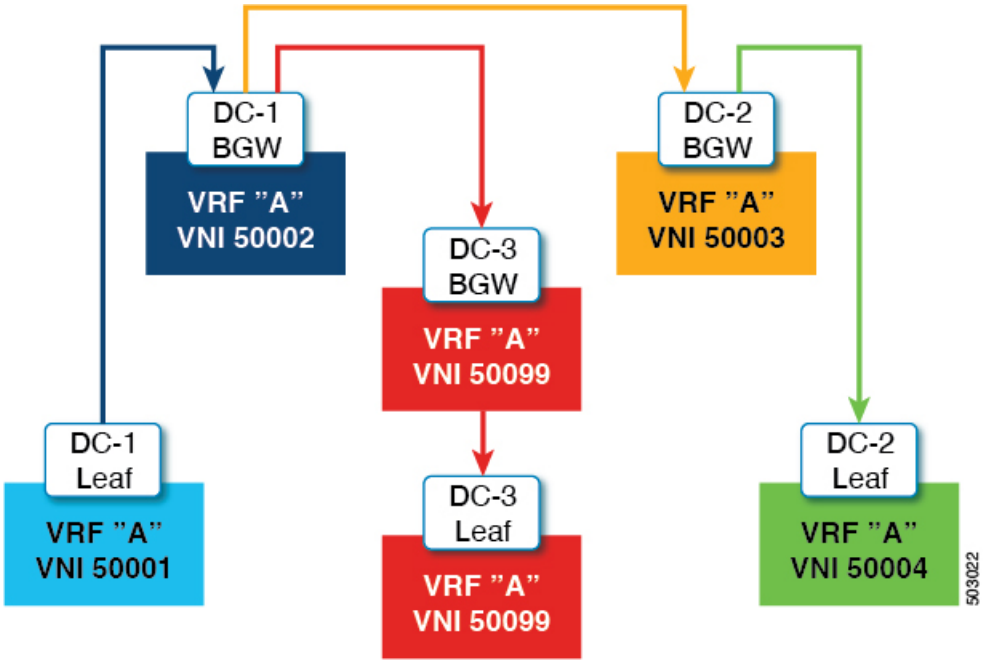


Multi-Site with Asymmetric VNIs

VXLAN EVPN with downstream VNI allows communication between sites that have different sets of VNIs. It does so by stitching the asymmetric VNIs at the border gateways.

In the following figure, DC-1 and DC-2 are asymmetric sites, and DC-3 is a symmetric site. Each site uses different VNIs within its site to communicate.

Figure 13: Multi-Site with Asymmetric VNIs



Guidelines and Limitations for VXLAN EVPN with Downstream VNI

VXLAN EVPN with downstream VNI has the following guidelines and limitations:

- Cisco Nexus 9300-FX/FX2/FXP platform switches and Cisco Nexus 9500 platform switches with -FX/FX3 line cards support VXLAN EVPN with downstream VNI.
- Beginning with Cisco NX-OS Release 9.3(7), Cisco Nexus 9300-GX platform switches support VXLAN EVPN with downstream VNI.
- Beginning with Cisco NX-OS Release 10.2(3)F, the VXLAN EVPN with downstream VNI is supported on the Cisco Nexus 9300-FX3/GX2 platform switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, the VXLAN EVPN with downstream VNI is supported on the Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, the VXLAN EVPN with downstream VNI is supported on the Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, the VXLAN EVPN with downstream VNI is supported on the Cisco Nexus 9364C-H1 switches.
- Beginning with Cisco NX-OS Release 10.5(2)F, VXLAN EVPN with downstream VNI is supported on IPv6 underlay on Cisco Nexus 9300-FX, FX2, FX3, GX, GX2, H2R, H1 Series switches and 9500 Series switches with 9700-FX, FX3, and GX line cards.

- Beginning with Cisco NX-OS Release 10.5(3)F, downstream VNI feature with VXLAN EVPN IPv4 underlay is supported on Cisco Nexus 9800 series switches acting as anycast BGW.
- Downstream VNI is configured based on route-target export and import. The following conditions must be met to leverage Downstream VNI:
  - Downstream VNI requires the usage of different VRF (MAC-VRF or IP-VRF), each VRF must have a different VNI (Asymmetric VNI).
  - To import routes of a foreign VRF (MAC-VRF or IP-VRF) the appropriate route-target for the import into the local VRF must be configured.
  - The configuration of only auto-derived route-targets will not result in downstream VNI.
  - The export of VRF prefixes can be done by static or auto-derived route-target configuration.
  - The import of a foreign VRF's auto-derived route-target is supported.
  - The import of a foreign VRFs statically configured route-target is supported.
- Downstream VNI is supported for the following underlay constellations:
  - For downstream VNI with Layer-3 VNI, the underlay can be ingress replication or multicast based.
  - For downstream VNI with Layer-2 VNI, the underlay must be in ingress replication. Multicast based underlay is not supported with downstream VNI of Layer-2 VNIs.
- Downstream VNI requires to have consistent configuration:
  - All multi-site Border Gateway (BGW) in a site must have a consistent configuration.
  - All vPC members in a vPC domain must have consistent configuration.
- The usage of downstream VNI with multi-site requires all BGW across all sites to run at least Cisco NX-OS Release 9.3(5).
- For existing centralized VRF route leaking deployments, a brief traffic loss might occur during ISSU to Cisco NX-OS Release 9.3(5) or later.
- For successful downgrade from Cisco NX-OS Release 9.3(5) to a prior release, ensure that the asymmetric VNI configuration has been removed. Downstream VNI is not supported before Cisco NX-OS Release 9.3(5) and hence traffic forwarding would be impacted.
- For successful downgrade from Cisco NX-OS Release 10.5(2)F to a prior release, ensure that the IPv6 underlay with downstream VNI configuration is removed, otherwise traffic forwarding may be impacted.
- Layer-3 VNIs (IP-VRF) can flexibly mapped between VNIs per peer.
  - VNI 50001 on VTEP1 can perform symmetric VNI with VNI 50001 and asymmetric VNI with VNI 50002 on VTEP2 at the same time.
  - VNI 50001 on VTEP1 can perform asymmetric VNI with VNI 50002 on VTEP2 and VNI 50003 on VTEP3.
  - VNI 50001 on VTEP1 can perform asymmetric VNI with VNI 50002 and VNI5003 on VTEP2 at the same time.
- Layer-2 VNIs (MAC-VRF) can only be mapped to one VNI per peer.

    - VNI 30001 on VTEP1 can perform asymmetric VNI with VNI 30002 on VTEP2 and VNI 30003 on VTEP3.
    - VNI 30001 on VTEP1 cannot perform asymmetric VNI with VNI 30002 and VNI 3003 on VTEP2 at the same time.

- iBGP sessions between vPC peer nodes in a VRF are not supported.
- BGP peering across VXLAN and Downstream VNI support the following constellations:
    - BGP peering between symmetric VNI is supported by using loopbacks.
    - BGP peering between asymmetric VNI is supported if the VNIs are in a direct message relationship. A loopback from VNI 50001 (on VTEP1) can peer with a loopback in VNI 50002 (on VTEP2).
    - BGP peering between asymmetric VNI is supported if the VNIs are in a direct message relationship but on different VTEPs. A loopback from VNI 50001 (on VTEP1) can peer with a loopback in VNI 50002 (on VTEP2 and VTEP3).
    - BGP peering between asymmetric VNI is not supported if the VNIs are in a 1:N relationship. A loopback in VNI 50001 (VTEP1) can't peer with a loopback in VNI 50002 (VTEP2) and VNI 50003 (VTEP3) at the same time.
- VXLAN consistency checker is not supported for VXLAN EVPN with downstream VNI.
- VXLAN EVPN with downstream VNI is currently not supported with the following feature combinations:
    - VXLAN static tunnels
    - TRM and TRM with Multi-Site
    - ESI-based multihoming
    - Seamless integration of EVPN with L3VPN (MPLS SR)
    - VXLAN policy-based routing (PBR)
    - IPv6 Underlay
- Beginning with Cisco NX-OS Release 10.5(3)F, downstream VNI is supported with CloudSec VXLAN EVPN Tunnel Encryption on Cisco Nexus 9300-FX/FX2/FX3/GX/GX2/H2R/H1 Series switches and 9500 Series switches with 9700-FX/GX/FX3 line cards.
- Make sure that you configure L2VNI SVI on Anycast BGW to enable DSVNI MAC-IP Layer 3 label translation in a multisite environment. The functionality of DSVNI is limited for reoriginated routes, which requires as association between L2VNI and VRF. You can associate using the VRF member command in L2VNI SVI.
- Beginning with Cisco NX-OS Release 10.5(3)F, the DSVNI feature is enhanced with the introduction of the new `associate-vrf vrf-name` command, which explicitly maps a Layer 2 VLAN to a VRF, eliminating the need for SVI configuration on the BGW. For more information on the command usage, see Associating VLAN to VRF - L2 to L3 Mapping, on page 146

  This feature enhancement is supported on Cisco Nexus 9300-FX/FX2/FX3/GX/GX2/H2R/H1 Series switches, 9800 Series switches, and 9500 Series switches with 9700-FX/GX/FX3 line cards.

  Beginning with Cisco NX-OS Release 10.6(1)F, this feature is supported on Cisco Nexus Nexus 9336C-SE1 switches.

Configuring VXLAN BGP EVPN

Enabling VXLAN

Enable VXLAN and the EVPN.

SUMMARY STEPS

1. **feature vn-segment**
2. **feature nv overlay**
3. **feature vn-segment-vlan-based**
4. **feature interface-vlan**
5. **nv overlay evpn**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**feature vn-segment**	Enable VLAN-based VXLAN
Step 2	**feature nv overlay**	Enable VXLAN
Step 3	**feature vn-segment-vlan-based**	Enable VN-Segment for VLANs.
Step 4	**feature interface-vlan**	Enable Switch Virtual Interface (SVI).
Step 5	**nv overlay evpn**	Enable the EVPN control plane for VXLAN.

Configuring VLAN and VXLAN VNI



Note Step 3 to Step 6 are optional for configuring the VLAN for VXLAN VNI and are only necessary in case of a custom route distinguisher or route-target requirement (not using auto derivation).

SUMMARY STEPS

1. **vlan** *number*
2. **vn-segment** *number*
3. **evpn**
4. **vni** *number* **l2**
5. **rd auto**
6. **route-target both** {**auto** | *rt*}

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**vlan** *number*	Specify VLAN.
Step 2	**vn-segment** *number*	Map VLAN to VXLAN VNI to configure Layer 2 VNI under VXLAN VLAN.
Step 3	**evpn**	Enter EVI (EVPN Virtual Instance) configuration mode.
Step 4	**vni** *number* **l2**	Specify the Service Instance (VNI) for the EVI.
Step 5	**rd auto**	Specify the MAC-VRF's route distinguisher (RD).
Step 6	**route-target both** {**auto** | *rt*}	Configure the route target (RT) for import and export of MAC prefixes. The RT is used for a per-MAC-VRF prefix import/export policy. If you enter an RT, the following formats are supported: ASN2:NN, ASN4:NN, or IPV4:NN. **Note** Specifying the auto option is applicable only for IBGP. Manually configured route targets are required for EBGP and for asymmetric VNIs.

Configuring New L3VNI Mode

Guidelines and Limitations for New L3VNI Mode

New L3VNI mode has the following configuration guidelines and limitations:

- Beginning with Cisco NX-OS Release 10.2(3)F, the new L3VNI mode is supported on Cisco Nexus 9300-X Cloud Scale Switches.
- **interface vni** config is optional (not needed if the PBR/NAT feature is not required).
- VRF-VNI-L3 new configuration will implicitly create the L3VNI interface. By default, it will not show up in the show running command.



Note Ensure that VRF-VNI-L3 is configured before configuring **interface vni**.

- Following configuration are allowed on **interface vni**:
  - PBR/NAT
  - no interface vni
  - default interface vni (will remove PBR/NAT configuration if present)

- The **shut/no shut** command is not allowed on **interface vni**. Performing **shut/no shut** command on VRF performs shut/no shut on L3VNI.
- Performing **no feature nv overlay** with the new L3VNI configuration removes all vrf-vni-l3 config under VRF and cleanup the PBR/NAT configuration, if present. Any existing VRF configuration will not be removed.
- VNI Configuration has the following guidelines and limitations:
  - Both old and new L3VNI mode configuration can coexist on the same switch.
  - For the VPC/VMCT system, same VNI config mode should be consistent across peers.
  - Post upgrade, the old L3VNI configuration holds good.
  - Beginning with Cisco NX-OS Release 10.3(1)F, TRM support for the new L3VNI is provided on Cisco Nexus 9300-X Cloud Scale Switches.
  - Config-replace and rollback are supported.
  - ISSU (ND) is supported for the new L3VNI.
- PBR/NAT configuration on the new L3VNI has the following guidelines and limitations:
  - NAT configuration can be applied on the new **interface vni**.
  - PBR encap side policy is still configured on encap node interface SVI as existing.
  - PBR decap side policy for the new L3VNI now applies on **interface vni** for the corresponding L3VNI.
  - PBR config syntax on the new L3VNI is similar to SVI interface.
  - The **no interface vni** removes the PBR/NAT config first and then remove the **interface vni**.
  - The **no interface vni** will only remove the CLI from config, as long as VRF-VNI-L3 config is still present, the **interface vni** is still present at the back-end.
- The following features are supported on the new L3VNI mode:
  - Leaf/VTEP features which use L3VNIs
    - VxLAN EVPN
      - IR and multicast.
      - IGMP Snooping
      - vPC
      - Distributed Anycast Gateway
    - MCT-less vPC
    - VxLAN Multisite
      - Cover all existing scenarios with Border Leaf, Border Spine and multi-site Border Gateway
      - Anycast BGW and vPC BGW

- DSVNI
- VxLAN NGOAM

- VXLAN supported features: PBR, NAT, and QoS
- VXLAN access features (QinVNI, SQinVNI, NIA, BUD-Node etc.)
- 4K scale L2VNI for VXLAN Port VLAN-Mapping VXLAN feature.

- Migration of L3VNI configuration has the following guidelines and limitations:
  - To migrate the L3VNI configuration from old to new, perform the following steps:
    1. Remove the VLAN, vlan-vnsegment and SVI configuration..
    2. Retain Interface nve1 member-vni-associate configuration.
    3. Add new VRF-VNI-L3 configuration. For more information, refer to Configuring New L3VNI Mode, on page 130.

  - To migrate the L3VNI configuration from new to old, perform the following steps:
    1. Remove new VRF-VNI-L3 configuration.
    2. Create VLAN and vlan-vnsegment configuration.
    3. Retain Interface nve1 member-vni-associate configuration.
    4. Create SVI configuration for the L3VNI.
    5. Add member-vni under VRF configuration.

- Upgrade and Downgrade have the following guidelines and limitations:
  - Upgrade:
    - The existing L3VNI configuration remains unchanged and continues to function properly.
    - You can configure additional L3VNIs with the new keyword **L3** without VLAN association.
    - You can choose to migrate the existing L3VNI config one by one to the new L3VNI without VLAN association.
    - If needed, you can revert from new L3VNI config to old L3VNI config (with VLAN association).
    - ND ISSU is supported for new L3VNI future releases.

  - Downgrade:
    - If the new L3 VNI is configured, check and disable the new L3VNI configuration before performing downgrade.
    - Downgrade will be allowed only after removing all new L3VNI configuration.

Configuring New L3VNI Mode

This procedure enables the new L3VNI mode on the switch:

SUMMARY STEPS

1. **configure terminal**
2. **vrf context** *vrf-name*
3. **vni** *number* **l3**
4. **member vni** *vni id* **associate-vrf**
5. (Optional) **{ip | ipv6} policy route-map** *map-name*
6. (Optional) **ip nat outside**

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# configure terminal`
	`switch(config)#`
		Enters global configuration mode.

Step 2	**vrf context** *vrf-name* **Example:** `switch(config)# vrf context vxlan-501`	Configures the VRF.
Step 3	**vni** *number* **l3** **Example:** `switch(config)# vni 500001  L3`	Specifies the VNI. **L3** is the new keyword which indicates the new L3VNI mode.

Step 4	**member vni** *vni id* **associate-vrf**
	Example:
	`switch(config)# interface nve1`
	`switch(config-intf)# no shutdown`
	`switch(config-intf)# member vni 500001 associate-vrf`
		Associates L3VNI to VRF.

Step 5	(Optional) **{ip | ipv6} policy route-map** *map-name*
	Example:
	`switch(config)# interface vni 500001`
	Example:
	For IPv4
	`switch(config-intf)# ip policy route-map IPV4_PBR_Appgroup`
	Example:
		Assigns a route map for IPv4 or IPv6 policy-based routing to L3VNI interface.

	Command or Action	Purpose
	For IPv6 `switch(config-intf)# ipv6 policy route-map IPV6_PBR_Appgroup`	

Step 6	(Optional) **ip nat outside**
	Example:
	`switch(config)# interface vni 500001`
	`switch(config-intf)# ip nat outside`
		Assigns a route map for NAT to L3VNI interface.

Verifying New L3VNI Mode Configuration

To display the new L3VNI mode configuration information, perform the following task:

Command	Purpose
Show nve vni	Displays corresponding new l3vni state

Configuring VRF for VXLAN Routing

Configure the tenant VRF.

Note Step 3 to step 6 are optional for configuring the VRF for VXLAN Routing and are only necessary in case of a custom route distinguisher or route-target requirement (not using auto derivation).

SUMMARY STEPS

1. **vrf context** *vrf-name*
2. **vni** *number*
3. **rd auto**
4. **address-family** {**ipv4** | **ipv6**} **unicast**
5. **route-target both** {**auto** | *rt*}
6. **route-target both** {**auto** | *rt*} **evpn**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**vrf context** *vrf-name*	Configure the VRF.
Step 2	**vni** *number*	Specify the VNI.
Step 3	**rd auto**	Specify the IP-VRF's route distinguisher (RD).

	Command or Action	Purpose
Step 4	**address-family** {**ipv4** | **ipv6**} **unicast**	Configure the IPv4 or IPv6 unicast address family.
Step 5	**route-target both** {**auto** | *rt*}	Configure the route target (RT) for import and export of IPv4 or IPv6 prefixes. The RT is used for a per-IP-VRF prefix import/export policy. If you enter an RT, the following formats are supported: ASN2:NN, ASN4:NN, or IPV4:NN. **Note** Specifying the **auto** option is applicable only for IBGP. Manually configured route targets are required for EBGP and for asymmetric VNIs.
Step 6	**route-target both** {**auto** | *rt*} **evpn**	Configure the route target (RT) for import and export of IPv4 or IPv6 prefixes. The RT is used for a per-VRF prefix import/export policy. If you enter an RT, the following formats are supported: ASN2:NN, ASN4:NN, or IPV4:NN. **Note** Specifying the **auto** option is applicable only for IBGP. Manually configured route targets are required for EBGP and for asymmetric VNIs.

Configuring VXLAN UDP Source Port

Configure the VXLAN UDP source port.

SUMMARY STEPS

1. [**no**] **vxlan udp src-port** [**high** | **rfc** | **low**]

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	[**no**] **vxlan udp src-port** [**high** | **rfc** | **low**]	Allows to select the VXLAN UDP source port number range for VXLAN encapsulated packets. **high**: This option sets the port number range to 0x8000-0xFFFF. **rfc**: Beginning with Cisco NX-OS Release 10.4(1)F, the **rfc** option is provided to set the port number range to 0xC000-0xFFFF. **Note**

	Command or Action	Purpose
		The **rfc** option is available only on Cisco Nexus 9332D-H2R, 9364C-H1, and 93400LD-H1 switches. **low**: Beginning with Cisco NX-OS Release 10.4(1)F, the **low** option is provided to set the port number range to default value (1024 to 32K-1). This is the default option. The **no** form of the **high** and **rfc** command is equivalent to the **low** command. **Note** The **low** option is available on all Cisco Nexus 9000 Series platform switches.

Configuring SVI for Core-facing VXLAN Routing

Configure the core-facing SVI VRF.

SUMMARY STEPS

1. **vlan** *number*
2. **vn-segment** *number*
3. **interface** *vlan-number*
4. **mtu** *number*
5. **vrf member** *vrf-name*
6. **no** {**ip** |**ipv6**} **redirects**
7. **ip forward**
8. **ipv6 address use-link-local-only**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**vlan** *number*	Specify VLAN.
Step 2	**vn-segment** *number*	Map VLAN to VXLAN VNI to configure Layer 3 VNI under VXLAN VLAN.
Step 3	**interface** *vlan-number*	Specify VLAN interface.
Step 4	**mtu** *number*	MTU size in bytes <68-9216>.
Step 5	**vrf member** *vrf-name*	Assign to VRF.
Step 6	**no** {**ip** |**ipv6**} **redirects**	Disable sending IP redirect messages for IPv4 and IPv6.
Step 7	**ip forward**	Enable IPv4 based lookup even when the interface VLAN has no IP address defined.

	Command or Action	Purpose
Step 8	**ipv6 address use-link-local-only**	Enable IPv6 forwarding. **Note** The IPv6 address use-link-local-only serves the same purpose as ip forward for IPv4. It enables the switch to perform an IP based lookup even when the interface VLAN has no IP address defined under it.

Configuring SVI for Host-Facing VXLAN Routing

Configure the SVI for hosts, acting as Distributed Default Gateway.

SUMMARY STEPS

1. **fabric forwarding anycast-gateway-mac** *address*
2. **vlan** *number*
3. **vn-segment** *number*
4. **interface** *vlan-number*
5. **vrf member** *vrf-name*
6. **ip address** *address*
7. **fabric forwarding mode anycast-gateway**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**fabric forwarding anycast-gateway-mac** *address*	Configure distributed gateway virtual MAC address. **Note** One virtual MAC per VTEP. **Note** All VTEPs should have the same virtual MAC address.
Step 2	**vlan** *number*	Specify VLAN.
Step 3	**vn-segment** *number*	Specify vn-segment.
Step 4	**interface** *vlan-number*	Specify VLAN interface.
Step 5	**vrf member** *vrf-name*	Assign to VRF.
Step 6	**ip address** *address*	Specify IP address.
Step 7	**fabric forwarding mode anycast-gateway**	Associate SVI with anycast gateway under VLAN configuration mode.

Configuring the NVE Interface and VNIs Using Multicast

SUMMARY STEPS

1. **interface** *nve-interface*
2. **source-interface loopback1**
3. **host-reachability protocol bgp**
4. **global mcast-group** *ip-address* **{L2 | L3}**
5. **member vni** *vni*
6. **mcast-group** *ip address*
7. **member vni** *vni* **associate-vrf**
8. **mcast-group** *address*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**interface** *nve-interface*	Configure the NVE interface.
Step 2	**source-interface loopback1**	Binds the NVE source-interface to a dedicated loopback interface.
Step 3	**host-reachability protocol bgp**	This defines BGP as the mechanism for host reachability advertisement
Step 4	**global mcast-group** *ip-address* **{L2 | L3}**	Configures the mcast group globally (for all VNI) on a per-NVE interface basis. This applies and gets inherited s to all Layer 2 or Layer 3 VNIs. **Note** Layer3 macst group is only used for Tenant Routed Multicast (TRM).
Step 5	**member vni** *vni*	Add Layer 2 VNIs to the tunnel interface.
Step 6	**mcast-group** *ip address*	Configure the mcast group on a per-VNI basis. Add Layer 2 VNI specific mcast group and override the global set configuration. **Note** Instead of a mcast group, ingress replication can be configured.
Step 7	**member vni** *vni* **associate-vrf**	Add Layer-3 VNIs, one per tenant VRF, to the overlay. **Note** Required for VXLAN routing only.

	Command or Action	Purpose
Step 8	**mcast-group** *address*	Configure the mcast group on a per-VNI basis. Add Layer 3 VNI specific mcast group and override the global set configuration.

Configuring the Delay Timer on NVE Interface

Configuring the delay timer on NVE interface allows BGP to delay the fabric route advertisement to VRF peers and VRF peer routes to fabric so that there are no transient traffic drops seen when border leaf nodes come up after a switch reload. Configure this timer on NX-OS border leaf and AnyCast border gateway.

The value of the delay timer on NVE interface depends on the scale values of NVE peers, VNIs, routes, and so on. To find the timer value to be configured, find the time it took to program the last NVE peer after reload and add buffer time of 100 seconds to it. This buffer time also provides time for route-advertisement. Use the **show forwarding internal trace nve-peer-history** command to display the time stamp of each NVE peer installed.

Also, convergence will not be improved for fabric isolation on NX-OS border leaf even when this timer is configured.

SUMMARY STEPS

1. **configure terminal**
2. **interface nve** *nve-interface*
3. **fabric-ready time** *seconds*
4. **show nve interface nve1 detail**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**interface nve** *nve-interface*	Configures the NVE interface.
Step 3	**fabric-ready time** *seconds*	Specifies the delay timer value for NVE interface. The default value is 135 seconds.
Step 4	**show nve interface nve1 detail**	Displays the configured timer value.

Configuring VXLAN EVPN Ingress Replication

For VXLAN EVPN ingress replication, the VXLAN VTEP uses a list of IP addresses of other VTEPs in the network to send BUM (broadcast, unknown unicast and multicast) traffic. These IP addresses are exchanged between VTEPs through the BGP EVPN control plane.

Note VXLAN EVPN ingress replication is supported on:

- Cisco Nexus Series 9300 Series switches (7.0(3)I1(2) and later).
- Cisco Nexus Series 9500 Series switches (7.0(3)I2(1) and later).

Before you begin: The following are required before configuring VXLAN EVPN ingress replication (7.0(3)I1(2) and later):

- Enable VXLAN.
- Configure VLAN and VXLAN VNI.
- Configure BGP on the VTEP.
- Configure RD and Route Targets for VXLAN Bridging.

SUMMARY STEPS

1. **interface** *nve-interface*
2. **host-reachability protocol bgp**
3. **global ingress-replication protocol bgp**
4. **member vni** *vni* **associate-vrf**
5. **member vni** *vni*
6. **ingress-replication protocol bgp**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**interface** *nve-interface*	Configure the NVE interface.
Step 2	**host-reachability protocol bgp**	This defines BGP as the mechanism for host reachability advertisement.
Step 3	**global ingress-replication protocol bgp**	Enables globally (for all VNI) the VTEP to exchange local and remote VTEP IP addresses on the VNI in order to create the ingress replication list. This enables sending and receiving BUM traffic for the VNI. **Note** Using ingress-replication protocol bgp avoids the need for any multicast configurations that might have been required for configuring the underlay.
Step 4	**member vni** *vni* **associate-vrf**	Add Layer-3 VNIs, one per tenant VRF, to the overlay. **Note** Required for VXLAN routing only.

	Command or Action	Purpose
Step 5	**member vni** *vni*	Add Layer 2 VNIs to the tunnel interface.
Step 6	**ingress-replication protocol bgp**	Enables the VTEP to exchange local and remote VTEP IP addresses on a oer VNI basis in order to create the ingress replication list. This enables sending and receiving BUM traffic for the VNI and override the global configuration. **Note** Instead of a ingress replication, mcast group can be configured. **Note** Using **ingress-replication protocol bgp** avoids the need for any multicast configurations that might have been required for configuring the underlay.

Configuring BGP on the VTEP

SUMMARY STEPS

1. **router bgp** *number*
2. **router-id** *address*
3. **neighbor** *address* **remote-as** *number*
4. **address-family l2vpn evpn**
5. (Optional) **Allowas-in**
6. **send-community extended**
7. **vrf** *vrf-name*
8. **address-family ipv4 unicast**
9. **maximum-paths path** {**ibgp**}
10. **address-family ipv6 unicast**
11. **maximum-paths path** {**ibgp**}

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**router bgp** *number*	Configure BGP.
Step 2	**router-id** *address*	Specify router address.
Step 3	**neighbor** *address* **remote-as** *number*	Define MPBGP neighbors. Under each neighbor define L2VPN EVPN.
Step 4	**address-family l2vpn evpn**	Configure address family Layer 2 VPN EVPN under the BGP neighbor.

	Command or Action	Purpose
		Note Address-family IPv4 EVPN for VXLAN host-based routing
Step 5	(Optional) **Allowas-in**	Only for EBGP deployment cases: Allows duplicate autonomous system (AS) numbers in the AS path. Configure this parameter on the leaf for eBGP when all leafs are using the same AS, but the spines have a different AS than leafs.
Step 6	**send-community extended**	Configures community for BGP neighbors.
Step 7	**vrf** *vrf-name*	Specify VRF.
Step 8	**address-family ipv4 unicast**	Configure the address family for IPv4.
Step 9	**maximum-paths path** {**ibgp**}	Enable ECMP for EVPN transported IP Prefixes within the IPv4 address-family of the respective VRF.
Step 10	**address-family ipv6 unicast**	Configure the address family for IPv6.
Step 11	**maximum-paths path** {**ibgp**}	Enable ECMP for EVPN transported IP Prefixes within the IPv6 address-family of the respective VRF.

Configuring iBGP for EVPN on the Spine

SUMMARY STEPS

1. **router bgp** *autonomous system number*
2. **neighbor** *address* **remote-as** *number*
3. **address-family l2vpn evpn**
4. **send-community extended**
5. **route-reflector-client**
6. **retain route-target all**
7. **address-family l2vpn evpn**
8. **disable-peer-as-check**
9. **route-map permitall out**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**router bgp** *autonomous system number*	Specify BGP.
Step 2	**neighbor** *address* **remote-as** *number*	Define neighbor.

	Command or Action	Purpose
Step 3	**address-family l2vpn evpn**	Configure address family Layer 2 VPN EVPN under the BGP neighbor.
Step 4	**send-community extended**	Configures community for BGP neighbors.
Step 5	**route-reflector-client**	Enable Spine as Route Reflector.
Step 6	**retain route-target all**	Configure retain route-target all under address-family Layer 2 VPN EVPN [global]. **Note** Required for eBGP. Allows the spine to retain and advertise all EVPN routes when there are no local VNI configured with matching import route targets.
Step 7	**address-family l2vpn evpn**	Configure address family Layer 2 VPN EVPN under the BGP neighbor.
Step 8	**disable-peer-as-check**	Disables checking the peer AS number during route advertisement. Configure this parameter on the spine for eBGP when all leafs are using the same AS but the spines have a different AS than leafs. **Note** Required for eBGP.
Step 9	**route-map permitall out**	Applies route-map to keep the next-hop unchanged. **Note** Required for eBGP.

Configuring eBGP for EVPN on the Spine

SUMMARY STEPS

1. **route-map NEXT-HOP-UNCH permit 10**
2. **set ip next-hop unchanged**
3. **router bgp** *autonomous system number*
4. **address-family l2vpn evpn**
5. **retain route-target all**
6. **neighbor** *address* **remote-as** *number*
7. **address-family l2vpn evpn**
8. **disable-peer-as-check**
9. **send-community extended**
10. **route-map NEXT-HOP-UNCH out**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**route-map NEXT-HOP-UNCH permit 10**	Configure route-map to keepthe next-hop unchanged for EVPN routes.
Step 2	**set ip next-hop unchanged**	Set next-hop address. **Note** When two next hops are enabled, next hop ordering is not maintained. If one of the next hops is a VXLAN next hop and the other next hop is local reachable via FIB/AM/Hmm, the local next hop reachable via FIB/AM/Hmm is always taken irrespective of the order. Directly/locally connected next hops are always given priority over remotely connected next hops.
Step 3	**router bgp** *autonomous system number*	Specify BGP.
Step 4	**address-family l2vpn evpn**	Configure address family Layer 2 VPN EVPN under the BGP neighbor.
Step 5	**retain route-target all**	Configure retain route-target all under address-family Layer 2 VPN EVPN [global]. **Note** Required for eBGP. Allows the spine to retain and advertise all EVPN routes when there are no local VNI configured with matching import route targets.
Step 6	**neighbor** *address* **remote-as** *number*	Define neighbor.
Step 7	**address-family l2vpn evpn**	Configure address family Layer 2 VPN EVPN under the BGP neighbor.
Step 8	**disable-peer-as-check**	Disables checking the peer AS number during route advertisement. Configure this parameter on the spine for eBGP when all leafs are using the same AS but the spines have a different AS than leafs.
Step 9	**send-community extended**	Configures community for BGP neighbors.
Step 10	**route-map NEXT-HOP-UNCH out**	Applies route-map to keep the next-hop unchanged.

Suppressing ARP

The ARP request from a host is normally flooded in the VLAN. It is possible to optimize this flooding behavior by maintaining an ARP cache locally on an attached access switch and generating an ARP response from the information in the local cache. Thus, the ARP broadcast on the overlay or VLAN is prevented.

The cache of remote hosts is built by learning the IP-host or MAC-address information through BGP EVPN MAC route advertisement.

Upon receiving an ARP request, the local cache can be consulted to see if the response can be locally generated. If the cache lookup fails, then the ARP request can be flooded. This helps in the detection of silent hosts.

SUMMARY STEPS

1. **interface nve 1**
2. **global suppress-arp**
3. **member vni** *vni-id*
4. **suppress-arp**
5. **suppress-arp disable**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**interface nve 1**	Create the network virtualization endpoint (NVE) interface.
Step 2	**global suppress-arp**	Configure to suppress ARP globally for all Layer 2 VNI.within the NVE interface.
Step 3	**member vni** *vni-id*	Specify VNI ID.
Step 4	**suppress-arp**	Configure to suppress ARP under Layer 2 VNI and overrides the global set default.
Step 5	**suppress-arp disable**	Disables the global setting of the ARP suppression on a specific VNI.

Disabling VXLANs

SUMMARY STEPS

1. **configure terminal**
2. **no nv overlay evpn**
3. **no feature vn-segment-vlan-based**
4. **no feature nv overlay**
5. (Optional) **copy running-config startup-config**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters configuration mode.
Step 2	**no nv overlay evpn**	Disables EVPN control plane.
Step 3	**no feature vn-segment-vlan-based**	Disables the global mode for all VXLAN bridge domains
Step 4	**no feature nv overlay**	Disables the VXLAN feature.
Step 5	(Optional) **copy running-config startup-config**	Saves the change persistently through reboots and restarts by copying the running configuration to the startup configuration.

Duplicate Detection for IP and MAC Addresses

For IP addresses:

Cisco NX-OS supports duplicate detection for IP addresses. This enables the detection of duplicate IP addresses based on the number of moves in a given time-interval (seconds), if host appears simultaneously under two VTEP's.

Simultaneous availability of host under two VTEP's is detected by host mobility logic with 600 msec refresh timeout for IPv4 hosts and default refresh time out logic for IPv6 addresses (default is 3 seconds).

The default is 5 moves in 180 seconds. (Default number of moves is 5 moves. Default time-interval is 180 seconds.)

After the 5th move within 180 seconds, the switch starts a 30 second lock (hold down timer) before checking to see if the duplication still exists (an effort to prevent an increment of the sequence bit). This 30 second lock can occur 5 times within 24 hours (this means 5 moves in 180 seconds for 5 times) before the switch permanently locks or freezes the duplicate entry. (**show fabric forwarding ip local-host-db vrf abc**)

Wherever a host IP address is permanently frozen, a syslog message is written by HMM.

```
2021 Aug 26 01:08:26 leaf hmm: (vrf-name) [IPv4] Freezing potential duplicate host
20.2.0.30/32, reached recover count (5) threshold
```

The following are example commands to help the configuration of the number of VM moves in a specific time interval (seconds) for duplicate IP-detection:

Command	Description

`switch(config)# fabric forwarding ?`
`anycast-gateway-mac`
`dup-host-ip-addr-detection`
	Available sub-commands:
	• Anycast gateway MAC of the switch.
	• To detect duplicate host addresses in n seconds.

Command	Description
`switch(config)# fabric forwarding dup-host-ip-addr-detection ?` `<1-1000>`	The number of host moves allowed in n seconds. The range is 1 to 1000 moves; default is 5 moves.
`switch(config)# fabric forwarding dup-host-ip-addr-detection 100 ?` `<2-36000>`	The duplicate detection timeout in seconds for the number of host moves. The range is 2 to 36000 seconds; default is 180 seconds.
`switch(config)# fabric forwarding dup-host-ip-addr-detection 100 10`	Detects duplicate host addresses (limited to 100 moves) in a period of 10 seconds.

For MAC addresses:

Cisco NX-OS supports duplicate detection for MAC addresses. This enables the detection of duplicate MAC addresses based on the number of moves in a given time-interval (seconds).

The default is 5 moves in 180 seconds. (Default number of moves is 5 moves. Default time-interval is 180 seconds.)

After the 5th move within 180 seconds, the switch starts a 30 second lock (hold down timer) before checking to see if the duplication still exists (an effort to prevent an increment of the sequence bit). This 30 second lock can occur 3 times within 24 hours (this means 5 moves in 180 seconds for 3 times) before the switch permanently locks or freezes the duplicate entry. (**show l2rib internal permanently-frozen-list**)

Wherever a MAC address is permanently frozen, a syslog message with written by L2RIB.

```
2017 Jul  5 10:27:34 leaf %$ VDC-1 %$  %USER-2-SYSTEM_MSG: Unfreeze limit (3) hit, MAC
0000.0033.3333in topo: 200 is permanently frozen - l2rib
2017 Jul  5 10:27:34 leaf %$ VDC-1 %$  %USER-2-SYSTEM_MSG: Detected duplicate host
0000.0033.3333, topology 200, during Local update, with host located at remote VTEP 1.2.3.4,
 VNI 2 - l2rib
2017 Jul  5 10:27:34 leaf %$ VDC-1 %$  %USER-2-SYSTEM_MSG: Unfreeze limit (3) hit, MAC
0000.0033.3334in topo: 200 is permanently frozen - l2rib
2017 Jul  5 10:27:34 leaf %$ VDC-1 %$  %USER-2-SYSTEM_MSG: Detected duplicate host
0000.0033.3334, topology 200, during Local update, with host l
```

MAC address remains in permanently frozen list until both local and remote entry exists.

Unconfiguring below commands will not disable permanently frozen functionality rather will change the parameters to default values.

- **l2rib dup-host-mac-detection**
- **l2rib dup-host-recovery**

The following are example commands to help the configuration of the number of VM moves in a specific time interval (seconds) for duplicate MAC-detection:

Command	Description

`switch(config)# l2rib dup-host-mac-detection ?`
`<1-1000>`
`default`
	Available sub-commands for L2RIB:
	• The number of host moves allowed in n seconds. The range is 1 to 1000 moves.
	• Default setting (5 moves in 180 in seconds).

`switch(config)# l2rib dup-host-mac-detection 100 ?` `<2-36000>`	The duplicate detection timeout in seconds for the number of host moves. The range is 2 to 36000 seconds; default is 180 seconds.
`switch(config)# l2rib dup-host-mac-detection 100 10`	Detects duplicate host addresses (limited to 100 moves) in a period of 10 seconds.

Configuring Event History Size for L2RIB

To set the event history size for the L2RIB component follow these steps:

SUMMARY STEPS

1. **configure terminal**
2. **l2rib event-history** { **mac** | **mac-ip** | **loop-detection** } **size** { **default** | **medium** | **high** | **very-high** }
3. **l2rib event-history** { **fl** | **imet** | **dme-oper** } **size** { **default** | **medium** | **high** | **very-high** }
4. **clear l2rib event-history** { **mac** | **mac-ip** | **loop-detection** } **size** { **default** | **medium** | **high** | **very-high** }

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.
Step 2	**l2rib event-history** { **mac**	**mac-ip**	**loop-detection** } **size** { **default**	**medium**	**high**	**very-high** } **Example:** `switch(config)# l2rib event-history mac size low`	Sets the event history size for the L2RIB component.
Step 3	**l2rib event-history** { **fl**	**imet**	**dme-oper** } **size** { **default**	**medium**	**high**	**very-high** } **Example:**	Generates the event logs for specified L2RIB objects: • **fl**: L2RIB VXLAN Flood-list

	Command or Action	Purpose
	`switch(config)# l2rib event-history fl size very-high`	• **imet**: L2RIB IMET • **dme-oper**: L2RIB DME OPER **Note** Ensure to enable buffer size to **very-high** in large scaling environment.
Step 4	**clear l2rib event-history** { **mac** | **mac-ip** | **loop-detection** } **size** { **default** | **medium** | **high** | **very-high** } **Example:** `switch(config)# clear l2rib event-history mac size low`	Clears the set event history size for the L2RIB component.

Associating VLAN to VRF - L2 to L3 Mapping

Beginning with Cisco NX-OS Release 10.5(3)F, the DSVNI feature is enhanced with the following new command for Layer 2 to Layer 3 mapping on the switch.

This feature has the following configuration limitations:

- Layer 2 to Layer 3 mapping can be configured in any of the following two ways:
  - Configuring an SVI for a given Layer 2 VLAN
  - Associating VRF under VLAN mode.
- If both mappings exist in a configuration file, preference is given to the SVI command.
- Both L3VM and VLAN Manager ensure only one mapping is allowed.
- If SVI config exists for a vlan, **associate-vrf** command will be rejected, and vice versa.
- For VLAN range commands, if SVI configuration does not exist for all vlans in the range, the **associate-vrf** command is accepted. Otherwise, it is rejected.



Note VLAN 1 is handled differently. If it has an **associate-vrf** command, the **feature interface vlan** cannot be enabled.

SUMMARY STEPS

1. **configure terminal**
2. **vlan** {*vlan-id* | *vlan-range*}
3. **associate-vrf** *vrf-name*

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# configure terminal`
	`switch(config)#`
		Enters global configuration mode.

Step 2	**vlan** {*vlan-id* | *vlan-range*}
	Example:
	`switch(config)# vlan 2`
	`switch(config-vlan)#`
		Creates a VLAN or a range or VLANs.

Step 3	**associate-vrf** *vrf-name* **Example:** `switch(config-vlan)# associate-vrf vrf-blue`	Maps a Layer 2 VLAN to VRF.

Verifying the VXLAN BGP EVPN Configuration

To display the VXLAN BGP EVPN configuration information, enter one of the following commands:

Command	Purpose
show nve vrf	Displays VRFs and associated VNIs
show bgp l2vpn evpn	Displays routing table information.
show ip arp suppression-cache [**detail** | **summary** | **vlan** *vlan* | **statistics**]	Displays ARP suppression information.
show vxlan interface	Displays VXLAN interface status.
show vxlan interface | count	Displays VXLAN VLAN logical port VP count. **Note** A VP is allocated on a per-port per-VLAN basis. The sum of all VPs across all VXLAN-enabled Layer 2 ports gives the total logical port VP count. For example, if there are 10 Layer 2 trunk interfaces, each with 10 VXLAN VLANs, then the total VXLAN VLAN logical port VP count is 10*10 = 100.
show l2route evpn mac [**all** | **evi** *evi* [**bgp** | **local** | **static** | **vxlan** | **arp**]]	Displays Layer 2 route information.
show l2route evpn fl all	Displays all fl routes.
show l2route evpn imet all	Displays all imet routes.

Command	Purpose
show l2route evpn mac-ip all **show l2route evpn mac-ip all detail**	Displays all MAC IP routes.
show l2route topology	Displays Layer 2 route topology.
show l2route evpn ethernet-segment all detail	Displays detailed information about all Ethernet Segment Identifiers (ESIs) in an EVPN (Ethernet VPN) environment.

Note Although the **show ip bgp** command is available for verifying a BGP configuration, as a best practice, it is preferable to use the **show bgp** command instead.

Verifying the VXLAN EVPN with Downstream VNI Configuration

To display the VXLAN EVPN with downstream VNI configuration information, enter one of the following commands:

Command	Purpose
show bgp evi *l2-evi*	Displays the VRF associated with an L2VNI.
show forwarding adjacency nve platform	Displays both symmetric and asymmetric NVE adjacencies with the corresponding DestInfoIndex.
show forwarding route vrf *vrf*	Displays the egress VNI or downstream VNI for each next-hop.
show ip route detail vrf *vrf*	Displays the egress VNI or downstream VNI for each next-hop.
show l2route evpn mac-ip all detail	Displays labeled next-hops that are present in the remote MAC routes.
show l2route evpn imet all detail	Displays the egress VNI associated with the remote peer.
show nve peers control-plane-vni peer-ip *ip-address*	Displays the egress VNI or downstream VNI for each NVE adjacency.

The following example shows sample output for the **show bgp evi** *l2-evi* command:

```
switch# show bgp evi 100
-----------------------------------------------
  L2VNI ID                    : 100 (L2-100)
  RD                          : 3.3.3.3:32867
  Secondary RD                : 1:100
  Prefixes (local/total)      : 1/6
  Created                     : Jun 23 22:35:13.368170
  Last Oper Up/Down           : Jun 23 22:35:13.369005 / never
  Enabled                     : Yes
```

```
  Associated IP-VRF            : vni100
  Active Export RT list        :
        100:100
  Active Import RT list        :
        100:100
```

The following example shows sample output for the **show forwarding adjacency nve platform** command:

```
switch# show forwarding adjacency nve platform
slot  1
=======
IPv4 NVE adjacency information

next_hop:12.12.12.12   interface:nve1 (0x49000001)  table_id:1
  Peer_id:0x49080002  dst_addr:12.12.12.12  src_addr:13.13.13.13  RefCt:1  PBRCt:0
Flags:0x440800
cp : TRUE, DCI peer: FALSE is_anycast_ip FALSE dsvni peer: FALSE
  HH:0x7a13f  DstInfoIndex:0x3002
    tunnel init: unit-0:0x3 unit-1:0x0

next_hop:12.12.12.12   interface:nve1 (0x49000001)  table_id:1
  Peer_id:0x49080002  dst_addr:12.12.12.12  src_addr:13.13.13.13  RefCt:1  PBRCt:0
Flags:0x10440800
cp : TRUE, DCI peer: FALSE is_anycast_ip FALSE dsvni peer: TRUE
  HH:0x7a142  DstInfoIndex:0x3ffd
    tunnel init: unit-0:0x6 unit-1:0x0
...
```

The following example shows sample output for the **show forwarding route vrf** *vrf* command:

```
switch# show forwarding route vrf vrf1000

slot  1
=======

IPv4 routes for table vrf1000/base

--------------+------------------+--------------+-----------------+-----------------
Prefix        | Next-hop         | Interface    | Labels          | Partial Install
--------------+------------------+--------------+-----------------+-----------------
…..
10.1.1.11/32    12.12.12.12         nve1        dsvni: 301000
10.1.1.20/32    123.123.123.123     nve1        dsvni: 301000
10.1.1.21/32    30.30.30.30         nve1        dsvni: 301000
10.1.1.30/32    10.1.1.30           Vlan10
```

The following example shows sample output for the **show ip route detail vrf** *vrf* command:

```
switch# show ip route detail vrf default
IP Route Table for VRF "default"
  '*' denotes best ucast next-hop
  '**' denotes best mcast next-hop
  '[x/y]' denotes [preference/metric]
  '%<string>' in via output denotes VRF <string>

193.0.1.0/24, ubest/mbest: 4/0
      *via 30.1.0.2, Eth1/1, [100/0], 00:00:05, urib_dt6-client1 segid: 6544, tunnelid:
0x7b9 encap: VXLAN

      *via 30.1.1.2, Eth1/1, [100/0], 00:00:05, urib_dt6-client1 segid: 6545, (Asymmetric)
 tunnelid: 0x7ba encap: VXLAN
```

```
      *via 30.1.2.2, Eth1/1, [100/0], 00:00:05, urib_dt6-client1 segid: 6546, (Asymmetric)
 tunnelid: 0x7bb encap: VXLAN
```

The following example shows sample output for the **show l2route evpn mac-ip all detail** command:

```
switch# show l2route evpn mac-ip all
Flags -(Rmac):Router MAC (Stt):Static (L):Local (R):Remote (V):vPC link
(Dup):Duplicate (Spl):Split (Rcv):Recv(D):Del Pending (S):Stale (C):Clear
(Ps):Peer Sync (Ro):Re-Originated (Orp):Orphan
Topology Mac Address    Host IP   Prod   Flags Seq No  Next-Hops
-------- -------------- --------  ------ ----- ------- ----------------------------
5        0000.0005.1301 1.3.13.1  BGP    --     0      102.1.13.1 (Label: 2000005)
5        0000.0005.1401 1.3.14.1  BGP    --     0      102.1.145.1 (Label: 2000005)
```

Beginning with Cisco NX-OS Release 10.5(3)F, the labeled nexthop and an asymmetric VNI flag is added as shown. For symmetric VNI, labels and flags will not be shown as part of nexthops for EAD and PL.

```
switch# show l2route mac-ip all detail
Flags -(Rmac):Router MAC (Stt):Static (L):Local (R):Remote
(Dup):Duplicate (Spl):Split (Rcv):Recv(D):Del Pending (S):Stale (C):Clear
(Ps):Peer Sync (Ro):Re-Originated (Orp):Orphan (Asy):Asymmetric (Gw):Gateway
(Bh):Blackhole
(Piporp): Directly connected Orphan to PIP based vPC BGW
(Pipporp): Orphan connected to peer of PIP based vPC BGW
Topology    Mac Address    Host IP                                     Prod   Flags
   Seq No     Next-Hops
----------- -------------- ------------------------------------------ ------ -----------------
 ---------- ---------------------------------------
100         abcd.ef12.3455 5.5.5.5                                     BGP    PS,
   2         1.1.1.10 (Label: 10000)(Flags: Asy)
             ESI : 0300.0000.0000.6400.0319
```

The following example shows sample output for the **show l2route evpn imet all detail** command:

```
switch# show l2route evpn imet all

Flags- (F): Originated From Fabric, (W): Originated from WAN

Topology ID VNI         Prod  IP Addr     Flags
----------- ----------- ----- ----------- ---------
3           2000003     BGP   102.1.13.1    -
3           2000003     BGP   102.1.31.1    -
3           2000003     BGP   102.1.32.1    -
3           2000003     BGP   102.1.145.1   -
```

The following example shows sample output for the **show nve peers control-plane-vni** command. In this example, 3000003 is the downstream VNI.

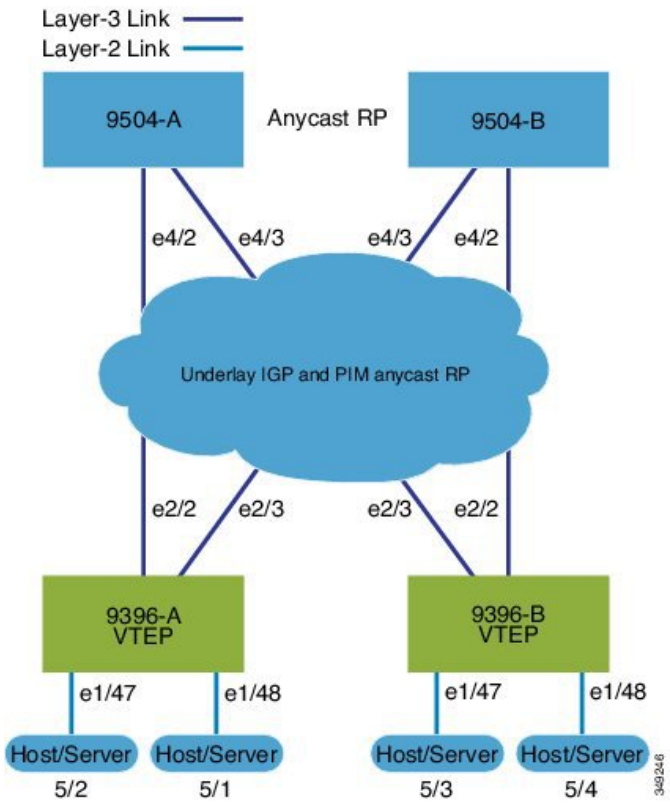
```
switch# show nve peers control-plane-vni peer-ip 203.1.1.1
Peer       VNI     Learn-Source Gateway-MAC     Peer-type  Egress-VNI SW-BD  State

---------  -----   ------------ --------------- ---------- ---------- -----
----------------------
203.1.1.1  2000003 BGP          f40f.1b6f.f8db   FAB        3000003      3005
peer-vni-add-complete
```

Example of VXLAN BGP EVPN (iBGP)

An example of a VXLAN BGP EVPN (iBGP):

Figure 14: VXLAN BGP EVPN Topology (iBGP)



iBGP between Spine and Leaf

- Spine (9504-A)
  - Enable the EVPN control plane

```
nv overlay evpn
```

  - Enable the relevant protocols

```
feature ospf
feature bgp
feature pim
```

  - Configure Loopback for local Router ID, PIM, and BGP

```
interface loopback0
  ip address 10.1.1.1/32
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
```

  - Configure Loopback for Anycast RP

```
interface loopback1
  ip address 100.1.1.1/32
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
```

- Configure Anycast RP

```
ip pim rp-address 100.1.1.1 group-list 224.0.0.0/4
ip pim anycast-rp 100.1.1.1 10.1.1.1
ip pim anycast-rp 100.1.1.1 20.1.1.1
```

- Enable OSPF for underlay routing

```
router ospf 1
```

- Configure interfaces for Spine-leaf interconnect

```
interface Ethernet4/2
  ip address 192.168.1.42/24
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  no shutdown

interface Ethernet4/3
  ip address 192.168.2.43/24
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  no shutdown
```

- Configure BGP

```
router bgp 65535
router-id 10.1.1.1
  neighbor 30.1.1.1 remote-as 65535
    update-source loopback0
    address-family l2vpn evpn
      send-community both
      route-reflector-client
  neighbor 40.1.1.1 remote-as 65535
    update-source loopback0
    address-family l2vpn evpn
      send-community both
      route-reflector-client
```

- Spine (9504-B)
  - Enable the EVPN control plane

```
nv overlay evpn
```

  - Enable the relevnt Protocols

```
feature ospf
feature bgp
feature pim
```

  - Configure Loopback for local Router ID, PIM, and BGP

```
interface loopback0
  ip address 20.1.1.1/32
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
```

- Configure Loopback for AnycastRP

```
interface loopback1
  ip address 100.1.1.1/32
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
```

- Configure Anycast RP

```
ip pim rp-address 100.1.1.1 group-list 224.0.0.0/4
ip pim anycast-rp 100.1.1.1 10.1.1.1
ip pim anycast-rp 100.1.1.1 20.1.1.1
```

- Enable OSPF for underlayrouting

```
router ospf 1
```

- Configure interfaces for Spine-leaf interconnect

```
interface Ethernet4/2
  ip address 192.168.3.42/24
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  no shutdown

interface Ethernet4/3
  ip address 192.168.4.43/24
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  no shutdown
```

- Configure BGP

```
router bgp 65535
  router-id 20.1.1.1
  neighbor 30.1.1.1 remote-as 65535
    update-source loopback0
    address-family l2vpn evpn
      send-community both
      route-reflector client
  neighbor 40.1.1.1 remote-as 65535
    update-source loopback0
    address-family l2vpn evpn
      send-community both
      route-reflector client
```

- Leaf (9396-A)
  - Enable the EVPN control plane

```
nv overlay evpn
```

  - Enable the relevant protocols

```
feature ospf
feature bgp
feature pim
feature interface-vlan
```

- Enable VXLAN with distributed anycast-gateway using BGP EVPN

```
feature vn-segment-vlan-based
feature nv overlay
fabric forwarding anycast-gateway-mac 0000.2222.3333
```

- Enabling OSPF for underlay routing

```
router ospf 1
```

- Configure Loopback for local Router ID, PIM, and BGP

```
interface loopback0
  ip address 30.1.1.1/32
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
```

- Configure Loopback for local VTEP IP

```
interface loopback1
  ip address 33.1.1.1/32
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
```

- Configure interfaces for Spine-leaf interconnect

```
interface Ethernet2/2
  no switchport
  ip address 192.168.1.22/24
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  no shutdown

interface Ethernet2/3
  no switchport
  ip address 192.168.3.23/24
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  shutdown
```

- Configure route-map to Redistribute Host-SVI (Silent Host)

```
route-map HOST-SVI permit 10
  match tag 54321
```

- Configure PIM RP

```
ip pim rp-address 100.1.1.1 group-list 224.0.0.0/4
```

- Create VLANs

```
vlan 1001-1002
```

- Create overlay VRF VLAN and configure vn-segment

```
vlan 101
  vn-segment 900001
```

- Create overlay VRF VLAN and configure vn-segment

```
vlan 101
  vn-segment 900001
```

- Configure Core-facing SVI for VXLAN routing

```
interface vlan101
 no shutdown
  vrf member vxlan-900001
  ip forward
  no ip redirects
  ipv6 address use-link-local-only
  no ipv6 redirects
```

- Create VLAN and provide mapping to VXLAN

```
vlan 1001
  vn-segment 2001001
vlan 1002
  vn-segment 2001002
```

- Create VRF and configure VNI

```
vrf context vxlan-900001
  vni 900001
  rd auto
```



Note The **rd auto** and **route-target** commands are automatically configured unless one or more are entered as overrides.

```
\
address-family ipv4 unicast
      route-target both auto
      route-target both auto evpn
    address-family ipv6 unicast
      route-target both auto
      route-target both auto evpn
```

- Create server facing SVI and enable distributed anycast-gateway.

```
interface vlan1001
  no shutdown
  vrf member vxlan-900001
  ip address 4.1.1.1/24 tag 54321
  ipv6 address 4:1:0:1::1/64 tag 54321
  fabric forwarding mode anycast-gateway

interface vlan1002
  no shutdown
  vrf member vxlan-900001
  ip address 4.2.2.1/24 tag 54321
  ipv6 address 4:2:0:1::1/64 tag 54321
```

```
  fabric forwarding mode anycast-gateway
```

- Configure ACL TCAM region for ARP suppression

Note The **hardware access-list tcam region arp-ether 256 double-wide** command is not needed for Cisco Nexus 9300-FX/FX2/FX3 and 9300-GX platform switches.

```
hardware access-list tcam region arp-ether 256 double-wide
```

Note You can choose either of the following two options for creating the NVE interface. Use Option 1 for a small number of VNIs. Use Option 2 to leverage the simplified configuration mode.

Create the network virtualization endpoint (NVE) interface

Option 1

```
interface nve1
no shutdown
  source-interface loopback1
  host-reachability protocol bgp
  member vni 900001 associate-vrf
  member vni 2001001
    mcast-group 239.0.0.1
  member vni 2001002
    mcast-group 239.0.0.1
```

Option 2

```
interface nve1
  source-interface loopback1
  host-reachability protocol bgp
  global mcast-group 239.0.0.1 L2
  member vni 2001001
  member vni 2001002
  member vni 2001007-2001010
```

- Configure interfaces for hosts/servers

```
interface Ethernet1/47
  switchport
  switchport access vlan 1002

interface Ethernet1/48
  switchport
  switchport access vlan 1001
```

- Configure BGP

```
router bgp 65535
  router-id 30.1.1.1
  neighbor 10.1.1.1 remote-as 65535
    update-source loopback0
    address-family l2vpn evpn
      send-community both
  neighbor 20.1.1.1 remote-as 65535
    update-source loopback0
    address-family l2vpn evpn
      send-community both
  vrf vxlan-900001
    address-family ipv4 unicast
      redistribute direct route-map HOST-SVI
    address-family ipv6 unicast
      redistribute direct route-map HOST-SVI
```

Note The following commands in EVPN mode do not need to be entered.

```
evpn
  vni 2001001 l2
  vni 2001002 l2
```

Note The **rd auto** and **route-target auto** commands are automatically configured unless one or more are entered as overrides.

```
rd auto
    route-target import auto
    route-target export auto
```

Note The **rd auto** and **route-target** commands are automatically configured unless you want to use them to override the **import** or **export** options.

Note The following commands in EVPN mode do not need to be entered.

```
evpn
  vni 2001001 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 2001002 l2
    rd auto
    route-target import auto
    route-target export auto
```

- Leaf (9396-B)
  - Enable the EVPN control plane

```
nv overlay evpn
```

- Enable the relevant protocols

```
feature ospf
feature bgp
feature pim
feature interface-vlan
```

- Enable VxLAN with distributed anycast-gateway using BGP EVPN

```
feature vn-segment-vlan-based
feature nv overlay
fabric forwarding anycast-gateway-mac 0000.2222.3333
```

- Enabling OSPF for underlayrouting

```
router ospf 1
```

- Configure Loopback for local Router ID, PIM, and BGP

```
interface loopback0
  ip address 40.1.1.1/32
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
```

- Configure Loopback for local VTEP IP

```
interface loopback1
  ip address 44.1.1.1/32
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
```

- Configure interfaces for Spine-leaf interconnect

```
interface Ethernet2/2
  no switchport
  ip address 192.168.3.22/24
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  no shutdown

interface Ethernet2/3
  no switchport
  ip address 192.168.4.23/24
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  shutdown
```

- Configure route-map to Redistribute Host-SVI (Silent Host)

```
route-map HOST-SVI permit 10
  match tag 54321
```

- Configure PIM RP

```
ip pim rp-address 100.1.1.1 group-list 224.0.0.0/4
```

- Create VLANs

```
vlan 1001-1002
```

- Create overlay VRF VLAN and configure vn-segment

```
vlan 101
  vn-segment 900001
```

- Configure Core-facing SVI for VXLAN routing

```
interface vlan101
 no shutdown
  vrf member vxlan-900001
  ip forward
  no ip redirects
  ipv6 address use-link-local-only
  no ipv6 redirects
```

- Create VLAN and provide mapping to VXLAN

```
vlan 1001
  vn-segment 2001001
vlan 1002
  vn-segment 2001002
```

- Create VRF and configure VNI

```
vrf context vxlan-900001
  vni 900001
  rd auto
```



Note The **rd auto** and **route-target** commands are automatically configured unless one or more are entered as overrides.

```
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
  address-family ipv6 unicast
    route-target both auto
    route-target both auto evpn
```

- Create server facing SVI and enable distributed anycast-gateway

```
interface vlan1001
  no shutdown
  vrf member vxlan-900001
  ip address 4.1.1.1/24
  ipv6 address 4:1:0:1::1/64
  fabric forwarding mode anycast-gateway

interface vlan1002
  no shutdown
  vrf member vxlan-900001
  ip address 4.2.2.1/24
```

```
  ipv6 address 4:2:0:1::1/64
  fabric forwarding mode anycast-gateway
```

- Configure ACL TCAM region for ARP suppression

Note The **hardware access-list tcam region arp-ether 256 double-wide** command is not needed for Cisco Nexus 9300-FX/FX2/FX3 and 9300-GX platform switches.

```
hardware access-list tcam region arp-ether 256 double-wide
```

Note You can choose either of the following two command procedures for creating the NVE interfaces. Use Option 1 for a small number of VNIs. Use Option 2 to leverage the simplified configuration mode.

Create the network virtualization endpoint (NVE) interface

Option 1

```
interface nve1
  no shutdown
  source-interface loopback1
  host-reachability protocol bgp
  member vni 900001 associate-vrf
  member vni 2001001
    mcast-group 239.0.0.1
  member vni 2001002
    mcast-group 239.0.0.1
```

Option 2

```
interface nve1
  interface nve1
  source-interface loopback1
  host-reachability protocol bgp
  global mcast-group 239.0.0.1 L2
  member vni 2001001
  member vni 2001002
  member vni 2001007-2001010
```

- Configure interfaces for hosts/servers

```
interface Ethernet1/47
  switchport
  switchport access vlan 1002

interface Ethernet1/48
  switchport
```

```
    switchport access vlan 1001
```

- Configure BGP

```
 router bgp 65535
   router-id 40.1.1.1
   neighbor 10.1.1.1 remote-as 65535
     update-source loopback0
     address-family l2vpn evpn
       send-community both
   neighbor 20.1.1.1 remote-as 65535
     update-source loopback0
     address-family l2vpn evpn
       send-community both
   vrf vxlan-900001
   vrf vxlan-900001
     address-family ipv4 unicast
       redistribute direct route-map HOST-SVI
     address-family ipv6 unicast
       redistribute direct route-map HOST-SVI
```

Note The following commands in EVPN mode do not need to be entered.

```
evpn
  vni 2001001 l2
  vni 2001002 l2
```

Note The **rd auto** and **route-target** commands are automatically configured unless one or more are entered as overrides.

```
rd auto
    route-target import auto
    route-target export auto
```

Note The following commands in EVPN mode do not need to be entered.

```
evpn
  vni 2001001 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 2001002 l2
    rd auto
    route-target import auto
    route-target export auto
```

- Configure interface vlan on Border Gateway (BGW)

```
interface vlan101
  no shutdown
  vrf member evpn-tenant-3103101
  no ip redirects
```

```
  ip address 101.1.0.1/16
  ipv6 address cafe:101:1::1/48
  no ipv6 redirects
  fabric forwarding mode anycast-gateway
```

Note When you have iBGP session between BGWs and EBGP fabric is used, you need to configure the route-map to make VIP or VIP_R route advertisement with higher AS-PATH when local VIP or VIP_R is down (due to reload or fabric link flap). A sample route-map configuration is provided below. In this example 192.0.2.1 is VIP address and 198.51.100.1 is BGP VIP route's nexthop learned from same BGW site.

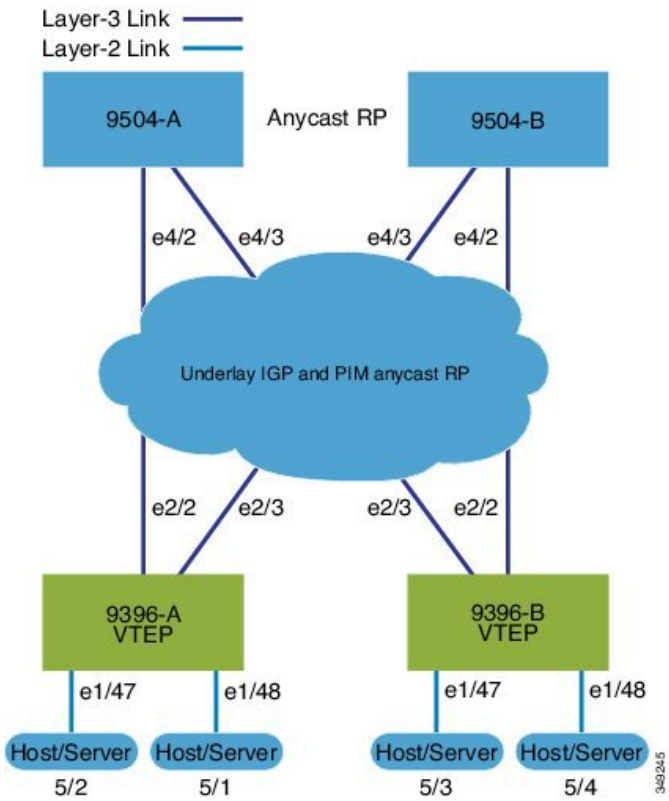
```
ip prefix-list vip_ip seq 5 permit 192.0.2.1/32
ip prefix-list vip_route_nh seq 5 permit 198.51.100.1/32

route-map vip_ip permit 5
  match ip address prefix-list vip_ip
  match ip next-hop prefix-list vip_route_nh
  set as-path prepend 5001 5001 5001
route-map vip_ip permit 10
```

Example of VXLAN BGP EVPN (eBGP)

An example of a VXLAN BGP EVPN (eBGP):

Figure 15: VXLAN BGP EVPN Topology (eBGP)



eBGP between Spine and Leaf

- Spine (9504-A)
  - Enable the EVPN control plane

```
nv overlay evpn
```

  - Enable the relevant protocols

```
feature bgp
feature pim
```

  - Configure Loopback for local Router ID, PIM, and BGP

```
interface loopback0
  ip address 10.1.1.1/32 tag 12345
  ip pim sparse-mode
```

  - Configure Loopback for Anycast RP

```
interface loopback1
  ip address 100.1.1.1/32 tag 12345
  ip pim sparse-mode
```

  - Configure Anycast RP

```
ip pim rp-address 100.1.1.1 group-list 224.0.0.0/4
ip pim anycast-rp 100.1.1.1 10.1.1.1
ip pim anycast-rp 100.1.1.1 20.1.1.1
```

  - Configure route-map used by eBGP for Spine

```
route-map NEXT-HOP-UNCH permit 10
  set ip next-hop unchanged
```

  - Configure route-map to Redistribute Loopback

```
route-map LOOPBACK permit 10
  match tag 12345
```

  - Configure interfaces for Spine-leaf interconnect

```
interface Ethernet4/2
  ip address 192.168.1.42/24
  ip pim sparse-mode
  no shutdown

interface Ethernet4/3
  ip address 192.168.2.43/24
  ip pim sparse-mode
  no shutdown
```

  - Configure the BGP overlay for the EVPN address family.

```
router bgp 100
  router-id 10.1.1.1
  address-family l2vpn evpn
    nexthop route-map NEXT-HOP-UNCH
    retain route-target all
```

```
 neighbor 30.1.1.1 remote-as 200
   update-source loopback0
   ebgp-multihop 3
   address-family l2vpn evpn
     send-community both
     disable-peer-as-check
     route-map NEXT-HOP-UNCH out
 neighbor 40.1.1.1 remote-as 200
   update-source loopback0
   ebgp-multihop 3
   address-family l2vpn evpn
     send-community both
     disable-peer-as-check
     route-map NEXT-HOP-UNCH out
```

- Configure BGP underlay for the IPv4 unicast address family.

```
address-family ipv4 unicast
   redistribute direct route-map LOOPBACK
 neighbor 192.168.1.22 remote-as 200
   update-source ethernet4/2
   address-family ipv4 unicast
     allowas-in
     disable-peer-as-check
 neighbor 192.168.2.23 remote-as 200
   update-source ethernet4/3
   address-family ipv4 unicast
     allowas-in
     disable-peer-as-check
```

- Spine (9504-B)
  - Enable the EVPN control plane

```
nv overlay evpn
```

  - Enable the relevant protocols

```
feature bgp
feature pim
```

  - Configure Loopback for local Router ID, PIM, and BGP

```
interface loopback0
  ip address 20.1.1.1/32 tag 12345
  ip pim sparse-mode
```

  - Configure Loopback for AnycastRP

```
interface loopback1
  ip address 100.1.1.1/32 tag 12345
  ip pim sparse-mode
```

  - Configure Anycast RP

```
ip pim rp-address 100.1.1.1 group-list 224.0.0.0/4
ip pim anycast-rp 100.1.1.1 10.1.1.1
ip pim anycast-rp 100.1.1.1 20.1.1.1
```

- Configure route-map used by eBGP for Spine

```
route-map NEXT-HOP-UNCH permit 10
  set ip next-hop unchanged
```

- Configure route-map to Redistribute Loopback

```
route-map LOOPBACK permit 10
  match tag 12345
```

- Configure interfaces for Spine-leaf interconnect

```
interface Ethernet4/2
  no switchport
  ip address 192.168.3.42/24
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  no shutdown

interface Ethernet4/3
  no switchport
  ip address 192.168.4.43/24
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  shutdown
```

- Configure BGP overlay for the EVPN address family

```
router bgp 100
  router-id 20.1.1.1
  address-family l2vpn evpn
    nexthop route-map NEXT-HOP-UNCH
    retain route-target all
  neighbor 30.1.1.1 remote-as 200
    update-source loopback0
    ebgp-multihop 3
    address-family l2vpn evpn
      send-community both
      disable-peer-as-check
      route-map NEXT-HOP-UNCH out
  neighbor 40.1.1.1 remote-as 200
    update-source loopback0
    ebgp-multihop 3
    address-family l2vpn evpn
      send-community both
      disable-peer-as-check
      route-map NEXT-HOP-UNCH out
```

- Configure the BGP underlay for the IPv4 unicast address family.

```
  address-family ipv4 unicast
    redistribute direct route-map LOOPBACK
  neighbor 192.168.3.22 remote-as 200
    update-source ethernet4/2
    address-family ipv4 unicast
```

```
    allowas-in
    disable-peer-as-check
  neighbor 192.168.4.43 remote-as 200
    update-source ethernet4/3
    address-family ipv4 unicast
      allowas-in
      disable-peer-as-check
```

- Leaf (9396-A)
  - Enable the EVPN control plane.

    ```
    nv overlay evpn
    ```

  - Enable the relevant protocols.

    ```
    feature bgp
    feature pim
    feature interface-vlan
    ```

  - Enable VXLAN with distributed anycast-gateway using BGP EVPN.

    ```
    feature vn-segment-vlan-based
    feature nv overlay
    fabric forwarding anycast-gateway-mac 0000.2222.3333
    ```

  - Enabling OSPF for underlay routing.

    ```
    router ospf 1
    ```

  - Configure Loopback for local Router ID, PIM, and BGP.

    ```
    interface loopback0
      ip address 30.1.1.1/32
      ip pim sparse-mode
    ```

  - Configure Loopback for VTEP.

    ```
    interface loopback1
      ip address 33.1.1.1/32
      ip pim sparse-mode
    ```

  - Configure interfaces for Spine-leaf interconnect.

    ```
    interface Ethernet2/2
      no switchport
      ip address 192.168.1.22/24
      ip pim sparse-mode
      no shutdown

    interface Ethernet2/3
      no switchport
      ip address 192.168.4.23/24
      ip pim sparse-mode
      shutdown
    ```

  - Configure route-map to Redistribute Host-SVI (Silent Host).

    ```
    route-map HOST-SVI permit 10
      match tag 54321
    ```

- Enable PIM RP.

```
ip pim rp-address 100.1.1.1 group-list 224.0.0.0/4
```

- Create VLANs.

```
vlan 1001-1002
```

- Create overlay VRF VLAN and configure vn-segment.

```
vlan 101
  vn-segment 900001
```

- Configure core-facing SVI for VXLAN routing.

```
interface vlan101
 no shutdown
  vrf member vxlan-900001
  ip forward
  no ip redirects
  ipv6 address use-link-local-only
  no ipv6 redirects
```

- Create VLAN and provide mapping toVXLAN.

```
vlan 1001
  vn-segment 2001001
vlan 1002
  vn-segment 2001002
```

- Create VRF and configure VNI

```
vrf context vxlan-900001
  vni 900001
  rd auto
```



Note The **rd auto** and **route-target** commands are automatically configured unless one or more are entered as overrides.

```
    address-family ipv4 unicast
      route-target both auto
      route-target both auto evpn
    address-family ipv6 unicast
      route-target both auto
      route-target both auto evpn
```

- Create server facing SVI and enable distributed anycast-gateway

```
interface vlan1001
  no shutdown
  vrf member vxlan-900001
  ip address 4.1.1.1/24 tag 54321
```

```
  ipv6 address 4:1:0:1::1/64 tag 54321
  fabric forwarding mode anycast-gateway

interface vlan1002
  no shutdown
  vrf member vxlan-900001
  ip address 4.2.2.1/24 tag 54321
  ipv6 address 4:2:0:1::1/64 tag 54321
  fabric forwarding mode anycast-gateway
```

- Configure ACL TCAM region for ARP suppression



Note The **hardware access-list tcam region arp-ether 256 double-wide** command is not needed for Cisco Nexus 9300-FX/FX2/FX3 and 9300-GX platform switches.

```
hardware access-list tcam region arp-ether 256 double-wide
```



Note You can choose either of the following two options for creating the NVE interface. Use Option 1 for a small number of VNIs. Use Option 2 to leverage the simplified configuration mode.

Create the network virtualization endpoint (NVE) interface

Option 1

```
interface nve1
  no shutdown
  source-interface loopback1
  host-reachability protocol bgp
  member vni 900001 associate-vrf
  member vni 2001001
    mcast-group 239.0.0.1
  member vni 2001002
    mcast-group 239.0.0.1
```

Option 2

```
interface nve1
  source-interface loopback1
  host-reachability protocol bgp
  global mcast-group 239.0.0.1 L2
  member vni 2001001
  member vni 2001002
  member vni 2001007-2001010
```

- Configure interfaces for hosts/servers.

```
interface Ethernet1/47
  switchport
  switchport access vlan 1002

interface Ethernet1/48
  switchport
  switchport access vlan 1001
```

- Configure BGP underlay for the IPv4 unicast address family.

```
router bgp 200
  router-id 30.1.1.1
  address-family ipv4 unicast
    redistribute direct route-map LOOPBACK
  neighbor 192.168.1.42 remote-as 100
    update-source ethernet2/2
    address-family ipv4 unicast
      allowas-in
      disable-peer-as-check
  neighbor 192.168.4.43 remote-as 100
    update-source ethernet2/3
    address-family ipv4 unicast
      allowas-in
      disable-peer-as-check
```

- Configure BGP overlay for the EVPN address family.

```
  address-family l2vpn evpn
    nexthop route-map NEXT-HOP-UNCH
    retain route-target all
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    ebgp-multihop 3
    address-family l2vpn evpn
      send-community both
      disable-peer-as-check
      route-map NEXT-HOP-UNCH out
  neighbor 20.1.1.1 remote-as 100
    update-source loopback0
    ebgp-multihop 3
    address-family l2vpn evpn
      send-community both
      disable-peer-as-check
      route-map NEXT-HOP-UNCH out
  vrf vxlan-900001
```



Note The following commands in EVPN mode do not need to be entered.

```
evpn
  vni 2001001 l2
  vni 2001002 l2
```

Note The **rd auto** and **route-target auto** commands are automatically configured unless one or more are entered as overrides.

```
rd auto
route-target import auto
route-target export auto
```

Note The following commands in EVPN mode do not need to be entered.

```
evpn
  vni 2001001 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 2001002 l2
    rd auto
    route-target import auto
    route-target export auto
```

- Leaf (9396-B)
  - Enable the EVPN control plane.

    ```
    nv overlay evpn
    ```

  - Enable the relevant protocols.

    ```
    feature bgp
    feature pim
    feature interface-vlan
    ```

  - Enable VXLAN with distributed anycast-gateway using BGP EVPN.

    ```
    feature vn-segment-vlan-based
    feature nv overlay
    fabric forwarding anycast-gateway-mac 0000.2222.3333
    ```

  - Enabling OSPF for underlay routing.

    ```
    router ospf 1
    ```

  - Configure Loopback for local Router ID, PIM, and BGP.

    ```
    interface loopback0
      ip address 40.1.1.1/32
      ip pim sparse-mode
    ```

  - Configure Loopback for VTEP.

    ```
    interface loopback1
      ip address 44.1.1.1/32
      ip pim sparse-mode
    ```

  - Configure interfaces for Spine-leaf interconnect.

```
interface Ethernet2/2
  no switchport
  ip address 192.168.3.22/24
  ip pim sparse-mode
  no shutdown

interface Ethernet2/3
  no switchport
  ip address 192.168.2.23/24
  ip pim sparse-mode
  shutdown
```

- Configure route-map to Redistribute Host-SVI (Silent Host).

```
route-map HOST-SVI permit 10
  match tag 54321
```

- Enable PIM RP

```
ip pim rp-address 100.1.1.1 group-list 224.0.0.0/4
```

- Create VLANs

```
vlan 1001-1002
```

- Create overlay VRF VLAN and configure vn-segment.

```
vlan 101
  vn-segment 900001
```

- Configure core-facing SVI for VXLAN routing.

```
interface vlan101
 no shutdown
  vrf member vxlan-900001
  ip forward
  no ip redirects
  ipv6 address use-link-local-only
  no ipv6 redirects
```

- Create VLAN and provide mapping to VXLAN.

```
vlan 1001
  vn-segment 2001001
vlan 1002
  vn-segment 2001002
```

- Create VRF and configure VNI

```
vrf context vxlan-900001
  vni 900001
  rd auto
```



Note The following commands are automatically configured unless one or more are entered as overrides.

```
    address-family ipv4 unicast
      route-target both auto
      route-target both auto evpn
    address-family ipv6 unicast
      route-target both auto
      route-target both auto evpn
```

- Create server facing SVI and enable distributed anycast-gateway.

```
interface vlan1001
  no shutdown
  vrf member vxlan-900001
  ip address 4.1.1.1/24 tag 54321
  ipv6 address 4:1:0:1::1/64 tag 54321
  fabric forwarding mode anycast-gateway

interface vlan1002
  no shutdown
  vrf member vxlan-900001
  ip address 4.2.2.1/24 tag 54321
  ipv6 address 4:2:0:1::1/64 tag 54321
  fabric forwarding mode anycast-gateway
```

- Configure ACL TCAM region for ARP suppression

Note The **hardware access-list tcam region arp-ether 256 double-wide** command is not needed for Cisco Nexus 9300-FX/FX2/FX3 and 9300-GX platform switches.

```
hardware access-list tcam region arp-ether 256 double-wide
```

Note You can choose either of the following two procedures for creating the NVE interface. Use Option 1 for a small number of VNIs. Use Option 2 to leverage the simplified configuration mode.

Create the network virtualization endpoint (NVE) interface.

Option 1

```
interface nve1
  no shutdown
  source-interface loopback1
  host-reachability protocol bgp
  member vni 900001 associate-vrf
  member vni 2001001
    mcast-group 239.0.0.1
  member vni 2001002
    mcast-group 239.0.0.1
```

Option 2

```
interface nve1
  source-interface loopback1
  host-reachability protocol bgp
  global mcast-group 239.0.0.1 L2
  member vni 2001001
  member vni 2001002
  member vni 2001007-2001010
```

- Configure interfaces for hosts/servers

```
interface Ethernet1/47
  switchport
  switchport access vlan 1002

interface Ethernet1/48
  switchport
  switchport access vlan 1001
```

- Configure BGP underlay for the IPv4 unicast address family.

```
router bgp 200
  router-id 40.1.1.1
  address-family ipv4 unicast
    redistribute direct route-map LOOPBACK
  neighbor 192.168.3.42 remote-as 100
    update-source ethernet2/2
    address-family ipv4 unicast
      allowas-in
      disable-peer-as-check
  neighbor 192.168.2.43 remote-as 100
    update-source ethernet2/3
    address-family ipv4 unicast
      allowas-in
      disable-peer-as-check
```

- Configure BGP overlay for the EVPN address family.

```
  address-family l2vpn evpn
    nexthop route-map NEXT-HOP-UNCH
    retain route-target all
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    ebgp-multihop 3
    address-family l2vpn evpn
      send-community both
      disable-peer-as-check
      route-map NEXT-HOP-UNCH out
  neighbor 20.1.1.1 remote-as 100
    update-source loopback0
    ebgp-multihop 3
    address-family l2vpn evpn
      send-community both
      disable-peer-as-check
      route-map NEXT-HOP-UNCH out
  vrf vxlan-900001
```

Note The following commands in EVPN mode do not need to be entered.

```
evpn
  vni 2001001 l2
  vni 2001002 l2
```

Note The **rd auto** and **route-target auto** commands are automatically configured unless one or more are entered as overrides.

```
rd auto
route-target import auto
route-target export auto
```

Note The following commands in EVPN mode do not need to be entered.

```
evpn
  vni 2001001 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 2001002 l2
    rd auto
    route-target import auto
    route-target export auto
```

Example Show Commands

- **show nve peers**

```
9396-B# show nve peers
Interface Peer-IP          State LearnType Uptime   Router-Mac
--------- ---------------  ----- --------- -------- -----------------
nve1      30.1.1.1         Up    CP        00:00:38 6412.2574.9f27
```

- **show nve vni**

```
9396-B# show nve vni
Codes: CP - Control Plane        DP - Data Plane
       UC - Unconfigured

Interface VNI      Multicast-group   State Mode Type [BD/VRF]      Flags
--------- -------- ----------------- ----- ---- ------------------ -----
nve1      900001   n/a               Up    CP   L3 [vxlan-900001]
nve1      2001001  225.4.0.1         Up    CP   L2 [1001]
nve1      2001002  225.4.0.1         Up    CP   L2 [1002]
```

- **show ip arp suppression-cache detail**

```
9396-B# show ip arp suppression-cache detail

Flags: + - Adjacencies synced via CFSoE
       L - Local Adjacency
       R - Remote Adjacency
       L2 - Learnt over L2 interface

Ip Address      Age      Mac Address    Vlan Physical-ifindex    Flags

4.1.1.54        00:06:41 0054.0000.0000 1001 Ethernet1/48        L
4.1.1.51        00:20:33 0051.0000.0000 1001 (null)              R
4.2.2.53        00:06:41 0053.0000.0000 1002 Ethernet1/47        L
4.2.2.52        00:20:33 0052.0000.0000 1002 (null)              R
```



Note The **show vxlan interface** command is not supported for the Cisco Nexus 9300-FX/FX2/FX3, and 9300-GX platform switches.

- **show vxlan interface**

```
9396-B# show vxlan interface
Interface       Vlan    VPL Ifindex     LTL             HW VP
=========       ====    ============    ===             =====
Eth1/47         1002    0x4c07d22e      0x10000         5697
Eth1/48         1001    0x4c07d02f      0x10001         5698
```

- **show bgp l2vpn evpn summary**

```
leaf3# show bgp l2vpn evpn summary
BGP summary information for VRF default, address family L2VPN EVPN
BGP router identifier 40.0.0.4, local AS number 10
BGP table version is 60, L2VPN EVPN config peers 1, capable peers 1
21 network entries and 21 paths using 2088 bytes of memory
BGP attribute entries [8/1152], BGP AS path entries [0/0]
BGP community entries [0/0], BGP clusterlist entries [1/4]

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down
State/PfxRcd
40.0.0.1        4    10    8570    8565       60    0    0    5d22h 6
leaf3#
```

- **show bgp l2vpn evpn**

```
leaf3# show bgp l2vpn evpn
BGP routing table information for VRF default, address family L2VPN EVPN
BGP table version is 60, local router ID is 40.0.0.4
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid,
>-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist,
I-injected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup

   Network            Next Hop            Metric     LocPrf     Weight Path
Route Distinguisher: 40.0.0.2:32868
*>i[2]:[0]:[10001]:[48]:[0000.8816.b645]:[0]:[0.0.0.0]/216
                      40.0.0.2                          100          0 i
*>i[2]:[0]:[10001]:[48]:[0011.0000.0034]:[0]:[0.0.0.0]/216
                      40.0.0.2                          100          0 i
```

- **show l2route evpn mac all**

```
leaf3# show l2route evpn mac all
Topology    Mac Address    Prod   Next Hop (s)
----------- -------------- ------ ---------------
101         0000.8816.b645 BGP    40.0.0.2
101         0001.0000.0033 Local  Ifindex 4362086
101         0001.0000.0035 Local  Ifindex 4362086
101         0011.0000.0034 BGP    40.0.0.2
```

- **show l2route evpn mac-ip all**

```
leaf3# show l2route evpn mac-ip all
Topology ID Mac Address    Prod Host IP                 Next Hop (s)
----------- -------------- ---- ---------------------------------------------------------
101         0011.0000.0034 BGP  5.1.3.2                      40.0.0.2
102         0011.0000.0034 BGP  5.1.3.2                      40.0.0.2
```

- **show l2route evpn startup-route all [detail]**

```
Flags - (V4):IPv4 Bcast address (V6):IPv6 Bcast address
(D):Del Pending (S):Stale (C):Clear
Topology ID Source Group (SG)
                           Flags       NFN Bitmap
-----------
-------------------------------------------------------------------------------------------------------
 ----------- ----------
100         (0.0.0.0, 255.255.255.255)
                           V4          8
            Delivery SG: (11.1.1.2, 21.1.1.2)
            DCI Delivery SG: (11.1.1.3, 21.1.1.3)
1162        (::, ff00::)
                           V6          8
            Delivery SG: (4901:11:11::11:11, ff1e:225:0:1::1)
            DCI Delivery SG: (4901:11:11::11:11, ::)
```

- **show l2route evpn startup-route evi <vpn-id> [detail]**

```
Flags - (V4):IPv4 Bcast address (V6):IPv6 Bcast address
(D):Del Pending (S):Stale (C):Clear
Topology ID Source Group (SG)
                           Flags       NFN Bitmap
-----------
-------------------------------------------------------------------------------------------------------
 ----------- ----------
1162        (::, ff00::)
                           V6          8
            Delivery SG: (4901:11:11::11:11, ff1e:225:0:1::1)
            DCI Delivery SG: (4901:11:11::11:11, ::)
```

- **show l2route evpn ethernet-segment all detail**

```
ESI                         Orig Rtr. IP Addr   Prod    Ifindex NFNBitmapDF Election
--------------------------  ----------------- ----    ------- ----------  --------
0300.0000.0000.0d00.0309    2.11.51.1           VXLAN   nve     11024       Modulus

0300.0000.0000.0dff.0309    2.11.51.1           VXLAN   nve     11024       Modulus

0300.0000.0000.0d00.0309    2.11.51.4           BGP     N/A     512         Per Flow
0300.0000.0000.0dff.0309    2.11.51.4           BGP     N/A     512         Per Flow
```

Configuring ND Suppression

ND Suppression on the Overlay

Multicast Neighbor Solicitation packets from host to another host are flooded over the BGP/EVPN VXLAN Core when hosts are behind two different VXLAN peers.

The ND Suppression cache is built by:

- Snooping NS request in the hosts and populating the ND Suppression cache with source IP and MAC bindings in the request.
- Learning IPv6-Host or MAC address information through BGP EVPN MAC route advertisements.

With ND Suppression, for host to host communication behind two different VXLAN peers, if the remote host is not learned in the suppression cache initially, then NS packets are flooded over the BGP/EVPN VXLAN Core. However, once the ND Suppression cache on a switch S1 is populated with the remote host, any subsequent Neighbor Solicitation request packet for the remote host in the hosts behind S1 are proxied by the Switch S1 thereby preventing the flooding of Neighbor Solicitation packet over the BGP-EVPN/VXLAN core

For ND Suppression cache scale values, see *Cisco Nexus 9000 Series NX-OS Verified Scalability Guide*.

Guidelines and Limitations for ND Suppression

ND suppression has the following configuration guidelines and limitations:

- Beginning with Cisco NX-OS Release 10.3(1)F, the Cisco Nexus 9300-X Cloud Scale switches supports the ND Suppression feature only on plain BGP EVPN.
- Beginning with Cisco NX-OS Release 10.5(2)F, ND suppression is supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.
- ND Suppression is not supported with BGP-EVPN feature variants like Multisite, Virtual MCT, IRB, Centralized Gateway, Firewall Clustering, vPC.
- For link-local addresses of hosts, ND Suppression is not supported and instead multicast NS for link local address of hosts are flooded over the core of BGP EVPN VXLAN network.
- ND Suppression gets enabled on all VNIs on which suppress-arp is enabled.
- ND Suppression CLI knob must be enabled only under the following conditions:
  - The suppress-arp must be enabled on a VNI and there must be an SVI associated with this VNI/VLAN. Also, this SVI must be in up state and must have both IPv4 and IPv6 address enabled.
  - ND Suppression will not work in the following conditions:
    - If SVI not present for the VLAN/VNI on which suppress-arp/suppress nd is enabled.
    - If SVI associated with VLAN VNI on which suppress-arp/suppress nd is enabled is down.
    - If SVI associated with VLAN/VNI on which suppress-arp/suppress nd is enabled has only IPv4 and no IPv6 address.

- If SVI associated with VLAN/VNI on which suppress-arp/suppress nd is enabled has only IPv6 and no IPv4 address.

  In all the above conditions, host to host traffic can potentially be dropped.

- For ND Suppression VACL to work, increase the SUP TCAM size to 768 or above using the **hardware access-list tcam region sup-tcam 768** command.
- If the installed Cisco NX-OS switch does not support ND suppression, ensure that Anycast Gateway MAC addresses across sites are identical.

Configuring ND Suppression

This procedure describes how to enable/disable the ND suppression feature on the NVE interface.

Before you begin

Ensure that ARP suppression is enabled.

SUMMARY STEPS

1. **configure terminal**
2. **hardware access-list tcam region ing-sup 768**
3. **copy running-config startup-config**
4. **reload**
5. **configure terminal**
6. **interface nve 1**
7. **[no]suppress nd**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**hardware access-list tcam region ing-sup 768** **Example:** `switch# hardware access-list tcam region ing-sup 768`	Carves the Ingress SUP TCAM size to 768.
Step 3	**copy running-config startup-config** **Example:** `switch# copy running-config startup-config`	Copies the running configuration to the startup configuration.

	Command or Action	Purpose
Step 4	**reload** **Example:** `switch# reload`	Reloads the switch.
Step 5	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.

Step 6	**interface nve 1**
	Example:
	`switch(config)# interface nve 1`
	`switch(config-if-nve)#`
		Enters interface nve configuration mode.

Step 7	**[no]suppress nd** **Example:** `switch(config-if-nve)# suppress nd`	Configures ND Suppression for all ARP enabled VNIs. Option **no** disables the ND Suppression for all ARP enabled VNIs.

Note

- When global **suppress arp** command is configured, ND Suppression is enabled on all VNIs.
- When global **suppress arp** command is not configured and instead per VNI **suppress arp** command is configured, then ND Suppression is enabled on all VNIs on which ARP suppression is configured.
- When enabling suppress arp command on a vPC pair, ensure steps 1-4 on both peers are complete before enabling the feature.

Verifying the ND Suppression Configuration

To display the ND Suppression configuration information, enter one of the following commands:

Command	Purpose
show run nv overlay	Displays the ND suppression configuration status.
show nve vni	Displays whether the ND suppression config has been enabled for ARP enabled VNIs.
show nve internal export nve	Displays whether the ND suppression config has been enabled or not in SDB.
show nve internal export vni	Displays the ND suppression state per VNI in SDB.
show ipv6 nd suppression-cache detail command.	Displays the ICMPv6 cache entries that are present in local.

Command	Purpose
show ipv6 nd suppression-cache remote	Displays the ICMPv6 cache entries that are present in remote.
show ipv6 nd suppression-cache summary	Displays the IPv6 cache entries summary of both local and remote.
show ipv6 nd suppression-cache statistics	Displays the IPv6 ND suppression cache statistics.
show ipv6 nd suppression-cache vlan "vlan_id"	Displays the details of IPv6 ND Suppression cache entries for a particular VLAN.

The following example shows sample output for the **show run nv overlay** command:

```
switch(config-if-nve)# sh run nv overlay
!Command: show running-config nv overlay
!Running configuration last done at: Sat Mar 19 01:07:49 2022
!Time: Sat Mar 19 01:10:00 2022

version 10.2(3) Bios:version 07.68
feature nv overlay

vlan 101-110,200-203,500-501

interface nve1
  no shutdown
  host-reachability protocol bgp
  suppress nd
  global suppress-arp
```

The following example shows sample output for the **show nve vni** command:

```
switch(config-if-nve-vni)# sh nve vni
Codes: CP - Control Plane        DP - Data Plane
       UC - Unconfigured         SA - Suppress ARP
       S-ND Suppress ND
       SU - Suppress Unknown Unicast
       Xconn - Crossconnect
       MS-IR - Multisite Ingress Replication
       HYB - Hybrid IRB mode

Interface VNI      Multicast-group   State Mode Type [BD/VRF]      Flags
--------- -------- ----------------- ----- ---- ------------------ -----
nve1      5000     239.2.0.2         Up    CP   L2 [500]           SA S-ND
```

The following example shows sample output for the **show nve internal export nve** command:

```
switch(config-if-nve-vni)# sh nve internal export nve

NVE Interface information.
+-----------------------------------------------------------+
Interface: nve1, Admin State: Up,
   State: nve-intf-add-complete, Encap: vxlan
   Source interface: loopback3, VRF: default,
   Anycast-interface: <none>
   Mcast-routing src intf <none>
   Primary IP: 4.4.4.4, Secondary IP: 0.0.0.0,
   VNI-VRF: default, Allow-Src-Lpbk-Down: No,
   Advertise MAC route: No,
   Virtual-rMAC: 0000.0000.0000,
   Mcast-routing Primary IP: 0.0.0.0
   Suppress ND: 1
```

```
   Host-reachability: CP
   unknown-peer-forwarding-mode: disable
   VNI assignment mode: n/a
   Multisite bgw-if: <none> (ip: 0.0.0.0, admin/oper state: Down/Down)
    src-node-last-notify: None
    anycast-node-last-notify: None
    mcast-src-node-last-notify: None
    multi-src-node-last-notify: None

+-----------------------------------------------------------+
```

The following example shows sample output for the **show nve internal export vni** command:

```
switch(config-if-nve-vni)# sh nve internal export vni

NVE VNI Information.
+-----------------------------------------------------------+
 VNI: 5000 [500] Mgroup: 239.2.0.2 Provision-State: vni-add-complete
  Primary: 4.4.4.4 Secondary: 0.0.0.0 SRC-VRF: default
  Encap: vxlan Repl-mode: Mcast
  Suppress ARP: SP Suppress ND: Enabled Mode: CP, VNI-VRF: <FALSE>  [vrf-id 0] [vrf flags
0x0]
  Suppress Unknown-Unicast: FALSE
  X-connect : Disabled
  [VNI local configs] SA : TRUE, Mcast-group : TRUE, IR proto BGP: FALSE
  Config Src: CLI, VNI flags: 0x0
  Spine-AGW: Disabled, HYBRID: Disabled
  Multisite optimized IR: Disabled
  Multisite DCI Group Unknown Address

+-----------------------------------------------------------+
```

The following example shows sample output for the **show ipv6 nd suppression-cache detail** command:

```
switch(config)# show ipv6 nd suppression-cache detail

Flags: + - Adjacencies synced via CFSoE
       L - Local Adjacency
       R - Remote Adjacency
       L2 - Learnt over L2 interface
       PS - Added via L2RIB, Peer Sync
       RO - Dervied from L2RIB Peer Sync Entry

IPv6 Address      Age      Mac Address    Vlan Physical-ifindex    Flags    Remote Vtep
Addrs

172:11:1:1::51  00:00:18 acf2.c5f6.7641   11 Ethernet1/51        L
172:11:1:1::201 00:06:14 0000.0011.1111   11 (null)              R        30.100.1.1
172:11:1:1::101 00:06:14 74a0.2f1d.d481   11 (null)              R        10.10.11.11
```

The following example shows sample output for the **show ipv6 nd suppression-cache local** command:

```
switch(config)# show ipv6 nd suppression-cache local

Flags: + - Adjacencies synced via CFSoE
       L - Local Adjacency
       R - Remote Adjacency
       L2 - Learnt over L2 interface

Ip Address      Age      Mac Address    Vlan Physical-ifindex    Flags

172:11:1:1::51  00:00:23 acf2.c5f6.7641   11 Ethernet1/51        L
```

The following example shows sample output for the **show ipv6 nd suppression-cache remote** command:

```
switch(config)# show ipv6 nd suppression-cache remote

Flags: + - Adjacencies synced via CFSoE
       L - Local Adjacency
       R - Remote Adjacency
       L2 - Learnt over L2 interface
       PS - Added via L2RIB, Peer Sync
       RO - Dervied from L2RIB Peer Sync Entry

IPv6 Address      Age      Mac Address    Vlan Physical-ifindex    Flags    Remote Vtep
Addrs

172:11:1:1::201 00:06:24 0000.0011.1111   11 (null)              R        30.100.1.1
172:11:1:1::101 00:06:24 74a0.2f1d.d481   11 (null)              R        10.10.11.11
```

The following example shows sample output for the **show ipv6 nd suppression-cache statistics** command:

```
switch(config)# show ipv6 nd suppression-cache statistics

ND packet statistics for suppression-cache

Suppressed:

Total: 1
L3 mode :       Requests 1, Replies 1
                Flood ND Probe 0

Received:
Total: 1
 L3 mode:       NS 1, Non-local NA 0
                Non-local NS 0

Mobility Requests:
Total: 0
 L3 mode:       Remote-to-local 0, Local-to-remote 0
                Remote-to-remote 0

RARP Signal Refresh: 0

ND suppression-cache Local entry statistics
Adds 3, Deletes 0
```

The following example shows sample output for the **show ipv6 nd suppression-cache summary** command:

```
switch(config)# show ipv6 nd suppression-cache summary

IPV6 ND suppression-cache Summary
Remote              :2
Local               :1
Total               :3
```

The following example shows sample output for the **show ipv6 nd suppression-cache vlan "vlan_id"** command:

```
switch(config)# show ipv6 nd suppression-cache vlan 11

Flags: + - Adjacencies synced via CFSoE
       L - Local Adjacency
       R - Remote Adjacency
       L2 - Learnt over L2 interface
       PS - Added via L2RIB, Peer Sync
       RO - Dervied from L2RIB Peer Sync Entry

IPv6 Address      Age      Mac Address    Vlan Physical-ifindex    Flags    Remote Vtep
Addrs
```

```
172:11:1:1::51  00:00:40 acf2.c5f6.7641   11 Ethernet1/51        L
172:11:1:1::201 00:06:36 0000.0011.1111   11 (null)              R        30.100.1.1
172:11:1:1::101 00:06:36 74a0.2f1d.d481   11 (null)              R        10.10.11.11
```

VXLAN BGP EVPN with RFC 5549 Underlay

VXLAN BGP EVPN with RFC 5549 Underlay Overview

The VXLAN BGP EVPN with RFC 5549 underlay feature is introduced for deployments where you already have a VXLAN IPv4 network infrastructure and must integrate IPv6 functionality.

The RFC 5549 enables the advertisement of underlay IPv4 prefixes using BGP with an IPv6 address as the next hop, effectively allowing IPv4 connectivity to be established over an IPv6 underlay network.

Benefits of VXLAN BGP EVPN with RFC 5549 Underlay

- Helps your network against the exhaustion of IPv4 addresses.
- Enhances scalability and efficiency by integrating IPv6 into the VXLAN IPv4 network infrastructure.
- Offers simpler, more direct communication between devices.
- Provides sub-second convergence.
- Supports the border spine gateway to leaf topology.

Supported Platform and Release for VXLAN BGP EVPN with RFC 5549 Underlay

Feature	Release	Platform
VXLAN BGP EVPN with RFC 5549 Underlay on single-site	10.2(3)F	Cisco Nexus 9000 Series switches

Functionalities of VXLAN BGP EVPN with RFC 5549 Underlay

This section describes how the VXLAN BGP EVPN with RFC 5549 underlay functions on a single-site environment.

VTEP IPv4 Address Advertisement

The VTEP addresses are advertised in BGP with RFC 5549 underlay such that the IPv4 VTEP addresses are routed using the following IPv6 enabled interfaces.

- **Link-local address (LLA)**: In this type of BGP peering configuration, BGP uses the interface's link-local address to establish IPv6 sessions, allowing for the advertisement of IPv4 addresses through them. This approach eliminates the necessity for configuring global addresses on interfaces.
- **Global IPv6 address**: In this type of BGP peering configuration, BGP creates a standard IPv6 neighbor relationship using the global IPv6 address of the directly connected interface. Therefore, it is required to configure a global IPv6 address on the directly connected interface for the peering to be established.

VXLAN EVPN Services Advertisement over IPv4 or IPv6 Session

When VTEPs IPv4 reachability is advertised with RFC 5549 underlay, an EVPN IPv4 or EVPN IPv6 BGP session must be created to advertise VXLANv4 services on a single-site environment.

VXLAN BGP EVPN with RFC 5549 Underlay on Single-Site

On Single-Site, the VXLAN BGP EVPN with RFC 5549 Underlay supports node to node communication as shown in the diagram.

The VTEP IPv4 address is advertised using IPv6 address (LLA or Global address) as the next-hop address in RFC 5549 underlay.

Figure 16: Leaf and Spine Fabric (iBGP)

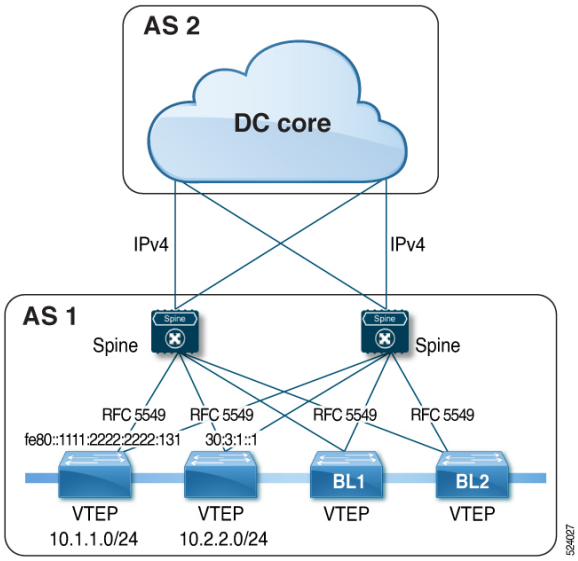
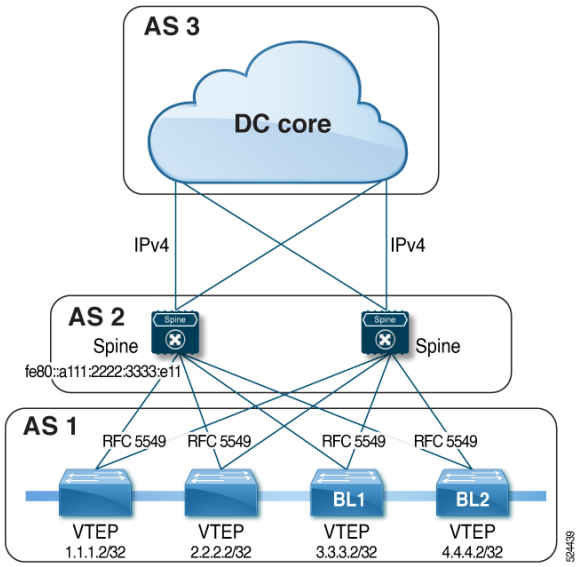


Figure 17: Leaf and Spine Fabric (eBGP)



The following use cases of EVPN VXLAN session over RFC5549 are supported:

1. **Use case 1:**
   - Underlay BGP session: BGP session uses IPv6 link-local or IPv6 global address.
   - Overlay EVPN session: EVPN session uses IPv4 between the loopbacks.
   - VTEP: IPv4 address.
2. **Use case 2:**
   - Underlay BGP session: BGP session uses IPv6 link-local or IPv6 global address.
   - Overlay EVPN session: EVPN session uses IPv6 between the loopbacks.
   - VTEP: IPv6 address.
3. **Use case 3:**
   - Underlay BGP session and Overlay EVPN session: Both uses Single TCP session over IPv6 link-local or IPv6 global address.
   - VTEP: IPv4 address.
4. **Use case 4:**
   - Underlay BGP session: BGP session uses IPv6 link-local or IPv6 global address.
   - Overlay EVPN session: EVPN session uses IPv6 between the loopbacks.
   - VTEP: IPv4 address.

Supported Combination of RFC 5549 and IPv4 over Core and DC Fabric

Use case	DC Fabric	Core	Supported Release
Single-Site	RFC 5549	IPv4	10.2(3)F

Guidelines for VXLAN BGP EVPN with RFC 5549 Underlay

- VTEP must be from the same address family of IPv4.
- VTEP loop-back address must be an IPv4 address.
- All the nodes including BGW, Core, and spine must be upgraded to Release 10.2(3)F or higher.
- Beginning with Cisco NX-OS Release 10.6(1)F, RFC5549 underlay is supported on Cisco Nexus 9364E-SG2-Q and 9364E-SG2-O switches.

Supported and Unsupported Features of VXLAN BGP EVPN with RFC 5549 Underlay

Following are the supported and unsupported features for VXLAN BGP EVPN with RFC 5549 Underlay on single-site:

Features	Supported Release	Supported or Not Supported
NGOAM	10.2(3)F	Supported
vPC	10.2(3)F	Supported
Ingress Replication	10.2(3)F	Supported
Overlay TCP Session	10.2(3)F	Supported
ARP Suppression	10.2(3)F	Supported
BFD	10.2(3)F	Supported
vPC VIP and PIP on leaf	10.2(3)F	Supported
IPv4 over RFC 5549 with default VRF	10.2(3)F	Supported
Overlay: eBGP EVPN IPv4 loopback peering with loopbacks	10.2(3)F	Supported
Route Leaking: Host addresses and between non default customer VRFs (On Leaf)	10.2(3)F	Supported
DSVNI	10.5(1)F	Supported
vMCT	10.5(1)F	Supported
Border Spine	10.5(2)F	Supported
EVPN IPv6 session with IPv4 VTEP over RFC5549 underlay	10.5(2)F	Supported
Flood and Learn deployment	10.5(3)F	Not Supported
IPv4 and IPv6 overlay multicast	10.5(3)F	Not Supported
vPC based multihoming	10.5(3)F	Not Supported
VLAN based Layer 3 VNI configuration	10.5(3)F	Not Supported
VXLAN IPv6 underlay	10.5(3)F	Not Supported
PIMv4 PIMv6 ASM /Bidir Multicast underlay	10.5(3)F	Not Supported

Features	Supported Release	Supported or Not Supported
Multicast underlay BUD node support	10.5(3)F	Not Supported
IGMP / MLD snooping	10.5(3)F	Not Supported
TRMv4/v6, and TRM data MDT support (ASIC limitation)	10.5(3)F	Not Supported
ND suppression	10.5(3)F	Not Supported
PVNF	10.5(3)F	Not Supported
VXLAN PBR support	10.5(3)F	Not Supported
VXLAN QoS policy	10.5(3)F	Not Supported
Base L3 QoS policies	10.5(3)F	Not Supported
VXLAN peer-based total packet/byte counters	10.5(3)F	Not Supported
Separate counters for unicast, multicast and broadcast traffic	10.5(3)F	Not Supported
DSVNI and route leak	10.5(3)F	Not Supported
Non-VXLAN access features • QinVNI, selective QinVNI • Port-VLAN mapping • Multitag • Xconnect • L2PT/VXLAN • Private VLAN • Port Security	10.5(3)F	Not Supported
VXLAN ESI multihoming	10.5(3)F	Not Supported
Firewall clustering with VXLAN	10.5(3)F	Not Supported
Southbound loop detection	10.5(3)F	Not Supported
SGACL	10.5(3)F	Not Supported
VXLAN Consistency Checker	10.5(3)F	Not Supported
VXLAN TE	10.5(3)F	Not Supported

Features	Supported Release	Supported or Not Supported
IPv4 over RFC 5549 with tenant VRF	-	Not Supported
IPv4 VNF: IPv4 services over IPv4 PE-CE BGP session	-	Not Supported
RFC 5549 VNF: IPv4 services over IPv6 PE-CE BGP session	-	Not Supported
IPv6 VNF: IPv6 services over IPv6 PE-CE BGP session	-	Not Supported
ND Suppression	-	Not Supported
Multicast/TRM	-	Not Supported
Underlay multicast	-	Not Supported
Policy Based Routing	-	Not Supported
New L3VNI config	-	Not Supported
Group Policy Option	-	Not Supported
First Hop Security	-	Not Supported
PVLAN over VXLAN	-	Not Supported

Configuring VXLAN BGP EVPN with RFC 5549 Underlay on Single-Site

You can perform the configuration of VXLAN BGP EVPN with RFC 5549 underlay using LLA or global IPv6 address on a single-site environment.

To configure VXLAN BGP EVPN with RFC 5549 underlay, follow these steps:

Procedure

Step 1 **configure terminal**

Example:

```
VTEP# config terminal
VTEP(config)#
```

Enters global configuration mode.

Step 2 **interface ethernet** *port*

Example:

```
VTEP(config)# interface ethernet1/2
VTEP(config-if)#
```

Enters Link local interface configuration mode for RFC 5549.

Step 3 **ip forward**

Example:

```
VTEP(config-if)# ip forward
```

Enables IPv4 based lookup even when the interface VLAN has no IP address defined.

Step 4 To generate link-local address, use any of the following options:

- Link-local IPv6 address

  OR
- Global IPv6 address

a) To configure LLA, use any of the following options:

- **ipv6 link-local** *LL_ipv6_address* (Manual option)

  OR
- **ipv6 address use-link-local-only** (Auto option)

  OR
- **ipv6 link-local use-bia** (Auto option)

Example:

```
VTEP(config-if)# ipv6 link-local fe80::1111:2222:2222:3101
```

OR

```
VTEP(config-if)# ipv6 address use-link-local-only
```

OR

```
VTEP(config-if)# ipv6 link-local use-bia
```

Configures the link-local IPv6 address for the interface based on the specified choice.

OR

a) Global IPv6 address: By default generates the LLA

Example:

ipv6 address *ipv6_address*

```
VTEP(config-if)# ipv6 address 2000:1:1::1/64
```

Configures the Global IPv6 address for the interface.

Step 5 **exit**

Example:

```
VTEP(config-if)# exit
VTEP(config)#
```

Exits the interface configuration mode.

Step 6 **router bgp** *as-number*

Example:

```
VTEP(config)# router bgp 1
VTEP(config-router)#
```

Enters BGP router configuration mode.

Step 7 **neighbor** [*ipv6_address* | *ethernet port*]

Example:

```
VTEP(config-router)# neighbor ethernet1/2
```

Configures a BGP neighbor for an interface.

Step 8 **remote-as** *value*

Example:

```
VTEP(config-router)# remote-as 2
VTEP(config-router-neighbor)#
```

Configures remote peer's autonomous system number for BGP neighbor.

Step 9 **address-family ipv4 unicast**

Example:

```
VTEP(config-router-neighbor)# address-family ipv4 unicast
VTEP(config-router-neighbor-af)#
```

Configures the IPv4 unicast address family as specified.

Step 10 **address-family ipv6 unicast**

Example:

```
VTEP(config-router-neighbor-af)# address-family ipv6 unicast
```

Configures the IPv6 unicast address family as specified.

Step 11 **address-family l2vpn evpn**

Example:

```
VTEP(config-router-neighbor-af)# address-family l2vpn evpn
```

Configures address family Layer 2 VPN EVPN under the BGP neighbor.

Step 12 (Optional) **Allowas-in** *number-of-occurrences-of-AS-number*

Example:

```
VTEP(config-router-neighbor-af)# Allowas-in 3
```

Only for eBGP deployment cases: Allows duplicate autonomous system (AS) numbers in the AS path. Configure this parameter on the leaf for eBGP when all leafs are using the same AS, but the spines have a different AS than leafs.

Step 13 **send-community both**

Example:

```
VTEP(config-router-neighbor-af)# send-community both
```

Configures community for BGP neighbors.

Step 14 **disable-peer-as-check**

Example:

```
VTEP(config-router-neighbor-af)# disable-peer-as-check
```

Disables checking the peer AS number during route advertisement. Configure this parameter on the spine for eBGP when all leafs are using the same AS but the spines have a different AS than leafs.

Note
This command is required for eBGP. For more information on eBGP configuration, see eBGP Underlay IP Network , on page 25.

Step 15 **router bgp** *as-number*

Example:

```
Spine(config)# router bgp 2
Spine(config-router)#
```

Enters BGP router configuration mode.

Step 16 **neighbor** [*ipv6_address* | *ethernet port*]

Example:

```
Spine(config-router)# neighbor ethernet1/1
Spine(config-router-neighbor)#
```

Configures a BGP neighbor for an interface.

Step 17 **remote-as** *value*

Example:

```
Spine(config-router-neighbor)# remote-as 1
```

Configures remote peer's autonomous system number for BGP neighbor.

Step 18 **address-family ipv4 unicast**

Example:

```
Spine(config-router-neighbor)# address-family ipv4 unicast
config-router-neighbor-af
```

Configures the IPv4 unicast address family as specified.

Step 19 **address-family ipv6 unicast**

Example:

```
Spine(config-router-neighbor-af)# address-family ipv6 unicast
```

Configures the IPv6 unicast address family as specified.

Step 20 **address-family l2vpn evpn**

Example:

```
Spine(config-router-neighbor-af)# address-family l2vpn evpn
```

Configures address family Layer 2 VPN EVPN under the BGP neighbor.

Step 21 **disable-peer-as-check**

Example:

```
Spine(config-router-neighbor-af)# disable-peer-as-check
```

Disables checking the peer AS number during route advertisement. Configure this parameter on the spine for eBGP when all leafs are using the same AS but the spines have a different AS than leafs.

Note
This command is required for eBGP. For more information on eBGP configuration, see eBGP Underlay IP Network , on page 25.

Step 22 **send-community both**

Example:

```
Spine(config-router-neighbor-af)# send-community both
```

Configures community for BGP neighbors.

Step 23 **route-map NEXT-HOP-UNCH out**

Example:

```
Spine(config-router-neighbor-af)# route-map NEXT-HOP-UNCH out
```

Applies route-map to keep the next-hop unchanged.

Examples of VXLAN BGP EVPN with RFC 5549 Underlay

Example of VXLAN BGP EVPN with RFC 5549 Underlay for Single-Site

- The following example shows how the IPv4 address is advertised over the IPv6 interface from Spine to Leaf using LLA address:

```
spine# show ip route 10.1.1.0
IP Route Table for VRF "default"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

10.1.1.0/24, ubest/mbest: 1/0
    *via fe80::1111:2222:2222:131%default, Eth1/2, [200/0], 6d09h, bgp-2, internal, tag
 2
```

- The following example shows how the IPv4 address is advertised over the IPv6 interface from Spine to Leaf using global IPv6 address:

```
spine# show ip route 10.2.2.0
IP Route Table for VRF "default"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

10.2.2.0/24, ubest/mbest: 1/0
    *via 30:3:1::1%default, Eth1/2, [200/0], 6d09h, bgp-2, external, tag 2
```

- The following example shows reachability of IPv4 VTEP over IPv6 LLA next-hop followed by an IPv4 overlay route with IPv4 VTEP as next-hop in RFC 5549 underlay:

```
vtep1# show nve peers
! shows all the IPv4 VTEPs

Interface Peer-IP                                 State LearnType Uptime   Router-Mac
```

```
--------- ------------------------------------  ----- --------- --------
-----------------
nve1      2.2.2.2                               Up    CP        1d12h    5214.bc4d.1b08

nve1      3.3.3.2                               Up    CP        1d12h    5267.54a7.1b08

nve1      4.4.4.2                               Up    CP        1d12h    52ad.917e.1b08


vtep1# show ip route 3.3.3.2
! The remote VTEP is reachable via the IPv6 LLA in underlay
IP Route Table for VRF "default"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

3.3.3.2/32, ubest/mbest: 1/0
    *via fe80::a111:2222:3333:e11%default, Eth1/1, [20/0], 1d12h, bgp-1, external, tag
 2

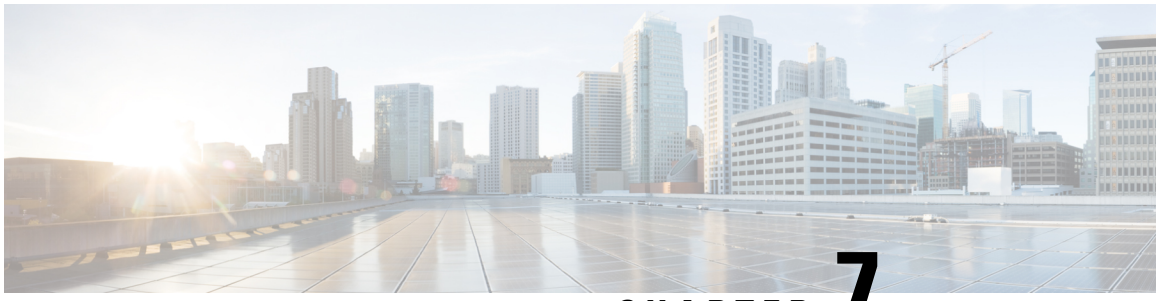
vtep1# show ip route 100.1.3.1 vrf vrF_Flow1_1
!The prefix route is reachable via the IPv4 VTEP
IP Route Table for VRF "VRF_Flow1_1"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

100.1.3.1/32, ubest/mbest: 1/0
   *via 3.3.3.2%default, [20/0], 1d12h, bgp-1, external, tag 2, segid: 501001 tunnelid:
 0x3030302 encap: VXLAN

! Following single session is used for both IPv6 underlay as well as L2VPN EVPN IPv6
session for the overlay
vtep1# show bgp sessions
! Shows the ipv6 link local sessions

Total peers 16, established peers 16
ASN 1
VRF default, local ASN 1
peers 2, established peers 2, local router-id 1.1.1.1
State: I-Idle, A-Active, O-Open, E-Established, C-Closing, S-Shutdown

Neighbor        ASN    Flaps LastUpDn|LastRead|LastWrit St Port(L/R)  Notif(S/R)
fe80::a111:2222:3333:e11%Ethernet1/1
                    2 0     1d12h   |00:00:28|00:00:31 E   179/43226      0/0
fe80::b111:2222:3333:e11%Ethernet1/2
                    2 0     1d12h   |00:00:27|00:00:21 E   179/26264      0/0
```



CHAPTER 7

EVPN Hybrid IRB Mode

• EVPN Hybrid IRB Mode, on page 195

EVPN Hybrid IRB Mode

Information About EVPN Hybrid IRB Mode

Cisco NX-OS Release 10.2(1)F introduces support for EVPN Hybrid IRB mode. This feature allows NX-OS VTEP devices operating in symmetric IRB mode to seamlessly integrate with asymmetric IRB VTEPs within the same fabric.

EVPN IRB Models

EVPN VXLAN supports Integrated Routing and Bridging (IRB) functionality which allows VTEPs in a VXLAN network to both bridge intra-subnet traffic and route inter-subnet traffic. Inter-subnet routing in an EVPN-IRB overlay network is implemented across fabric VTEPs in two ways:

- Asymmetric IRB
- Symmetric IRB

Asymmetric IRB

Asymmetric IRB uses EVPN purely as a Layer-2 VPN overlay, with inter-subnet traffic routed only at the ingress VTEP. As a result, ingress VTEP performs both routing and bridging, while the egress VTEP performs only bridging. On the ingress VTEP, packet is bridged towards the Default Gateway in the source subnet, then routed into the destination subnet local on the ingress VTEP. From that ingress routing operation, traffic is bridge via the Layer-2 VPN (VNI) tunnel. Post receiving and de-encapsulation on the egress VTEP, the packet is simply bridged to the destination end point. In essence, all packet processing associated with inter-subnet forwarding semantics is confined to the ingress VTEP. This model requires all Layer-2 VPNs to exist on all IRB VTEPs that are involved in the inter-subnet procedure for an IP VRF with consistent ARP/ND population across the fabric.

Symmetric IRB

Symmetric IRB uses EVPN as a Layer-2 and Layer-3 VPN overlay, with distributed inter-subnet traffic routed at any VTEP, ingress and egress. As a result, ingress and egress VTEP performs both routing and bridging. On the ingress VTEP, packet is bridged towards the Default Gateway in the source subnet, then routed into the destination VRF local on the ingress VTEP. From that ingress routing operation, traffic is routed via the Layer-3 VPN (VNI) tunnel. Post receiving and de-encapsulation on the egress VTEP, the packet is first routed

and then bridged to the destination end point. In essence, all packet processing associated with inter-subnet forwarding semantics is truly distributed across all VTEPs. This model allows only locally attached Layer-2 VPNs to exist on IRB VTEPs that are involved in the inter-subnet procedure for an IP VRF; the ARP/ND consumption is local to where the end point is attached.

Asymmetric and Symmetric Interop

NX-OS supports EVPN-IRB using symmetric IRB mode. While control plane and data plane is needed to enable intra subnet bridging, the procedure is identical across symmetric and asymmetric IRB modes. While the intra subnet approach is the same, the inter subnet procedure between the two IRB modes are incompatible. As a result, inter subnet routing between a symmetric IRB VTEP and a asymmetric IRB VTEP within the same fabric is not possible.

With Cisco's Hybrid IRB mode, the symmetric IRB VTEPs will support an incremental enhancement that allows to seamlessly inter-operate with VTEPs running in asymmetric IRB mode in the same fabric. NX-OS VTEPs enabled with this hybrid mode will continue to operate in the more scalable symmetric IRB mode, whenever communicating with hybrid or symmetric IRB VTEPs. In addition, the hybrid IRB will at the same time inter-operate with the asymmetric IRB VTEPs, if any exist in the same fabric.

EVPN hybrid feature is supported on the Cisco Nexus 9300-FX, FX2, FX3, GX, N9K-C9504 TOR and Modular platforms.

Beginning with Cisco NX-OS Release 10.5(2)F, VXLAN QoS is supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.

Inter-op Control Plane

Main difference between asymmetric and symmetric IRB control plane is with respect to how host MAC+IP routes (EVPN route type 2) are formatted. In asymmetric IRB, MAC+IP host routes are advertised with only layer-2 VNI encapsulation and MAC VRF route targets (RT). In symmetric IRB, MAC+IP host routes are advertised with “additional” layer-3 VNI and with “additional” IP VRF RTs to enable inter-subnet routing.

- NX-OS VTEPs provisioned in hybrid mode continue to advertise local MAC+IP routes using symmetric IRB route type 2 format with additional L3 VNI information and IP VRF RTs, such that hybrid mode NX-OS VTEPs can continue to use symmetric routing between them.
- VTEPs operating in asymmetric mode simply ignore these additional L3 VNI and IP VRF RT fields and handle these routes using asymmetric route procedure by installing layer-3 adjacencies, and host routes via these adjacencies in IP VRF. Layer-3 adjacency is a ARP/ND entry.
- NX-OS VTEPs provisioned in hybrid mode handle MAC+IP routes received from an asymmetric VTEP using asymmetric route handling. As a result, they install layer-3 adjacencies, and host routes via these adjacencies for remote hosts advertised from an asymmetric VTEP.
- Note that as a result, on an NX-OS hybrid VTEP, layer-3 adjacencies are still only installed towards hosts behind asymmetric VTEPs, and not towards hosts behind other NX-OS hybrid VTEPs.

Inter-op Provisioning Requirements

- NX-OS symmetric IRB VTEPs must be provisioned with all subnets in an IP VRF that are stretched to asymmetric VTEPs in the fabric.
- NX-OS symmetric IRB VTEPs must be provisioned with subnets in an IP VRF that are stretched to asymmetric VTEPs in “hybrid” mode using “fabric forwarding mode anycast-gateway hybrid” CLI under the subnet SVI interface.
- All symmetric IRB VTEPs must have the hybrid mode enabled when interoperating with asymmetric VTEPs in each fabric.

Inter-op Data Plane

As a result of the above requirements:

- NX-OS VTEP continues to follow symmetric routing data path with other NX-OS hybrid VTEPs in both directions. Traffic is bridged in source subnet and routed in IP VRF on ingress VTEP with L3 VNI encapsulation and then routed in IP VRF and bridged in destination subnet on the egress VTEP.
- NX-OS VTEP follows asymmetric routing data path and encapsulation towards hosts behind asymmetric VTEPs. Traffic is bridged in source subnet, routed in IP VRF with host MAC rewrite, and then bridged in destination subnet on source VTEP, while it is simply bridged in destination subnet on the egress VTEP.

Supported Features

- Hybrid mode can be enabled per L3 interfaces.
- IPv4 and IPv6 overlay end points
- Host mobility is supported with hybrid mode
- Both Ingress replication as well as multicast underlay is supported.
- Co-existance of multicast and IR underlay is supported across different VLANs
- Distributed Anycast Gateway
- vPC

Guidelines and Limitations

- Hybrid mode is not supported with DCI Border gateway.
- In Distributed Anycast Gateway mode, asymmetric IRB also needs to be provisioned with same anycast gateway MAC and IP.

Configuration Example: EVPN Hybrid IRB Mode

The following example provides the configuration of EVPN Hybrid IRB Mode:

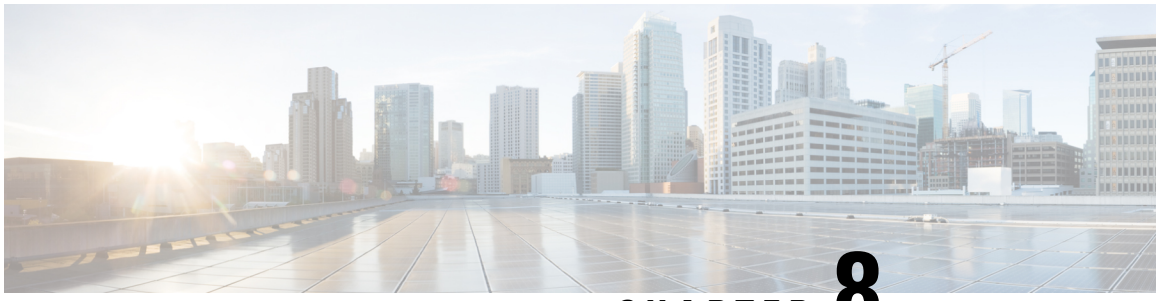
```
vlan 201
vn-segment 20001
interface vlan201
no shutdown
vrf member vrf_30001
ip address 10.1.1.1/16
fabric forwarding mode anycast-gateway hybrid
```

The following example display the VNIs and the Hybrid IRB Mode:

```
switch# show nve vni
Codes: CP - Control Plane DP - Data Plane
UC - Unconfigured SA - Suppress ARP
SU - Suppress Unknown Unicast
Xconn - Crossconnect
MS-IR - Multisite Ingress Replication
HYB - Hybrid IRB Mode
```

EVPN Hybrid IRB Mode

```
Interface VNI Multicast-group State Mode Type [BD/VRF] Flags
--------- -------- ----------------- ----- ---- ------------------ -----
nve1 5001 234.1.1.1 Up CP L2 [1001]
nve1 5002 234.1.1.1 Up CP L2 [1002]
nve1 5010 225.1.1.1 Up CP L2 [3003] HYB
nve1 6010 n/a Up CP L3 [vni_6010]
nve1 10001 n/a Up CP L3 [vni_10001]
nve1 30001 234.1.1.1 Up CP L2 [3001] HYB
nve1 30002 234.1.1.1 Up CP L2 [3002] HYB
```



CHAPTER 8

Default Gateway Coexistence of HSRP and Anycast Gateway (VXLAN EVPN)

This chapter contains these sections:

- Default Gateway Coexistence of HSRP and Anycast Gateway (VXLAN EVPN), on page 199
- Guidelines and Limitations for Migrating from Classic Ethernet / FabricPath to VXLAN, on page 200
- Configuring Classic Ethernet / FabricPath to VXLAN Migration, on page 202
- Configuring an External Port on Border Leaf for Migration, on page 203
- Configuring External IP Address for Migration, on page 204

Default Gateway Coexistence of HSRP and Anycast Gateway (VXLAN EVPN)

This feature provides coexistence between traditional Default Gateways using First Hop Gateway Protocol (HSRP being the mode supported in this release), and Distributed Anycast Gateway (DAG) for VXLAN EVPN fabrics. Instead of a disruptive cut-over or inefficient hair pinning, Default Gateways with HSRP can now be active at the same time as VXLAN EVPNs DAG, as long as the common Default Gateway MAC and IP is configured. The functionality as part of this feature provides ease for migration and coexistence between Classic Ethernet / FabricPath and VXLAN EVPN fabrics. This functionality is solely enabled on the VXLAN EVPN side, more specifically on the Border nodes neighboring the Classic Ethernet / FabricPath network. This feature allows more efficient routing and less disruptive migrations without the requirement for Software or Hardware upgrades on the Classic Ethernet / FabricPath side.

Migration can now be performed with minimal traffic impact even when both DAG is functional on VXLAN network and HSRP gateway is functional on Classic Ethernet / FabricPath network for the same VLAN after the premigration step is performed on the Classic Ethernet / FabricPath HSRP gateway. For more information, see details for premigration step in Configuring Classic Ethernet / FabricPath to VXLAN Migration, on page 202.

Coexistence of both DAG and HSRP gateway was not possible earlier for the same VLAN even after the premigration step was performed. This coexistence will enable optimal routing for the Layer 3 workloads that are migrated to VXLAN network during migration.

Layer 2 Interconnection

- Interconnecting the two networks via Layer 2 is crucial to facilitate seamless workload migration from Classic Ethernet / FabricPath to VXLAN.

- The border leaf on VXLAN network is connected via a Layer 2 interface to the Classic Ethernet / FabricPath network.
- The Layer 2 link can be a port channel trunk or a physical Ethernet trunk.
- The VXLAN border leaf switch can be a vPC or a NX-OS switch and the switch can be a TOR or an EOR. Similarly, the Classic Ethernet / FabricPath border-edge switch can be a vPC or a NX-OS switch. The switch could also host the HSRP gateway for the Classic Ethernet / FabricPath network.

For migration, you must configure the following on the VXLAN border leaf:

- The Layer 2 ports connecting the two network infrastructures must be configured as **port-type external**. These ports are referred as *external* interfaces.
- A unique Burned In Address (BIA) address for IPv4 and IPv6 must be configured on the SVI of each VXLAN border leaf during migration of the VLAN.
- If the VXLAN border leaf is in a vPC configuration, then the BIA address for the SVI must be different on both switches.

The following table provides few Layer 2 interconnection combinations:

Table 5: Layer 2 Interconnection Combinations

VXLAN Border Leaf	Classic Ethernet / FabricPath Border Edge Switch
VPC	VPC
NX-OS switch	NX-OS switch
NX-OS switch	VPC
VPC	NX-OS switch

Guidelines and Limitations for Migrating from Classic Ethernet / FabricPath to VXLAN

- Ingress PACL region must be carved and made available before configuring the migration of workloads for 9300-FX/FX2/FX3/GX/GX2B platforms deployed as VXLAN border leaf nodes.

  For example: You need to verify if the PACL region is carved before configuring the **port-type external** command on the ports connecting the VXLAN and Classic Ethernet / FabricPath networks. You can verify if the ingress PACL region is configured by using the **show hardware access-list tcam region** command. If the region is unavailable, configure the region using the **hardware access-list tcam region ing-ifacl 512** command. Ensure that you reload the switch after the PACL region is configured.
- Verify that there is no ingress PACL policy configured on the external interfaces before migration. If they are configured, you must remove them before configuring the **port-type external** command.
- vPC Fabric Peering, Egress CNTACL, VRRP, and VXLAN Flood and Learn are not supported with this migration. Also, this migration does not support moving workloads that are multicast sources or receivers.
- It is recommend that you configure only up to six external interfaces.

- For migration, ensure that you do not have the *Extended IFACL* feature configured using the **hardware access-list tcam label ing-ifacl 6** command.
- Migration of IPv4 and IPv6 applications must be performed sequentially as mentioned below:
  1. Premigration step must be performed on HSRP gateway for IPv4 gateway-IP for a particular VLAN. For more information, see details for premigration step in Configuring Classic Ethernet / FabricPath to VXLAN Migration, on page 202.
  2. The migration procedure in terms of configuring SVIs with BIA address for IPv4 must be performed on each VXLAN border leaf node connecting to the Classic Ethernet / FabricPath network.
  3. Migrate all the IPv4 hosts from Classic Ethernet / FabricPath to VXLAN side.
  4. After all the IPv4 hosts for all VLANs are migrated from Classic Ethernet / FabricPath to VXLAN, the premigration step and migration procedure has to be repeated for IPv6.



Note It is recommended that you limit the migration of concurrent host to a maximum of 1000 hosts. Start the next migration only after the previous migration of hosts is complete.

- If we have a vPC VXLAN border leaf configured, Layer 3 peer-router needs to be enabled.
- If the Suppress ARP or Suppress ND feature is enabled on the VXLAN network during Classic Ethernet / FabricPath to VXLAN migration, the host must be learned in the respective ARP or ND tables on the VXLAN border leaf. You can send a GARP/ND before moving the host to VXLAN.

  If adjacency is not learned for the host that is moved to VXLAN, then traffic from the host behind Classic Ethernet / FabricPath network to this host can fail on the Classic Ethernet / FabricPath network.

  For example:

  - When host 10.10.1.8 is being moved to VXLAN, initially, it is not learned as shown:

```
switch# sh ip arp 10.10.1.8 vrf vrf1501

IP ARP Table
Total number of entries: 0
Address         Age       MAC Address     Interface       Flags
switch#

switch(config)# sh ip route 10.10.1.8 vrf vrf1501

10.10.1.0/24, ubest/mbest: 2/0, attached
    *via 10.10.1.1, Vlan1001, [0/0], 22:55:42, direct
    *via 10.10.1.4, Vlan1001, [0/0], 22:55:42, direct
```

  - After sending GARP from host 10.10.1.8, the ARP table output of the border leaf switch is as shown:

```
switch# sh ip arp 10.10.1.8 vrf vrf1501

Flags: * - Adjacencies learnt on non-active FHRP router
       + - Adjacencies synced via CFSoE
       # - Adjacencies Throttled for Glean
       CP - Added via L2RIB, Control plane Adjacencies
       PS - Added via L2RIB, Peer Sync
       RO - Re-Originated Peer Sync Entry
       D - Static Adjacencies attached to down interface
```

```
IP ARP Table
Total number of entries: 1
Address         Age       MAC Address     Interface       Flags
10.10.1.8       00:00:04  0000.8aa9.79d3  Vlan1001

switch(config)# sh ip route 10.10.1.8 vrf vrf1501

10.10.1.8/32, ubest/mbest: 1/0, attached
    *via 10.10.1.8, Vlan1001, [190/0], 00:00:14, hmm
```

- After GARP, the host is moved to leaf in the VXLAN network as shown:

```
switch(config)# sh ip route 10.10.1.8 vrf vrf1501

10.10.1.8/32, ubest/mbest: 1/0
    *via 2.2.2.5%default, [200/0], 00:00:23, bgp-200, internal, tag 200, segid:
11501 tunnelid: 0x2020205 encap: VXLAN
```

- Beginning with Cisco NX-OS Release 10.5(2)F, the following features are supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card:
  - HSRP
  - VRRP
  - VXLAN Flood and Learn (Static Ingress Replication)
  - FabricPath to VXLAN Migration

Configuring Classic Ethernet / FabricPath to VXLAN Migration

To migrate workloads from Classic Ethernet / FabricPath to VXLAN, perform these steps:

Note Check if PACL region was carved using the **show hardware access-list tcam region** command for EX/FX/FX2 platforms. If not, ensure that PACL region is carved and made available before configuring migration of workloads.

Procedure

Step 1 Ensure that you have a Layer 2 interconnection between the VXLAN and the Classic Ethernet / FabricPath networks. As specified in Table 5: Layer 2 Interconnection Combinations, on page 200, this can be between a VXLAN border leaf (with or without vPC configuration) and the Classic Ethernet / FabricPath edge switch (with or without vPC configuration). This interface can be a physical Ethernet Layer 2 port or a Layer 2 port channel. For more information, see Configuring VXLAN BGP EVPN, on page 113.

Step 2 If there is a vPC VXLAN border leaf, ensure that **peer-gateway** and **layer3 peer-router** commands are configured.

Step 3 As part of the premigration step, configure the Anycast gateway MAC address (value present on VXLAN fabric) in HSRP for a particular VLAN on the Classic Ethernet / FabricPath network using the **mac-address** *address* {**ipv4** | **ipv6**} under HSRP.

With this premigration step configured, a GARP is triggered and it will update all hosts in that VLAN with the Anycast gateway MAC address.

Step 4 Configure a port on VXLAN border leaf as an external port using the **port-type external** for the Layer 2 port connecting the two fabrics.

Step 5 Ensure that the SVI for the VLAN that is to be migrated is configured on all the VXLAN leafs including border leaf. This step is required if there is a routed traffic for the VLAN. Ensure that you keep the SVI in the shutdown state.

Step 6 On the VXLAN border leaf, ensure that the SVI is configured with the IPv4 and/or IPv6 BIA address.

This configuration is required so that a proxy-ARP or ND request can be sent using this BIA IP address as the source-IP address and VDC-MAC as source-MAC over the external interfaces to Classic Ethernet / FabricPath network. This configuration ensures that you do not use the regular gateway-IP and the Anycast gateway MAC. This configuration will prevent collision of MACs after the premigration step.

Step 7 The IPv4 or IPv6 BIA address must be in the same subnet as the source address on the SVI of the VXLAN border leaf.

Step 8 Bring up the SVI using the **no shutdown** command under the SVI on all Leafs of VXLAN.

With this configuration, when a workload on a VLAN is moved from Classic Ethernet / FabricPath to VXLAN, it can route on the source VXLAN leaf following the VXLAN Distributed Anycast Gateway (DAG) paradigm.

Step 9 Hosts for the VLAN that continue to exist on Classic Ethernet / FabricPath side will route at the HSRP gateway. With this, both DAG and HSRP are coexisting and functional for the VLAN.

Step 10 Move all hosts from Classic Ethernet / FabricPath to VXLAN for a given VLAN.

Step 11 Ensure that all the hosts in one address family (IPv4 or IPv6) are migrated completely before migrating the other address family.

Step 12 After all the hosts for a VLAN are moved from Classic Ethernet / FabricPath to VXLAN, the HSRP gateway SVI can be removed from the Classic Ethernet / FabricPath side for the VLAN.

Step 13 After all the VLANs have been migrated from Classic Ethernet / FabricPath to VXLAN for both address families (IPv4 and IPv6), run the **no port-type external** command on the Layer 2 interfaces connecting the two fabrics. The BIA address are no longer required and can be removed from the SVI of border leafs.

The migration will now be complete.

Configuring an External Port on Border Leaf for Migration

For migrating applications or workloads from Classic Ethernet / FabricPath to VXLAN, you must configure ports on border leaf as an external port for Layer 2 interconnection.

Before you begin

For migrating hosts in a VLAN from Classic Ethernet / FabricPath to VXLAN, ensure that you complete the premigration step for the VLAN on the Classic Ethernet / FabricPath side. For this, configure an Anycast gateway MAC address in HSRP for Classic Ethernet / FabricPath network for the VLAN.

SUMMARY STEPS

1. **configure terminal**
2. **interface port-channel** *number*

3. **port-type external**

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# configure terminal`
	`switch(config)#`
		Enters configuration mode.

Step 2	**interface port-channel** *number*
	Example:
	`switch(config)# interface port-channel 40`
	`switch(config-if)#`
		Enters configuration mode and configures a port channel interface.

Step 3	**port-type external**
	Example:
	`switch(config-if)# port-type external`
	`switch(config-if)#`
		Configures the interface to be the external interface that connects to a Classic Ethernet / FabricPath network.

What to do next

As mentioned in the steps, we need to configure a BIA address for IPv4 or IPv6 on the SVI where VLAN-hosts are being moved from Classic Ethernet / FabricPath to VXLAN. For configuration this, see Configuring External IP Address for Migration, on page 204.

Configuring External IP Address for Migration

SUMMARY STEPS

1. **configure terminal**
2. **interface vlan** *vlan-id*
3. **vrf member** *vrf-name*
4. **ip address** *address netmask*
5. **ip address** *address netmask* **secondary use-bia**
6. **ipv6 address** *address netmask*
7. **ipv6 address** *address netmask* **use-bia**

DETAILED STEPS

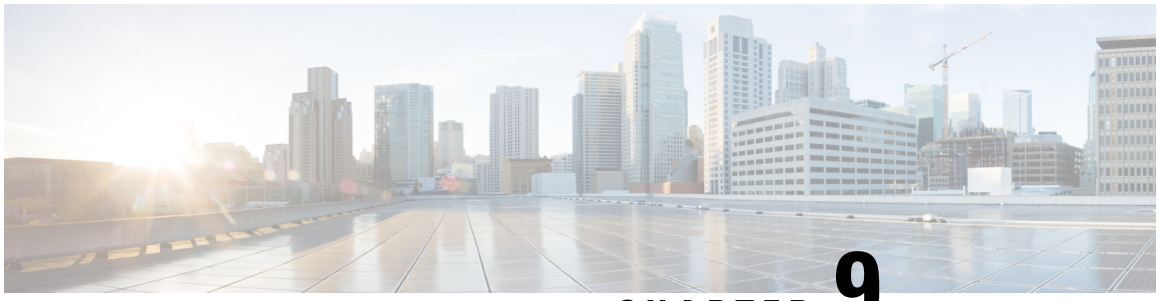
Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# configure terminal`
	`switch(config)#`
		Enters configuration mode.

Step 2	**interface vlan** *vlan-id*
	Example:
	`switch(config)# interface vlan 1100`
	`switch(config-if)#`
		Creates a VLAN interface and enters the interface configuration mode.

Step 3	**vrf member** *vrf-name* **Example:** `switch(config-if)# vrf member vrf50`	Adds this interface to VRF.
Step 4	**ip address** *address netmask* **Example:** `switch(config-if)# ip address 192.168.1.1/24`	Assigns an IPv4 address to the interface.
Step 5	**ip address** *address netmask* **secondary use-bia** **Example:** `switch(config-if)# ip address 192.168.1.10/24 secondary use-bia`	Configures external IPv4 address.
Step 6	**ipv6 address** *address netmask* **Example:** `switch(config-if)# ipv6 address 2001:DB8:1::1/64`	Assigns an IPv6 address to the interface.
Step 7	**ipv6 address** *address netmask* **use-bia** **Example:** `switch(config-if)# ipv6 address 2001:DB8:1::10/64 use-bia`	Configures external IPv6 address.



CHAPTER 9

Configuring External VRF Connectivity and Route Leaking

This chapter contains these sections:

- Configuring External VRF Connectivity, on page 207
- Configuring Route Leaking, on page 225

Configuring External VRF Connectivity

About External Layer-3 Connectivity for VXLAN BGP EVPN Fabrics

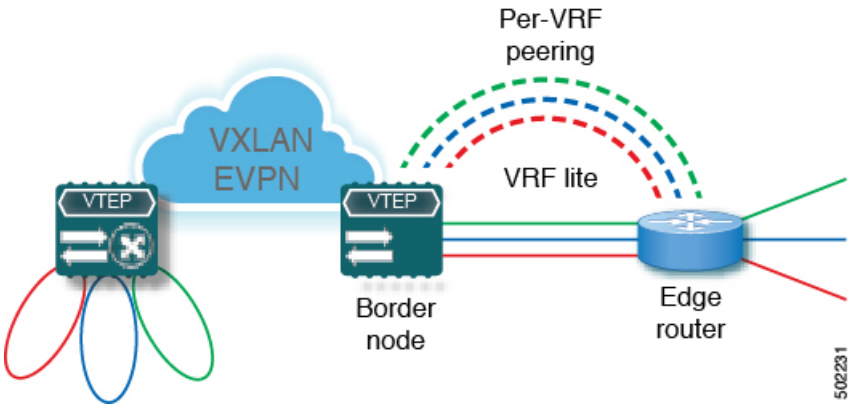
A VXLAN BGP EVPN fabric can be extended by using per-VRF IP routing to achieve external connectivity. The approach that is used for the Layer-3 extensions is commonly referred to as VRF Lite, while the functionality itself is more accurately defined as Inter-AS Option A or back-to-back VRF connectivity.

VXLAN BGP EVPN - VRF-lite brief

Some pointers are given below:

- The VXLAN BGP EVPN fabrics is depicted on the left in the following figure.
- Routes within the fabric are exchanged between all Edge-Devices (VTEPs) as well as Route-Reflectors; the control-plane used is MP-BGP with EVPN address-family.
- The Edge-Devices (VTEPs) acting as border nodes are configured to pass on prefixes to the external router (ER). This is achieved by exporting prefixes from MP-BGP EVPN to IPv4/IPv6 per-VRF peerings.
- Various routing protocols can be used for the per-VRF peering. While eBGP is the protocol of choice, IGPs like OSPF, IS-IS or EIGRP can be leveraged but require redistribution

Figure 18: External Layer-3 Connectivity - VRF-lite



Guidelines and Limitations for External VRF Connectivity and Route Leaking

The following guidelines and limitations apply to external Layer 3 connectivity for VXLAN BGP EVPN fabrics:

- Support is added for Cisco Nexus 9504 and 9508 platform switches with Cisco Nexus 96136YC-R and 9636C-RX line cards.
- A physical Layer 3 interface (parent interface) can be used for external Layer 3 connectivity (that is, VRF default).
- The parent interface to multiple subinterfaces cannot be used for external Layer 3 connectivity (that is, Ethernet1/1 for a VRF default). You can use a subinterface instead.
- Beginning with Cisco NX-OS Release 9.3(5), VTEPs support VXLAN-encapsulated traffic over parent interfaces if subinterfaces are configured.
- VTEPs do not support VXLAN-encapsulated traffic over subinterfaces, regardless of VRF participation or IEEE 802.1Q encapsulation.
- Mixing subinterfaces for VXLAN and non-VXLAN VLANs is not supported.
- The **import map** command applied under address-family ipv4 unicast does not control what gets imported into the EVPN table L3VNI counterpart.
- If TRM is configured, SVIs must not be used to interconnect to the external router.

Configuring VXLAN BGP EVPN with eBGP for VRF-lite

Configuring VRF for VXLAN Routing and External Connectivity using BGP

Configure the VRF on the border node.

SUMMARY STEPS

1. **configure terminal**
2. **vrf context** *vrf-name*
3. **vni** *number*
4. **rd** {**auto** | *rd*}

5. **address-family** {**ipv4** | **ipv6**} **unicast**
6. **route-target both** {**auto** | *rt*}
7. **route-target both** {**auto** | *rt*} **evpn**
8. Repeat Step 1 through Step 7 for every L3VNI.

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enter global configuration mode.
Step 2	**vrf context** *vrf-name*	Configure the VRF.
Step 3	**vni** *number*	Specify the VNI. The VNI associated with the VRF is often referred to as a Layer 3 VNI, L3VNI, or L3VPN. The L3VNI is configured as the common identifier across the participating VTEPs.
Step 4	**rd** {**auto**	*rd*}	Specify the VRF's route distinguisher (RD). The RD uniquely identifies a VTEP within an L3VNI. If you enter an RD, the following formats are supported: ASN2:NN, ASN4:NN, or IPV4:NN.
Step 5	**address-family** {**ipv4**	**ipv6**} **unicast**	Configure the IPv4 or IPv6 unicast address family.
Step 6	**route-target both** {**auto**	*rt*}	Configure the route target (RT) for import and export of IPv4 prefixes. The RT is used for a per-VRF prefix import/export policy. If you enter an RT, the following formats are supported: ASN2:NN, ASN4:NN, or IPV4:NN. Manually configured RTs are required to support asymmetric VNIs.
Step 7	**route-target both** {**auto**	*rt*} **evpn**	Configure the route target (RT) for import and export of IPv4 prefixes. The RT is used for a per-VRF prefix import/export policy. If you enter an RT, the following formats are supported: ASN2:NN, ASN4:NN, or IPV4:NN. Manually configured RTs are required to support asymmetric VNIs.
Step 8	Repeat Step 1 through Step 7 for every L3VNI.	

Configuring the L3VNI's Fabric Facing VLAN and SVI on the Border Node

SUMMARY STEPS

1. **configure terminal**
2. **vlan** *number*
3. **vn-segment** *number*
4. **interface** *vlan-number*

5. **mtu** *value*
6. **vrf member** *vrf-name*
7. **ip forward**
8. **no ip redirects**
9. **ipv6** *ip-address*
10. **no ipv6 redirects**
11. Repeat Step 2 through Step 10 for every L3VNI.

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enter configuration mode.
Step 2	**vlan** *number*	Specify the VLAN id that is used for the L3VNI.
Step 3	**vn-segment** *number*	Map the L3VNI to the VLAN for VXLAN EVPN routing.
Step 4	**interface** *vlan-number*	Specify the SVI (Switch Virtual Interface) for VXLAN EVPN routing.
Step 5	**mtu** *value*	Specify the MTU for the L3VNI.
Step 6	**vrf member** *vrf-name*	Map the SVI to the matching VRF context.
Step 7	**ip forward**	Enable IPv4 forwarding for the L3VNI.
Step 8	**no ip redirects**	Disable ICMP redirects
Step 9	**ipv6** *ip-address*	Enable IPv6 forwarding for the L3VNI.
Step 10	**no ipv6 redirects**	Disable ICMPv6 redirects.
Step 11	Repeat Step 2 through Step 10 for every L3VNI.	

Configuring the VTEP on the Border Node

SUMMARY STEPS

1. **configure terminal**
2. **interface nve1**
3. **member vni** *vni* **associate-vrf**
4.

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enter global configuration mode.
Step 2	**interface nve1**	Configure the NVE interface.
Step 3	**member vni** *vni* **associate-vrf**	Add Layer-3 VNIs, one per tenant VRF, to the overlay.
Step 4		Repeat Step 3 for every L3VNI.

Configuring the BGP VRF Instance on the Border Node for IPv4 per-VRF Peering

SUMMARY STEPS

1. **configure terminal**
2. **router bgp** *autonomous-system-number*
3. **vrf** *vrf-name*
4. **address-family ipv4 unicast**
5. **advertise l2vpn evpn**
6. **maximum-paths ibgp** *number*
7. **maximum-paths** *number*
8. **neighbor** *address* **remote-as** *number*
9. **update-source** *type/id*
10. **address-family ipv4 unicast**
11. Repeat Step 3 through Step 10 for every L3VNI that requires external connectivity for IPv4.

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enter global configuration mode.
Step 2	**router bgp** *autonomous-system-number*	Configure BGP. The range of the *autonomous-system-number* is from 1 to 4294967295.
Step 3	**vrf** *vrf-name*	Specify the VRF.
Step 4	**address-family ipv4 unicast**	Configure address family for IPv4.
Step 5	**advertise l2vpn evpn**	Enable the advertisement of EVPN routes within IPv4 address-family.
Step 6	**maximum-paths ibgp** *number*	Enabling equal cost multipathing (ECMP) for iBGP prefixes. The range for *number* if 1 to 64. The default is 1.

	Command or Action	Purpose
Step 7	**maximum-paths** *number*	Enabling equal cost multipathing (ECMP) for eBGP prefixes.
Step 8	**neighbor** *address* **remote-as** *number*	Define eBGP neighbor IPv4 address and remote Autonomous-System (AS) number.
Step 9	**update-source** *type/id*	Define interface for eBGP peering.
Step 10	**address-family ipv4 unicast**	Activate the IPv4 address family for IPv4 prefix exchange.
Step 11	Repeat Step 3 through Step 10 for every L3VNI that requires external connectivity for IPv4.	

Configuring the BGP VRF Instance on the Border Node for IPv6 per-VRF Peering

SUMMARY STEPS

1. **configure terminal**
2. **router bgp** *autonomous-system-number*
3. **vrf** *vrf-name*
4. **address-family ipv6 unicast**
5. **advertise l2vpn evpn**
6. **maximum-paths ibgp** *number*
7. **maximum-paths** *number*
8. **neighbor** *address* **remote-as** *number*
9. **update-source** *type/id*
10. **address-family ipv6 unicast**
11. Repeat Step 3 Through Step 10 for every L3VNI that requires external connectivity for IPv6.

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enter global configuration mode.
Step 2	**router bgp** *autonomous-system-number*	Configure BGP.
Step 3	**vrf** *vrf-name*	Specify the VRF.
Step 4	**address-family ipv6 unicast**	Configure address family for IPv4.
Step 5	**advertise l2vpn evpn**	Enable the advertisement of EVPN routes within IPv6 address-family.
Step 6	**maximum-paths ibgp** *number*	Enabling equal cost multipathing (ECMP) for iBGP prefixes.

	Command or Action	Purpose
Step 7	**maximum-paths** *number*	Enabling equal cost multipathing (ECMP) for eBGP prefixes.
Step 8	**neighbor** *address* **remote-as** *number*	Define eBGP neighbor IPv6 address and remote Autonomous-System (AS) number.
Step 9	**update-source** *type/id*	Define interface for eBGP peering.
Step 10	**address-family ipv6 unicast**	Configure address family for IPv6.
Step 11	Repeat Step 3 Through Step 10 for every L3VNI that requires external connectivity for IPv6.	

Configuring the Sub-Interface Instance on the Border Node for Per-VRF Peering - Version 1

SUMMARY STEPS

1. **configure terminal**
2. interface *type/id*
3. **no switchport**
4. **no shutdown**
5. **exit**
6. interface *type/id*
7. **encapsulation dot1q** *number*
8. **vrf member** *vrf-name*
9. **ip address** *address*
10. **no shutdown**
11. Repeat Step 5 through Step 9 for every per-VRF peering.

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	interface *type/id*	Configure parent interface.
Step 3	**no switchport**	Disable Layer-2 switching mode on interface.
Step 4	**no shutdown**	Bring up parent interface.
Step 5	**exit**	Exit interface configuration mode.
Step 6	interface *type/id*	Define the Sub-Interface instance.
Step 7	**encapsulation dot1q** *number*	Configure the VLAN ID for the sub-interface. The *number* argument can have a value from 1 to 3967.

	Command or Action	Purpose
Step 8	**vrf member** *vrf-name*	Map the Sub-Interface to the matching VRF context.
Step 9	**ip address** *address*	Configure the Sub-Interfaces IP address.
Step 10	**no shutdown**	Bring up Sub-Interface.
Step 11	Repeat Step 5 through Step 9 for every per-VRF peering.	

VXLAN BGP EVPN - Default-Route, Route Filtering on External Connectivity

About Configuring Default Routing for External Connectivity

For default-route advertisement into a VXLAN BGP EVPN fabric, we have to ensure that the default-route advertised into the fabric is at the same time not advertised outside of the fabric. For this case, it is necessary to have route filtering in place that prevents this eventuality.

Configuring the Default Route in the Border Nodes VRF

SUMMARY STEPS

1. **configure terminal**
2. **vrf context** *vrf-name*
3. **ip route 0.0.0.0/0** *next-hop*
4. **ipv6 route 0::/0** *next-hop*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**vrf context** *vrf-name*	Configure the VRF.
Step 3	**ip route 0.0.0.0/0** *next-hop*	Configure the IPv4 default-route.
Step 4	**ipv6 route 0::/0** *next-hop*	Configure the IPv6 default-route.

Configuring the BGP VRF Instance on the Border Node for IPv4/IPv6 Default-Route Advertisement

SUMMARY STEPS

1. **configure terminal**
2. **router bgp** *autonomous-system-number*
3. **vrf** *vrf-name*
4. **address-family ipv4 unicast**
5. **network 0.0.0.0/0**
6. **address-family ipv6 unicast**

7. **network 0::/0**
8. **neighbor** *address***remote-as** *number*
9. **update-source** *type/id*
10. **address-family {ipv4 | ipv6} unicast**
11. **route-map** *name* **out**
12. Repeat Step 3 through Step 11 for every L3VNI that requires external connectivity with default-route filtering.

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**router bgp** *autonomous-system-number*	Configure BGP.
Step 3	**vrf** *vrf-name*	Specify the VRF.
Step 4	**address-family ipv4 unicast**	Configure the IPv4 Unicast address-family. Required for IPv6 over VXLAN with IPv4 underlay.
Step 5	**network 0.0.0.0/0**	Creating IPv4 default-route network statement.
Step 6	**address-family ipv6 unicast**	Configure the IPv6 unicast address-family.
Step 7	**network 0::/0**	Creating IPv6 default-route network statement.
Step 8	**neighbor** *address***remote-as** *number*	Define eBGP neighbor IPv4 address and remote Autonomous-System (AS) number.
Step 9	**update-source** *type/id*	Define interface for eBGP peering
Step 10	**address-family {ipv4 | ipv6} unicast**	Activate the IPv4 or IPv6 address family for IPv4/IPv6 prefix exchange.
Step 11	**route-map** *name* **out**	Attach route-map for egress route filtering.
Step 12	Repeat Step 3 through Step 11 for every L3VNI that requires external connectivity with default-route filtering.	

Configuring Route Filtering for IPv4 Default-Route Advertisement

You can configure route filtering for IPv4 default-route advertisement.

SUMMARY STEPS

1. **configure terminal**
2. **ip prefix-list** *name* **seq 5 permit 0.0.0.0/0**
3. **route-map** *name* **deny 10**
4. **match ip address prefix-list** *name*

5. **route-map** *name* **permit 1000**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**ip prefix-list** *name* **seq 5 permit 0.0.0.0/0**	Configure IPv4 prefix-list for default-route filtering.
Step 3	**route-map** *name* **deny 10**	Create route-map with leading deny statement to prevent the default-route of being advertised via External Connectivity.
Step 4	**match ip address prefix-list** *name*	Match against the IPv4 prefix-list that contains the default-route.
Step 5	**route-map** *name* **permit 1000**	Create route-map with trailing allow statement to advertise non-matching routes via External Connectivity.

Configuring Route Filtering for IPv6 Default-Route Advertisement

SUMMARY STEPS

1. **configure terminal**
2. **ipv6 prefix-list** *name* **seq 5 permit 0::/0**
3. **route-map** *name* **deny 10**
4. **match ipv6 address prefix-list** *name*
5. **route-map** *name* **permit 1000**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**ipv6 prefix-list** *name* **seq 5 permit 0::/0**	Configure IPv6 prefix-list for default-route filtering.
Step 3	**route-map** *name* **deny 10**	Create route-map with leading deny statement to prevent the default-route of being advertised via External Connectivity.
Step 4	**match ipv6 address prefix-list** *name*	Match against the IPv6 prefix-list that contains the default-route.
Step 5	**route-map** *name* **permit 1000**	Create route-map with trailing allow statement to advertise non-matching routes via External Connectivity.

About Configuring Default-Route Distribution and Host-Route Filter

Per-default, a VXLAN BGP EVPN fabric always advertises all known routes via the External Connectivity. As not in all circumstances it is beneficial to advertise IPv4 /32 or IPv6 /128 Host-Routes, a respective route filtering approach can become necessary.

Configuring the BGP VRF Instance on the Border Node for IPv4/IPv6 Host-Route Filtering

SUMMARY STEPS

1. **configure terminal**
2. **router bgp** *autonomous-system-number*
3. **vrf** *vrf-name*
4. **neighbor** *address* **remote-as** *number*
5. **update-source** *type/id*
6. **address-family {ipv4 | ipv6} unicast**
7. **route-map** *name* **out**
8. Repeat Step 3 through Step 7 for every L3VNI that requires external connectivity with host-route filtering.

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**router bgp** *autonomous-system-number*	Configure BGP.
Step 3	**vrf** *vrf-name*	Specify the VRF.
Step 4	**neighbor** *address* **remote-as** *number*	Define eBGP neighbor IPv4/IPv6 address and remote Autonomous-System (AS) number.
Step 5	**update-source** *type/id*	Define interface for eBGP peering.
Step 6	**address-family {ipv4 | ipv6} unicast**	Activate the IPv4 or IPv6 address family for IPv4/IPv6 prefix exchange.
Step 7	**route-map** *name* **out**	Attach route-map for egress route filtering.
Step 8	Repeat Step 3 through Step 7 for every L3VNI that requires external connectivity with host-route filtering.	

Configuring Route Filtering for IPv4 Host-Route Advertisement

SUMMARY STEPS

1. **configure terminal**
2. **ip prefix-list** *name* **seq 5 permit 0.0.0.0/0 eq 32**
3. **route-map** *name* **deny 10**
4. **match ip address prefix-list** *name*

5. **route-map** *name* **permit 1000**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**ip prefix-list** *name* **seq 5 permit 0.0.0.0/0 eq 32**	Configure IPv4 prefix-list for host-route filtering.
Step 3	**route-map** *name* **deny 10**	Create route-map with leading deny statement to prevent the default-route of being advertised via External Connectivity.
Step 4	**match ip address prefix-list** *name*	Match against the IPv4 prefix-list that contains the host-route.
Step 5	**route-map** *name* **permit 1000**	Create route-map with trailing allow statement to advertise non-matching routes via external connectivity.

Configuring Route Filtering for IPv6 Host-Route Advertisement

SUMMARY STEPS

1. **configure terminal**
2. **ipv6 prefix-list** *name* **seq 5 permit 0::/0 eq 128**
3. **route-map** *name* **deny 10**
4. **match ipv6 address prefix-list** *name*
5. **route-map** *name* **permit 1000**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**ipv6 prefix-list** *name* **seq 5 permit 0::/0 eq 128**	Configure IPv4 prefix-list for host-route filtering.
Step 3	**route-map** *name* **deny 10**	Create route-map with leading deny statement to prevent the default-route of being advertised via External Connectivity.
Step 4	**match ipv6 address prefix-list** *name*	Match against the IPv4 prefix-list that contains the host-route.
Step 5	**route-map** *name* **permit 1000**	Create route-map with trailing allow statement to advertise non-matching routes via External Connectivity.

Example - Configuring VXLAN BGP EVPN with eBGP for VRF-lite

An example of external connectivity from VXLAN BGP EVPN to an external router using VRF-lite.

Configuring VXLAN BGP EVPN Border Node

The VXLAN BGP EVPN Border Node acts as neighbor device to the External Router. The VRF Name is purely localized and can be different to the VRF Name on the External Router, only significance is the L3VNI must be consistent across the VXLAN BGP EVPN fabric. For the ease of reading, the VRF and interface enumeration will be consistently used.

The configuration examples represents a IPv4 and IPv6 dual-stack approach; IPv4 or IPv6 can be substituted of each other.

```
vrf context myvrf_50001
  vni 50001
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
  address-family ipv6 unicast
    route-target both auto
    route-target both auto evpn
!
vlan 2000
  vn-segment 50001
!
interface Vlan2000
  no shutdown
  mtu 9216
  vrf member myvrf_50001
  no ip redirects
  ip forward
  ipv6 address use-link-local-only
  no ipv6 redirects
!
interface nve1
  no shutdown
  host-reachability protocol bgp
  source-interface loopback1
  member vni 50001 associate-vrf
!
router bgp 65002
  vrf myvrf_50001
    router-id 10.2.0.6
    address-family ipv4 unicast
      advertise l2vpn evpn
      maximum-paths ibgp 2
      maximum-paths 2
    address-family ipv6 unicast
      advertise l2vpn evpn
      maximum-paths ibgp 2
      maximum-paths 2
    neighbor 10.31.95.95
      remote-as 65099
      address-family ipv4 unicast
    neighbor 2001::95/64
      remote-as 65099
      address-family ipv4 unicast
!
interface Ethernet1/3
  no switchport
  no shutdown
```

```
interface Ethernet1/3.2
  encapsulation dot1q 2
  vrf member myvrf_50001
  ip address 10.31.95.31/24
  ipv6 address 2001::31/64
  no shutdown
```

Configuring Default-Route, Route Filtering on External Connectivity

The VXLAN BGP EVPN Border Node has the ability to advertise IPv4 and IPv6 default-route within the fabric. In cases where it is not beneficial to advertise the Host Routes from the VXLAN BGP EVPN fabric to the External Router, these IPv4 /32 and IPv6 /128 can be filtered at the External Connectivity peering configuration.

```
ip prefix-list default-route seq 5 permit 0.0.0.0/0 le 1
ipv6 prefix-list default-route-v6 seq 5 permit 0::/0
!
ip prefix-list host-route seq 5 permit 0.0.0.0/0 eq 32
ipv6 prefix-list host-route-v6 seq 5 permit 0::/0 eq 128
!
route-map extcon-rmap-filter deny 10
  match ip address prefix-list default-route
route-map extcon-rmap-filter deny 20
  match ip address prefix-list host-route
route-map extcon-rmap-filter permit 1000
!
route-map extcon-rmap-filter-v6 deny 10
  match ipv6 address prefix-list default-route-v6
route-map extcon-rmap-filter-v6 deny 20
  match ip address prefix-list host-route-v6
route-map extcon-rmap-filter-v6 permit 1000
!
vrf context myvrf_50001
  ip route 0.0.0.0/0 10.31.95.95
  ipv6 route 0::/0 2001::95/64
!
router bgp 65002
  vrf myvrf_50001
    address-family ipv4 unicast
      network 0.0.0.0/0
    address-family ipv6 unicast
      network 0::/0


    neighbor 10.31.95.95
      remote-as 65099
      address-family ipv4 unicast
        route-map extcon-rmap-filter out
    neighbor 2001::95/64
      remote-as 65099
      address-family ipv4 unicast
        route-map extcon-rmap-filter-v6 out
```

Configuring External Router

The External Router performs as a neighbor device to the VXLAN BGP EVPN border node. The VRF Name is purely localized and can be different to the VRF Name on the VXLAN BGP EVPN Fabric. For the ease of reading, the VRF and interface enumeration will be consistently used.

The configuration examples represents a IPv4 and IPv6 dual-stack approach; IPv4 or IPv6 can be substituted of each other.

```
vrf context myvrf_50001
!
router bgp 65099
  vrf myvrf_50001
    address-family ipv4 unicast
      maximum-paths 2
    address-family ipv6 unicast
      maximum-paths 2
    neighbor 10.31.95.31
      remote-as 65002
      address-family ipv4 unicast
    neighbor 2001::31/64
      remote-as 65002
      address-family ipv4 unicast
!
interface Ethernet1/3
  no switchport
  no shutdown
interface Ethernet1/3.2
  encapsulation dot1q 2
  vrf member myvrf_50001
  ip address 10.31.95.95/24
  Ipv6 address 2001::95/64
  no shutdown
```

Configuring VXLAN BGP EVPN with OSPF for VRF-lite

Configuring VRF for VXLAN Routing and External Connectivity using OSPF

Configure the BGP VRF instance on the border node for OSPF per-VRF peering.

SUMMARY STEPS

1. **configure terminal**
2. **router bgp** *autonomous-system-number*
3. **vrf** *vrf-name*
4. **address-family ipv4 unicast**
5. **advertise l2vpn evpn**
6. **maximum-paths ibgp** *number*
7. **redistribute ospf** *name* **route-map** *name*
8. Repeat Step 3 through Step 7 for every per-VRF peering.

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enter global configuration mode.
Step 2	**router bgp** *autonomous-system-number*	Configure BGP.
Step 3	**vrf** *vrf-name*	Specify the VRF.
Step 4	**address-family ipv4 unicast**	Configure the IPv4 address family.

	Command or Action	Purpose
Step 5	**advertise l2vpn evpn**	Enable the advertisement of EVPN routes within the address family.
Step 6	**maximum-paths ibgp** *number*	Enabling equal-cost multipathing (ECMP) for iBGP prefixes.
Step 7	**redistribute ospf** *name* **route-map** *name*	Define redistribution from OSPF into BGP.
Step 8	Repeat Step 3 through Step 7 for every per-VRF peering.	

Configuring the Route-Map for BGP to OSPF Redistribution

SUMMARY STEPS

1. **configure terminal**
2. **route-map** *name* **permit 10**
3. **match route-type internal**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enter global configuration mode.
Step 2	**route-map** *name* **permit 10**	Create route-map for BGP to OSPF redistribution
Step 3	**match route-type internal**	Redistribution route-map must allow the matching of BGP internal route-types if iBGP is used in the VXLAN BGP EVPN fabric.

Configuring the OSPF on the Border Node for Per-VRF Peering

SUMMARY STEPS

1. **configure terminal**
2. **router ospf** *instance*
3. **vrf** *vrf-name*
4. **redistribute bgp** *autonomous-system-number* **route-map** *name*
5. Repeat Step 3 through Step 4 for every per-VRF peering.

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enter global configuration mode.
Step 2	**router ospf** *instance*	Configure OSPF.
Step 3	**vrf** *vrf-name*	Specify the VRF.
Step 4	**redistribute bgp** *autonomous-system-number* **route-map** *name*	Define redistribution from BGP to OSPF.
Step 5	Repeat Step 3 through Step 4 for every per-VRF peering.	

Configuring the Sub-Interface Instance on the Border Node for Per-VRF Peering - Version 2

SUMMARY STEPS

1. **configure terminal**
2. **interface** *type/id*
3. **no switchport**
4. **no shutdown**
5. **exit**
6. **interface** *type/id*
7. **encapsulation dot1q** *number*
8. **vrf member** *vrf-name*
9. **ip address** *address*
10. **ip ospf network point-to-point**
11. **ip router ospf** *name* **area** *area-id*
12. **no shutdown**
13. Repeat Step 5 through Step 12 for every per-VRF peering.

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**interface** *type/id*	Configure parent interface.
Step 3	**no switchport**	Disable Layer-2 switching mode on interface.
Step 4	**no shutdown**	Bring up parent interface.
Step 5	**exit**	Exit interface configuration mode.

	Command or Action	Purpose
Step 6	**interface** *type/id*	Define the Sub-Interface instance.
Step 7	**encapsulation dot1q** *number*	Configure the VLAN ID for the sub-interface. The range is from 2 to 4093.
Step 8	**vrf member** *vrf-name*	Map the Sub-Interface to the matching VRF context.
Step 9	**ip address** *address*	Configure the Sub-Interfaces IP address.
Step 10	**ip ospf network point-to-point**	Define OSPF network-type for sub-interface.
Step 11	**ip router ospf** *name* **area** *area-id*	Configure the OSPF instance.
Step 12	**no shutdown**	Bring up Sub-Interface.
Step 13	Repeat Step 5 through Step 12 for every per-VRF peering.	

Example - Configuration VXLAN BGP EVPN with OSPF for VRF-lite

An example of external connectivity from VXLAN BGP EVPN to an External Router using VRF-lite.

Configuring VXLAN BGP EVPN Border Node with OSPF

The VXLAN BGP EVPN Border Node acts as neighbor device to the External Router. The VRF Name is purely localized and can be different to the VRF Name on the External Router, only significance is the L3VNI must be consistent across the VXLAN BGP EVPN fabric. For the ease of reading, the VRF and interface enumeration will be consistently used.

The configuration examples represents a IPv4 approach with OSPFv2.

```
route-map extcon-rmap-BGP-to-OSPF permit 10
  match route-type internal
route-map extcon-rmap-OSPF-to-BGP permit 10
!
vrf context myvrf_50001
  vni 50001
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
!
vlan 2000
  vn-segment 50001
!
interface Vlan2000
  no shutdown
  mtu 9216
  vrf member myvrf_50001
  no ip redirects
  ip forward
!
interface nve1
  no shutdown
  host-reachability protocol bgp
  source-interface loopback1
  member vni 50001 associate-vrf
!
router bgp 65002
```

```
  vrf myvrf_50001
    router-id 10.2.0.6
    address-family ipv4 unicast
      advertise l2vpn evpn
      maximum-paths ibgp 2
      maximum-paths 2
      redistribute ospf EXT route-map extcon-rmap-OSPF-to-BGP
!
router ospf EXT
  vrf myvrf_50001
    redistribute bgp 65002 route-map extcon-rmap-BGP-to-OSPF
!
interface Ethernet1/3
  no switchport
  no shutdown
interface Ethernet1/3.2
  encapsulation dot1q 2
  vrf member myvrf_50001
  ip address 10.31.95.31/24
  ip ospf network point-to-point
  ip router ospf EXT area 0.0.0.0
  no shutdown
```

Configuring Route Leaking

About Centralized VRF Route-Leaking for VXLAN BGP EVPN Fabrics

VXLAN BGP EVPN uses MP-BGP and its route-policy concept to import and export prefixes. The ability of this very extensive route-policy model allows to leak routes from one VRF to another VRF and vice-versa; any combination of custom VRF or VRF default can be used. VRF route-leaking is a switch-local function at specific to a location in the network, the location where the cross-VRF route-target import/export configuration takes place (leaking point). The forwarding between the different VRFs follows the control-plane, the location of where the configuration for the route-leaking is performed - hence Centralized VRF route-leaking. With the addition of VXLAN BGP EVPN, the leaking point requires to advertise the cross-VRF imported/exported route and advertise them towards the remote VTEPs or External Routers.

The advantage of Centralized VRF route-leaking is that only the VTEP acting as leaking point requires the special capabilities needed, while all other VTEPs in the network are neutral to this function.

Guidelines and Limitations for Centralized VRF Route-Leaking

The following are the guidelines and limitations for Centralized VRF Route-Leaking:

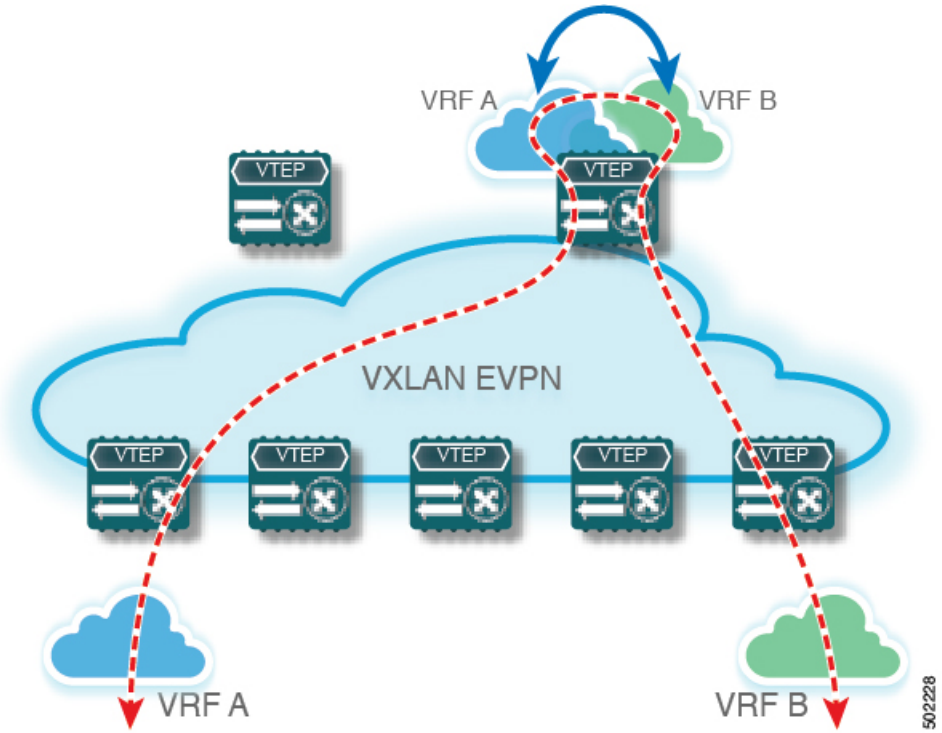
- Each prefix must be imported into each VRF for full cross-VRF reachability.
- The **feature bgp** command is required for the **export vrf default** command.
- If a VTEP has a less specific local prefix in its VRF, the VTEP might not be able to reach a more specific prefix in a different VRF.
- VXLAN routing in hardware and packet reencapsulation at VTEP is required for Centralized VRF Route-Leaking with BGP EVPN.
- Beginning with Cisco NX-OS Release 9.3(5), asymmetric VNIs are used to support Centralized VRF Route-Leaking. For more information, see About VXLAN EVPN with Downstream VNI, on page 121.

Centralized VRF Route-Leaking Brief - Specific Prefixes Between Custom VRF

Some pointers are given below:

- The Centralized VRF route-leaking for VXLAN BGP EVPN fabrics is depicted within Figure 2.
- BGP EVPN prefixes are cross-VRF leaked by exporting them from VRF Blue with an import into VRF Red and vice-versa. The Centralized VRF route-leaking is performed on the centralized Routing-Block (RBL) and could be any or multiple VTEPs.
- Configured less specific prefixes (aggregates) are advertised from the Routing-Block to the remaining VTEPs in the respective destination VRF.
- BGP EVPN does not export prefixes that were previously imported to prevent the occurrence of routing loops.

Figure 19: Centralized VRF Route-Leaking - Specific Prefixes with Custom VRF



Configuring Centralized VRF Route-Leaking - Specific Prefixes between Custom VRF

Configuring VRF Context on the Routing-Block VTEP

This procedure applies equally to IPv6.

SUMMARY STEPS

1. **configure terminal**
2. **vrf context** *vrf-name*
3. **vni** *number*
4. **rd auto**
5. **address-family ipv4 unicast**
6. **route-target both** {**auto** | *rt*}
7. **route-target both** {**auto** | *rt*} **evpn**
8. **route-target import** *rt-from-different-vrf*
9. **route-target import** *rt-from-different-vrf* **evpn**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enter global configuration mode.
Step 2	**vrf context** *vrf-name*	Configure the VRF.
Step 3	**vni** *number*	Specify the VNI. The VNI associated with the VRF is often referred to as Layer 3 VNI, L3VNI, or L3VPN. The L3VNI is configured as a common identifier across the participating VTEPs.
Step 4	**rd auto**	Specify the VRF's route distinguisher (RD). The RD uniquely identifies a VTEP within an L3VNI.
Step 5	**address-family ipv4 unicast**	Configure the IPv4 unicast address family.
Step 6	**route-target both** {**auto** | *rt*}	Configure the route target (RT) for import and export of IPv4 prefixes. The RT is used for a per-VRF prefix import/export policy. If you enter an RT, the following formats are supported: ASN2:NN, ASN4:NN, or IPV4:NN. Manually configured RTs are required to support asymmetric VNIs.
Step 7	**route-target both** {**auto** | *rt*} **evpn**	Configure the route target (RT) for import and export of IPv4 prefixes. The RT is used for a per-VRF prefix import/export policy. If you enter an RT, the following formats are supported: ASN2:NN, ASN4:NN, or IPV4:NN. Manually configured RTs are required to support asymmetric VNIs.
Step 8	**route-target import** *rt-from-different-vrf*	Configure the RT for importing IPv4 prefixes from the leaked-from VRF. The following formats are supported: ASN2:NN, ASN4:NN, or IPV4:NN.

	Command or Action	Purpose
Step 9	**route-target import** *rt-from-different-vrf* **evpn**	Configure the RT for importing IPv4 prefixes from the leaked-from VRF. The following formats are supported: ASN2:NN, ASN4:NN, or IPV4:NN.

Configuring the BGP VRF instance on the Routing-Block

This procedure applies equally to IPv6.

SUMMARY STEPS

1. **configure terminal**
2. **router bgp** *autonomous-system number*
3. **vrf** *vrf-name*
4. **address-family ipv4 unicast**
5. **advertise l2vpn evpn**
6. **aggregate-address** *prefix/mask*
7. **maximum-paths ibgp** *number*
8. **maximum-paths** *number*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**router bgp** *autonomous-system number*	Configure BGP.
Step 3	**vrf** *vrf-name*	Specify the VRF.
Step 4	**address-family ipv4 unicast**	Configure address family for IPv4
Step 5	**advertise l2vpn evpn**	Enable the advertisement of EVPN routes within IPv4 address-family.
Step 6	**aggregate-address** *prefix/mask*	Create less specific prefix aggregate into the destination VRF.
Step 7	**maximum-paths ibgp** *number*	Enabling equal cost multipathing (ECMP) for iBGP prefixes.
Step 8	**maximum-paths** *number*	Enabling equal cost multipathing (ECMP) for eBGP prefixes

Example - Configuration Centralized VRF Route-Leaking - Specific Prefixes Between Custom VRF

Configuring VXLAN BGP EVPN Routing-Block

The VXLAN BGP EVPN Routing-Block acts as centralized route-leaking point. The leaking configuration is localized such that control-plane leaking and data-path forwarding follow the same path. Most significantly is the VRF configuration of the Routing-Block and the advertisement of the less specific prefixes (aggregates) into the respective destination VRFs.

```
vrf context Blue
  vni 51010
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
    route-target import 65002:51020
    route-target import 65002:51020 evpn
!
vlan 2110
  vn-segment 51010
!
interface Vlan2110
  no shutdown
  mtu 9216
  vrf member Blue
  no ip redirects
  ip forward
!
vrf context Red
  vni 51020
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
    route-target import 65002:51010
    route-target import 65002:51010 evpn
!
vlan 2120
  vn-segment 51020
!
interface Vlan2120
  no shutdown
  mtu 9216
  vrf member Blue
  no ip redirects
  ip forward
!
interface nve1
  no shutdown
  host-reachability protocol bgp
  source-interface loopback1
  member vni 51010 associate-vrf
  member vni 51020 associate-vrf
!
router bgp 65002
  vrf Blue
    address-family ipv4 unicast
      advertise l2vpn evpn
      aggregate-address 10.20.0.0/16
      maximum-paths ibgp 2
      Maximum-paths 2
  vrf Red
    address-family ipv4 unicast
```

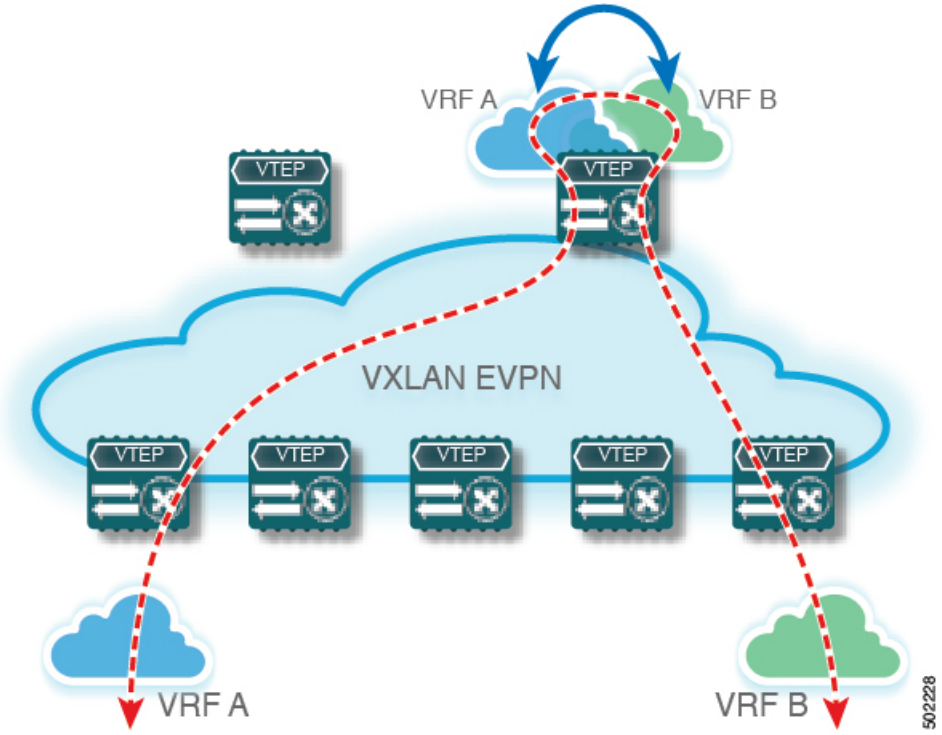
```
advertise l2vpn evpn
aggregate-address 10.10.0.0/16
maximum-paths ibgp 2
Maximum-paths 2
```

Centralized VRF Route-Leaking Brief - Shared Internet with Custom VRF

Some pointers follow:

- The Shared Internet with VRF route-leaking for VXLAN BGP EVPN fabrics is depicted in the following figure.
- The default-route is made exported from the Shared Internet VRF and re-advertisement within VRF Blue and VRF Red on the Border Node.
- Ensure the default-route in VRF Blue and VRF Red is not leaked to the Shared Internet VRF.
- The less specific prefixes for VRF Blue and VRF Red are exported for the Shared Internet VRF and re-advertised as necessary.
- Configured less specific prefixes (aggregates) that are advertised from the Border Node to the remaining VTEPs to the destination VRF (Blue or Red).
- BGP EVPN does not export prefixes that were previously imported to prevent the occurrence of routing loops.

Figure 20: Centralized VRF Route-Leaking - Shared Internet with Custom VRF



Configuring Centralized VRF Route-Leaking - Shared Internet with Custom VRF

Configuring Internet VRF on Border Node

This procedure applies equally to IPv6.

SUMMARY STEPS

1. **configure terminal**
2. **vrf context** *vrf-name*
3. **vni** *number*
4. **ip route 0.0.0.0/0** *next-hop*
5. **rd auto**
6. **address-family ipv4 unicast**
7. **route-target both** {**auto** | *rt*}
8. **route-target both** *shared-vrf-rt* **evpn**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enter global configuration mode.
Step 2	**vrf context** *vrf-name*	Configure the VRF.
Step 3	**vni** *number*	Specify the VNI. The VNI associated with the VRF is often referred to as Layer 3 VNI, L3VNI, or L3VPN. The L3VNI is configured as a common identifier across the participating VTEPs.
Step 4	**ip route 0.0.0.0/0** *next-hop*	Configure the default route in the shared internet VRF to the external router.
Step 5	**rd auto**	Specify the VRF's route distinguisher (RD). The RD uniquely identifies a VTEP within an L3VNI.
Step 6	**address-family ipv4 unicast**	Configure the IPv4 unicast address family. This configuration is required for IPv4 over VXLAN with IPv4 underlay.
Step 7	**route-target both** {**auto**	*rt*}	Configure the route target (RT) for the import and export of EVPN and IPv4 prefixes. If you enter an RT, the following formats are supported: ASN2:NN, ASN4:NN, or IPV4:NN. Manually configured RTs are required to support asymmetric VNIs.

	Command or Action	Purpose
Step 8	**route-target both** *shared-vrf-rt* **evpn**	Configure a special route target (RT) for the import and export of the shared IPv4 prefixes. An additional import/export map for further qualification is supported.

Configuring Shared Internet BGP Instance on the Border Node

This procedure applies equally to IPv6.

SUMMARY STEPS

1. **configure terminal**
2. **router bgp** *autonomous-system number*
3. **vrf** *vrf-name*
4. **address-family ipv4 unicast**
5. **advertise l2vpn evpn**
6. **aggregate-address** *prefix/mask*
7. **maximum-paths ibgp** *number*
8. **maximum-paths** *number*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**router bgp** *autonomous-system number*	Configure BGP.
Step 3	**vrf** *vrf-name*	Specify the VRF.
Step 4	**address-family ipv4 unicast**	Configure address family for IPv4
Step 5	**advertise l2vpn evpn**	Enable the advertisement of EVPN routes within IPv4 address-family.
Step 6	**aggregate-address** *prefix/mask*	Create less specific prefix aggregate into the destination VRF.
Step 7	**maximum-paths ibgp** *number*	Enabling equal cost multipathing (ECMP) for iBGP prefixes.
Step 8	**maximum-paths** *number*	Enabling equal cost multipathing (ECMP) for eBGP prefixes.

Configuring Custom VRF on Border Node

This procedure applies equally to IPv6

SUMMARY STEPS

1. **configure terminal**
2. **ip prefix-list** *name* **seq 5 permit 0.0.0.0/0**
3. **route-map** *name* **deny 10**
4. **match ip address prefix-list** *name*
5. **route-map** *name* **permit 20**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**ip prefix-list** *name* **seq 5 permit 0.0.0.0/0**	Configure IPv4 prefix-list for default-route filtering.
Step 3	**route-map** *name* **deny 10**	Create route-map with leading deny statement to prevent the default-route of being leaked.
Step 4	**match ip address prefix-list** *name*	Match against the IPv4 prefix-list that contains the default-route.
Step 5	**route-map** *name* **permit 20**	Create route-map with trailing allow statement to advertise non-matching routes via route-leaking.

Configuring Custom VRF Context on the Border Node - 1

This procedure applies equally to IPv6.

SUMMARY STEPS

1. **configure terminal**
2. **vrf context** *vrf-name*
3. **vni** *number*
4. **rd auto**
5. **ip route 0.0.0.0/0 Null0**
6. **address-family ipv4 unicast**
7. **route-target both** {**auto** | *rt*}
8. **route-target both** {**auto** | *rt*} **evpn**
9. **import map** *name*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.

	Command or Action	Purpose
Step 2	**vrf context** *vrf-name*	Configure the VRF.
Step 3	**vni** *number*	Specify the VNI. The VNI associated with the VRF is often referred to as Layer 3 VNI, L3VNI, or L3VPN. The L3VNI is configured as the common identifier across the participating VTEPs.
Step 4	**rd auto**	Specify the VRF's route distinguisher (RD). The RD uniquely identifies a VTEP within an L3VNI.
Step 5	**ip route 0.0.0.0/0 Null0**	Configure default-route in common VRF to attract traffic towards Border Node with Shared Internet VRF.
Step 6	**address-family ipv4 unicast**	Configure the IPv4 address family. This configuration is required for IPv4 over VXLAN with IPv4 underlay.
Step 7	**route-target both** {**auto** | *rt*}	Configure the route target (RT) for the import and export of IPv4 prefixes within the IPv4 address family The RT is used for a per-VRF prefix import/export policy. If you enter an RT, the following formats are supported: ASN2:NN, ASN4:NN, or IPV4:NN. Manually configured RTs are required to support asymmetric VNIs.
Step 8	**route-target both** {**auto** | *rt*} **evpn**	Configure the route target (RT) for the import and export of IPv4 prefixes within the IPv4 address family The RT is used for a per-VRF prefix import/export policy. If you enter an RT, the following formats are supported: ASN2:NN, ASN4:NN, or IPV4:NN. Manually configured RTs are required to support asymmetric VNIs.
Step 9	**import map** *name*	Apply a route-map on routes being imported into this routing table.

Configuring Custom VRF Instance in BGP on the Border Node

This procedure applies equally to IPv6.

SUMMARY STEPS

1. **configure terminal**
2. **router bgp** *autonomous-system-number*
3. **vrf** *vrf-name*
4. **address-family ipv4 unicast**
5. **advertise l2vpn evpn**
6. **network 0.0.0.0/0**
7. **maximum-paths ibgp** *number*
8. **maximum-paths** *number*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**router bgp** *autonomous-system-number*	Configure BGP.
Step 3	**vrf** *vrf-name*	Specify the VRF.
Step 4	**address-family ipv4 unicast**	Configure address family for IPv4.
Step 5	**advertise l2vpn evpn**	Enable the advertisement of EVPN routes within IPv4 address-family.
Step 6	**network 0.0.0.0/0**	Creating IPv4 default-route network statement.
Step 7	**maximum-paths ibgp** *number*	Enabling equal cost multipathing (ECMP) for iBGP prefixes.
Step 8	**maximum-paths** *number*	Enabling equal cost multipathing (ECMP) for eBGP prefixes.

Example - Configuration Centralized VRF Route-Leaking - Shared Internet with Custom VRF

An example of Centralized VRF route-leaking with Shared Internet VRF

Configuring VXLAN BGP EVPN Border Node for Shared Internet VRF

The VXLAN BGP EVPN Border Node provides a centralized Shared Internet VRF. The leaking configuration is localized such that control-plane leaking and data-path forwarding following the same path. Most significantly is the VRF configuration of the Border Node and the advertisement of the default-route and less specific prefixes (aggregates) into the respective destination VRFs.

```
vrf context Shared
  vni 51099
  ip route 0.0.0.0/0 10.9.9.1
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
    route-target both 99:99
    route-target both 99:99 evpn
!
vlan 2199
  vn-segment 51099
!
interface Vlan2199
  no shutdown
  mtu 9216
  vrf member Shared
  no ip redirects
  ip forward
!
ip prefix-list PL_DENY_EXPORT seq 5 permit 0.0.0.0/0
```

```
!
route-map RM_DENY_IMPORT deny 10
 match ip address prefix-list PL_DENY_EXPORT
route-map RM_DENY_IMPORT permit 20
!
vrf context Blue
  vni 51010
  ip route 0.0.0.0/0 Null0
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
    route-target both 99:99
    route-target both 99:99 evpn
    import map RM_DENY_IMPORT
!
vlan 2110
  vn-segment 51010
!
interface Vlan2110
  no shutdown
  mtu 9216
  vrf member Blue
  no ip redirects
  ip forward
!
vrf context Red
  vni 51020
  ip route 0.0.0.0/0 Null0
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
    route-target both 99:99
    route-target both 99:99 evpn
    import map RM_DENY_IMPORT
!
vlan 2120
  vn-segment 51020
!
interface Vlan2120
  no shutdown
  mtu 9216
  vrf member Blue
  no ip redirects
  ip forward
!
interface nve1
  no shutdown
  host-reachability protocol bgp
  source-interface loopback1
  member vni 51099 associate-vrf
  member vni 51010 associate-vrf
  member vni 51020 associate-vrf
!
router bgp 65002
  vrf Shared
    address-family ipv4 unicast
      advertise l2vpn evpn
      aggregate-address 10.10.0.0/16
      aggregate-address 10.20.0.0/16
      maximum-paths ibgp 2
      maximum-paths 2
  vrf Blue
```

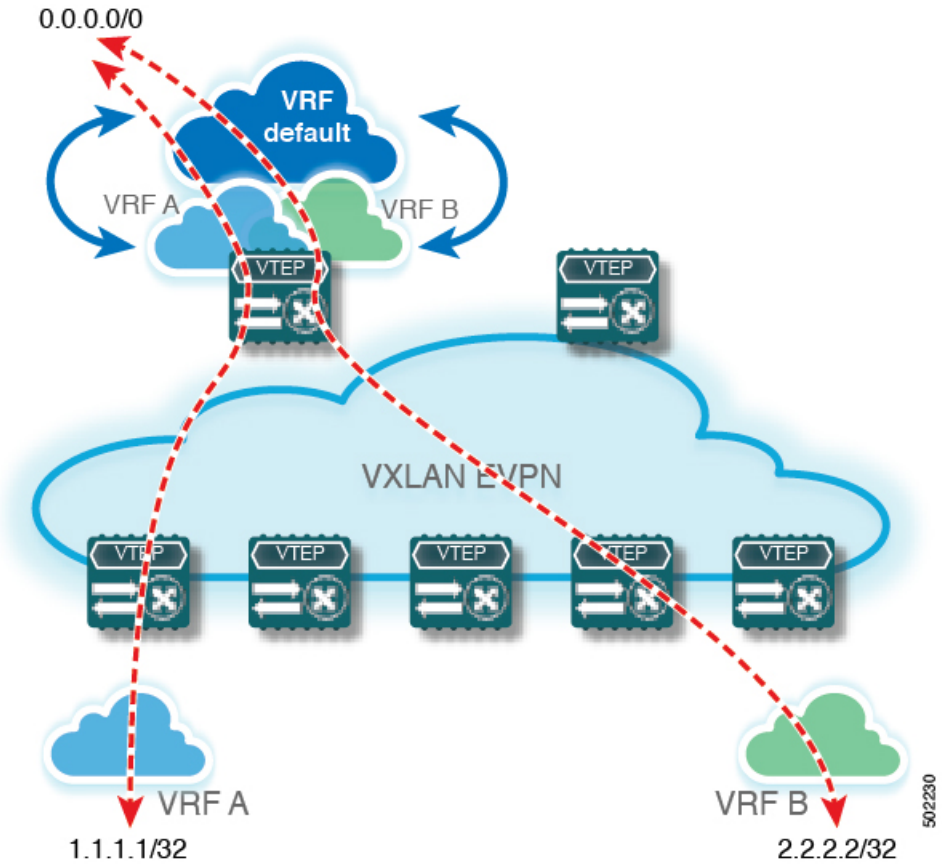
```
  address-family ipv4 unicast
    advertise l2vpn evpn
    network 0.0.0.0/0
    maximum-paths ibgp 2
    maximum-paths 2
vrf Red
  address-family ipv4 unicast
    advertise l2vpn evpn
    network 0.0.0.0/0
    maximum-paths ibgp 2
    maximum-paths 2
```

Centralized VRF Route-Leaking Brief - Shared Internet with VRF Default

Some pointers are given below:

- The Shared Internet with VRF route-leaking for VXLAN BGP EVPN fabrics is depicted within Figure 4.
- The default-route is made exported from VRF default and re-advertisement within VRF Blue and VRF Red on the Border Node.
- Ensure the default-route in VRF Blue and VRF Red is not leaked to the Shared Internet VRF
- The less specific prefixes for VRF Blue and VRF Red are exported to VRF default and re-advertised as necessary.
- Configured less specific prefixes (aggregates) that are advertised from the Border Node to the remaining VTEPs to the destination VRF (Blue or Red).
- BGP EVPN does not export prefixes that were previously imported to prevent the occurrence of routing loops.

Figure 21: Centralized VRF Route-Leaking - Shared Internet with VRF Default



Configuring Centralized VRF Route-Leaking - Shared Internet with VRF Default

Configuring VRF Default on Border Node

This procedure applies equally to IPv6.

SUMMARY STEPS

1. **configure terminal**
2. **ip route 0.0.0.0/0** *next-hop*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.

	Command or Action	Purpose
Step 2	**ip route 0.0.0.0/0** *next-hop*	Configure default-route in VRF default to external router (example)

Configuring BGP Instance for VRF Default on the Border Node

This procedure applies equally to IPv6.

SUMMARY STEPS

1. **configure terminal**
2. **router bgp** *autonomous-system number*
3. **address-family ipv4 unicast**
4. **aggregate-address** *prefix/mask*
5. **maximum-paths** *number*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**router bgp** *autonomous-system number*	Configure BGP.
Step 3	**address-family ipv4 unicast**	Configure address family for IPv4.
Step 4	**aggregate-address** *prefix/mask*	Create less specific prefix aggregate in VRF default.
Step 5	**maximum-paths** *number*	Enabling equal cost multipathing (ECMP) for eBGP prefixes.

Configuring Custom VRF on Border Node

This procedure applies equally to IPv6

SUMMARY STEPS

1. **configure terminal**
2. **ip prefix-list** *name* **seq 5 permit 0.0.0.0/0**
3. **route-map** *name* **deny 10**
4. **match ip address prefix-list** *name*
5. **route-map** *name* **permit 20**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**ip prefix-list** *name* **seq 5 permit 0.0.0.0/0**	Configure IPv4 prefix-list for default-route filtering.
Step 3	**route-map** *name* **deny 10**	Create route-map with leading deny statement to prevent the default-route of being leaked.
Step 4	**match ip address prefix-list** *name*	Match against the IPv4 prefix-list that contains the default-route.
Step 5	**route-map** *name* **permit 20**	Create route-map with trailing allow statement to advertise non-matching routes via route-leaking.

Configuring Filter for Permitted Prefixes from VRF Default on the Border Node

This procedure applies equally to IPv6.

SUMMARY STEPS

1. **configure terminal**
2. **route-map** *name* **permit 10**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**route-map** *name* **permit 10**	Create route-map with allow statement to advertise routes via route-leaking to the customer VRF and subsequently remote VTEPs.

Configuring Custom VRF Context on the Border Node - 2

This procedure applies equally to IPv6.

SUMMARY STEPS

1. **configure terminal**
2. **vrf context** *vrf-name*
3. **vni** *number*
4. **rd auto**
5. **ip route 0.0.0.0/0 Null0**

6. **address-family ipv4 unicast**
7. **route-target both** {**auto** | *rt*}
8. **route-target both** {**auto** | *rt*} **evpn**
9. **route-target both** *shared-vrf-rt*
10. **route-target both** *shared-vrf-rt* **evpn**
11. **import vrf default map** *name*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enter global configuration mode.
Step 2	**vrf context** *vrf-name*	Configure the VRF.
Step 3	**vni** *number*	Specify the VNI. The VNI associated with the VRF is often referred to as Layer 3 VNI, L3VNI, or L3VPN. The L3VNI is configured as the common identifier across the participating VTEPs.
Step 4	**rd auto**	Specify the VRF's route distinguisher (RD). The RD uniquely identifies a VTEP within an L3VNI.
Step 5	**ip route 0.0.0.0/0 Null0**	Configure default-route in common VRF to attract traffic towards Border Node with Shared Internet VRF.
Step 6	**address-family ipv4 unicast**	Configure the IPv4 address family. This configuration is required for IPv4 over VXLAN with IPv4 underlay.
Step 7	**route-target both** {**auto** | *rt*}	Configure the route target (RT) for the import and export of EVPN and IPv4 prefixes within the IPv4 address family. If you enter an RT, the following formats are supported: ASN2:NN, ASN4:NN, or IPV4:NN. Manually configured RTs are required to support asymmetric VNIs.
Step 8	**route-target both** {**auto** | *rt*} **evpn**	Configure the route target (RT) for the import and export of EVPN and IPv4 prefixes within the IPv4 address family. If you enter an RT, the following formats are supported: ASN2:NN, ASN4:NN, or IPV4:NN. Manually configured RTs are required to support asymmetric VNIs.
Step 9	**route-target both** *shared-vrf-rt*	Configure a special route target (RT) for the import/export of the shared IPv4 prefixes. An additional import/export map for further qualification is supported.
Step 10	**route-target both** *shared-vrf-rt* **evpn**	Configure a special route target (RT) for the import/export of the shared IPv4 prefixes. An additional import/export map for further qualification is supported.

	Command or Action	Purpose
Step 11	**import vrf default map** *name*	Permits all routes, from VRF default, from being imported into the custom VRF according to the specific route-map.

Configuring Custom VRF Instance in BGP on the Border Node

This procedure applies equally to IPv6.

SUMMARY STEPS

1. **configure terminal**
2. **router bgp** *autonomous-system-number*
3. **vrf** *vrf-name*
4. **address-family ipv4 unicast**
5. **advertise l2vpn evpn**
6. **network 0.0.0.0/0**
7. **maximum-paths ibgp** *number*
8. **maximum-paths** *number*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**router bgp** *autonomous-system-number*	Configure BGP.
Step 3	**vrf** *vrf-name*	Specify the VRF.
Step 4	**address-family ipv4 unicast**	Configure address family for IPv4.
Step 5	**advertise l2vpn evpn**	Enable the advertisement of EVPN routes within IPv4 address-family.
Step 6	**network 0.0.0.0/0**	Creating IPv4 default-route network statement.
Step 7	**maximum-paths ibgp** *number*	Enabling equal cost multipathing (ECMP) for iBGP prefixes.
Step 8	**maximum-paths** *number*	Enabling equal cost multipathing (ECMP) for eBGP prefixes.

Example - Configuration Centralized VRF Route-Leaking - VRF Default with Custom VRF

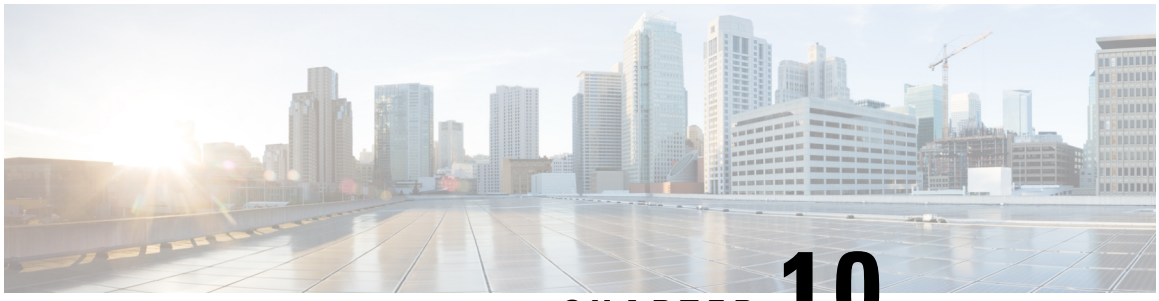
An example of Centralized VRF route-leaking with VRF default

Configuring VXLAN BGP EVPN Border Node for VRF Default

The VXLAN BGP EVPN Border Node provides centralized access to VRF default. The leaking configuration is localized such that control-plane leaking and data-path forwarding following the same path. Most significantly is the VRF configuration of the Border Node and the advertisement of the default-route and less specific prefixes (aggregates) into the respective destination VRFs.

```
ip route 0.0.0.0/0 10.9.9.1
!
ip prefix-list PL_DENY_EXPORT seq 5 permit 0.0.0.0/0
!
route-map permit 10
match ip address prefix-list PL_DENY_EXPORT
route-map RM_DENY_EXPORT permit 20
route-map RM_PERMIT_IMPORT permit 10
!
vrf context Blue
  vni 51010
  ip route 0.0.0.0/0 Null0
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
    import vrf default map RM_PERMIT_IMPORT
    export vrf default 100 map RM_DENY_EXPORT allow-vpn
!
vlan 2110
  vn-segment 51010
!
interface Vlan2110
  no shutdown
  mtu 9216
  vrf member Blue
  no ip redirects
  ip forward
!
vrf context Red
  vni 51020
  ip route 0.0.0.0/0 Null0
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
    import vrf default map RM_PERMIT_IMPORT
    export vrf default 100 map RM_DENY_EXPORT allow-vpn
!
vlan 2120
  vn-segment 51020
!
interface Vlan2120
  no shutdown
  mtu 9216
  vrf member Blue
  no ip redirects
  ip forward
!
interface nve1
  no shutdown
  host-reachability protocol bgp
  source-interface loopback1
  member vni 51010 associate-vrf
  member vni 51020 associate-vrf
!
router bgp 65002
```

```
address-family ipv4 unicast
    aggregate-address 10.10.0.0/16
    aggregate-address 10.20.0.0/16
    maximum-paths 2
    maximum-paths ibgp 2
vrf Blue
  address-family ipv4 unicast
    advertise l2vpn evpn
    network 0.0.0.0/0
    maximum-paths ibgp 2
    maximum-paths 2
vrf Red
  address-family ipv4 unicast
    advertise l2vpn evpn
    network 0.0.0.0/0
    maximum-paths ibgp 2
    maximum-paths 2
```



CHAPTER 10

Configuring Seamless Integration of EVPN with L3VPN (MPLS LDP)

This chapter contains these sections:

- Information About Configuring Seamless Integration of EVPN with L3VPN (MPLS LDP), on page 245
- Guidelines and Limitations for Configuring Seamless Integration of EVPN with L3VPN (MPLS LDP), on page 245
- Configuring Seamless Integration of EVPN with L3VPN (MPLS LDP), on page 246

Information About Configuring Seamless Integration of EVPN with L3VPN (MPLS LDP)

Data center deployments have adopted VXLAN EVPN for its benefits like EVPN control-plane learning, multitenancy, seamless mobility, redundancy, and easier POD additions. Similarly, the Core is either an LDP-based MPLS L3VPN network or transitioning from traditional an MPLS L3VPN LDP-based underlay to a more sophisticated solution like segment routing (SR). Segment routing is adopted for its benefits like unified IGP and MPLS control planes, simpler traffic engineering methods, easier configuration, and SDN adoption.

With two different technologies, a Border Leaf or a Shared PE router acting as the DCI Nodes within the data centers, it is natural to handoff from VXLAN to an MPLS-based core at the Border Leaf. These nodes which sit on the edge of the DC domain, interfacing with the Core edge router.

Guidelines and Limitations for Configuring Seamless Integration of EVPN with L3VPN (MPLS LDP)

The following are the guidelines and limitations for Configuring Seamless Integration of EVPN with L3VPN (MPLS LDP):

The following features are supported:

- Cisco Nexus 9504 and 9508 switches with -R and -RX line cards.
- Layer 3 orphans

- 256 peers/nodes within a VXLAN DC domain
- 24,000 ECMP routes is supported on -RX line cards.



Note If you enter the **no hardware profile mpls extended-ecmp** command, the mode is switched to 4 K ECMP routes. This is applicable only when the line card is -RX and the ECMP group has exactly 2 paths.

- The Egress RACL (e-RACL) TCAM and MPLS Extended ECMP features are mutually exclusive. To enable MPLS Extended ECMP (**hardware profile mpls extended-ecmp)** on the Cisco Nexus N9K-X9636C-RX line card, set the e-RACL TCAM carving to 0.
- Beginning with Cisco NX-OS Release 10.3(3)F, Type-6 encryption for MPLS LDP user password is supported on Cisco NX-OS switches.
- When configuring MPLS LDP, it is strongly recommended to use the **router-id force** command. Specify the router ID with a loopback interface to ensure consistent LDP operation and avoid issues with the order of interface initialization.

The following features are not supported:

- Subnet stretches across the DC domain
- vPC
- SVI/Subinterfaces

Configuring Seamless Integration of EVPN with L3VPN (MPLS LDP)

These configuration steps are required on a Border Leaf switch to import and re-originate the routes from a VXLAN domain to an MPLS domain and back to a VXLAN domain.

SUMMARY STEPS

1. **configure terminal**
2. **[no] install feature-set mpls**
3. **[no] feature-set mpls**
4. **feature mpls l3vpn**
5. **feature mpls ldp**
6. **mpls ip**
7. **nv overlay evpn**
8. **router bgp** *number*
9. **address-family ipv4 unicast**
10. **redistribute direct route-map** *route-map-name*
11. **exit**
12. **address-family l2vpn evpn**

13. **exit**
14. **neighbor** *address* **remote-as** *number*
15. **update-source** *type/id*
16. **ebgp-multihop** *ttl-value*
17. **address-family ipv4 unicast**
18. **send-community extended**
19. **exit**
20. **address-family ipv4 labeled-unicast**
21. **send-community extended**
22. **address-family vpnv4 unicast**
23. **send-community extended**
24. **import l2vpn evpn reoriginate**
25. **neighbor** *address* **remote-as** *number*
26. **address-family ipv4 unicast**
27. **send-community extended**
28. **address-family ipv6 unicast**
29. **send-community extended**
30. **address-family l2vpn evpn**
31. **send-community extended**
32. **import vpn unicast reoriginate**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**[no] install feature-set mpls** **Example:** `switch# install feature-set mpls`	Installs the MPLS feature set. The no form of this command uninstalls the MPLS feature set.
Step 3	**[no] feature-set mpls** **Example:** `switch# feature-set mpls`	Installs the MPLS feature set. The no form of this command uninstalls the MPLS feature set.
Step 4	**feature mpls l3vpn** **Example:** `switch# feature mpls l3vpn`	Enables the MPLS Layer 3 VPN feature.
Step 5	**feature mpls ldp** **Example:**	Enables the MPLS Label Distribution Protocol (LDP).

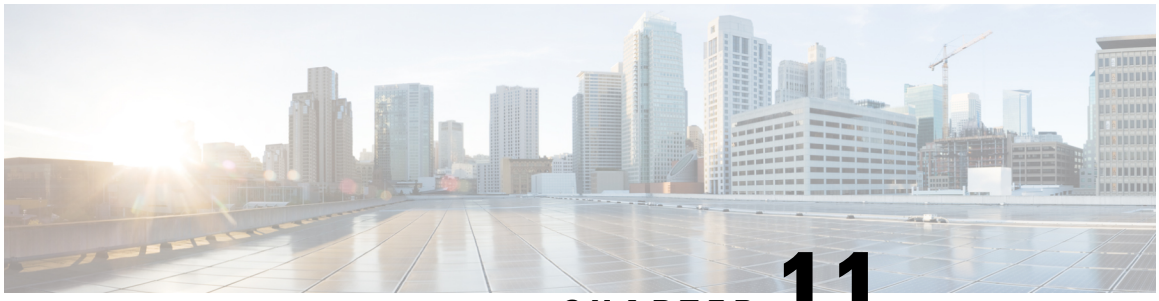
	Command or Action	Purpose
	`switch# feature mpls ldp`	

Step 6	**mpls ip**
	Example:
	`switch# interface Ethernet1/1`
	`switch(config-if)# mpls ip`
		Enables MPLS on the specified interfaces that are MPLS links.

Step 7	**nv overlay evpn** **Example:** `switch(config)# nv overlay evpn`	Enables the EVPN control plane for VXLAN.
Step 8	**router bgp** *number* **Example:** `switch(config)# router bgp 100`	Configures BGP. The value of the *number* argument is from 1 to 4294967295.
Step 9	**address-family ipv4 unicast** **Example:** `switch(config-router)# address-family ipv4 unicast`	Configures the address family for IPv4.
Step 10	**redistribute direct route-map** *route-map-name* **Example:** `switch(config-router-af)# redistribute direct route-map passall`	Configures the directly connected route map.
Step 11	**exit** **Example:** `switch(config-router-af)# exit`	Exits command mode.
Step 12	**address-family l2vpn evpn** **Example:** `switch(config-router)# address-family l2vpn evpn`	Configures the L2VPN address family.
Step 13	**exit** **Example:** `switch(config-router-af)# exit`	Exits command mode.
Step 14	**neighbor** *address* **remote-as** *number* **Example:** `switch(config-router)# neighbor 108.108.108.108 remote-as 22`	Configures a BGP neighbor. The range of the *number* argument is from 1 to 65535.
Step 15	**update-source** *type/id* **Example:** `switch(config-router-neighbor)# update-source loopback100`	Specifies the source of the BGP session and updates.

	Command or Action	Purpose
Step 16	**ebgp-multihop** *ttl-value* **Example:** `switch(config-router-neighbor)# ebgp-multihop 10`	Specifies the multihop TTL for the remote peer. The range of *ttl-value* is from 2 to 255.
Step 17	**address-family ipv4 unicast** **Example:** `switch(config-router-neighbor)# address-family ipv4 unicast`	Configures the unicast sub-address family.
Step 18	**send-community extended** **Example:** `switch(config-router-neighbor-af)# send-community extended`	Configures the community attribute for this neighbor.
Step 19	**exit** **Example:** `switch(config-router-neighbor-af)# exit`	Exits command mode.
Step 20	**address-family ipv4 labeled-unicast** **Example:** `switch(config-router-neighbor)# address-family ipv4 labeled-unicast`	Advertises the labeled IPv4 unicast routes as specified in RFC 3107.
Step 21	**send-community extended** **Example:** `switch(config-router-neighbor-af)# send-community extended`	Sends the extended community attribute.
Step 22	**address-family vpnv4 unicast** **Example:** `switch(config-router-neighbor)# address-family vpnv4 unicast`	Configures the address family for IPv4.
Step 23	**send-community extended** **Example:** `switch(config-router)# send-community extended`	Sends the extended community attribute.
Step 24	**import l2vpn evpn reoriginate** **Example:** `switch(config-router)# import l2vpn evpn reoriginate`	Reoriginates the route with a new RT.
Step 25	**neighbor** *address* **remote-as** *number* **Example:** `switch(config-router)# neighbor 175.175.175.2 remote-as 1`	Defines the neighbor.

	Command or Action	Purpose
Step 26	**address-family ipv4 unicast** **Example:** `switch(config-router)# address-family ipv4 unicast`	Configures the address family for IPv4.
Step 27	**send-community extended** **Example:** `switch(config-router)# send-community extended`	Configures the community for BGP neighbors.
Step 28	**address-family ipv6 unicast** **Example:** `switch(config-router)# address-family ipv6 unicast`	Configures the IPv6 unicast address family. This is required for IPv6 over VXLAN with an IPv4 underlay.
Step 29	**send-community extended** **Example:** `switch(config-router)# send-community extended`	Configures the community for BGP neighbors.
Step 30	**address-family l2vpn evpn** **Example:** `switch(config-router)# address-family l2vpn evpn`	Configures the L2VPN address family.
Step 31	**send-community extended** **Example:** `switch(config-router)# send-community extended`	Configures the community for BGP neighbors.
Step 32	**import vpn unicast reoriginate** **Example:** `switch(config-router)# import vpn unicast reoriginate`	Reoriginates the route with a new RT.



CHAPTER 11

Configuring Seamless Integration of EVPN with L3VPN (MPLS SR)

This chapter contains these sections:

- Information About Configuring Seamless Integration of EVPN with L3VPN (MPLS SR), on page 251
- Guidelines and Limitations for Configuring Seamless Integration of EVPN with L3VPN (MPLS SR) , on page 254
- Configuring Seamless Integration of EVPN with L3VPN (MPLS SR), on page 257
- Example Configuration for Configuring Seamless Integration of EVPN with L3VPN (MPLS SR), on page 261
- Configuring DSCP Based SR-TE Flow Steering, on page 271

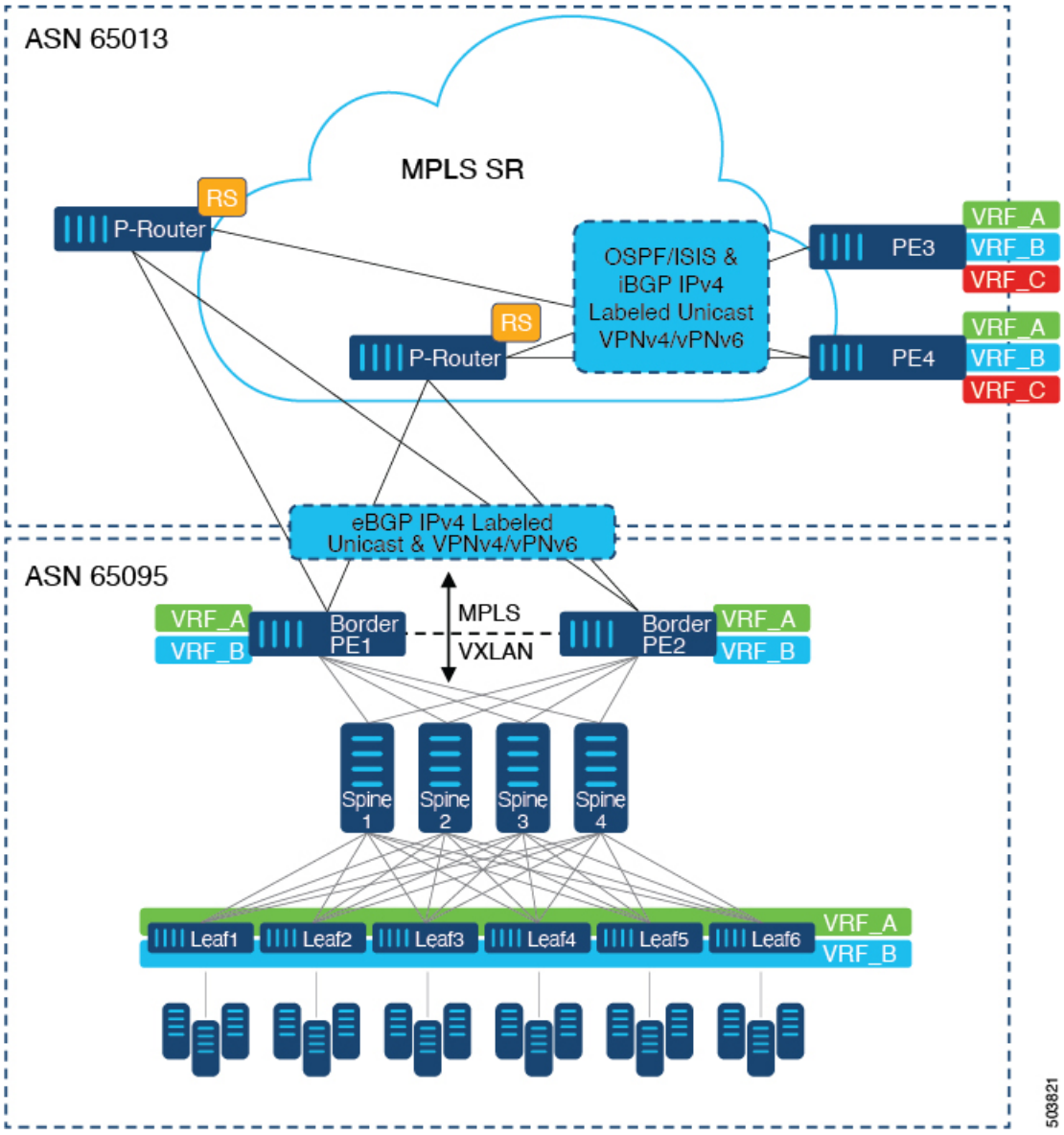
Information About Configuring Seamless Integration of EVPN with L3VPN (MPLS SR)

Data Center (DC) deployments have adopted VXLAN EVPN for its benefits such as EVPN control-plane learning, multi tenancy, seamless mobility, redundancy, and easier horizontal scaling. Similarly, the Core network transitions to different technologies with their respective capabilities. MPLS with Label Distribution Protocol (LDP) and Layer-3 VPN (L3VPN) is present in many Core networks interconnecting Data Centers. With the technology evolution, a transformation from the traditional MPLS L3VPN with LDP-based underlay to MPLS-based Segment Routing (SR) with L3VPN, became available. Segment Routing is adopted for its benefits such as:

- Unified IGP and MPLS control planes
- Simpler traffic engineering methods

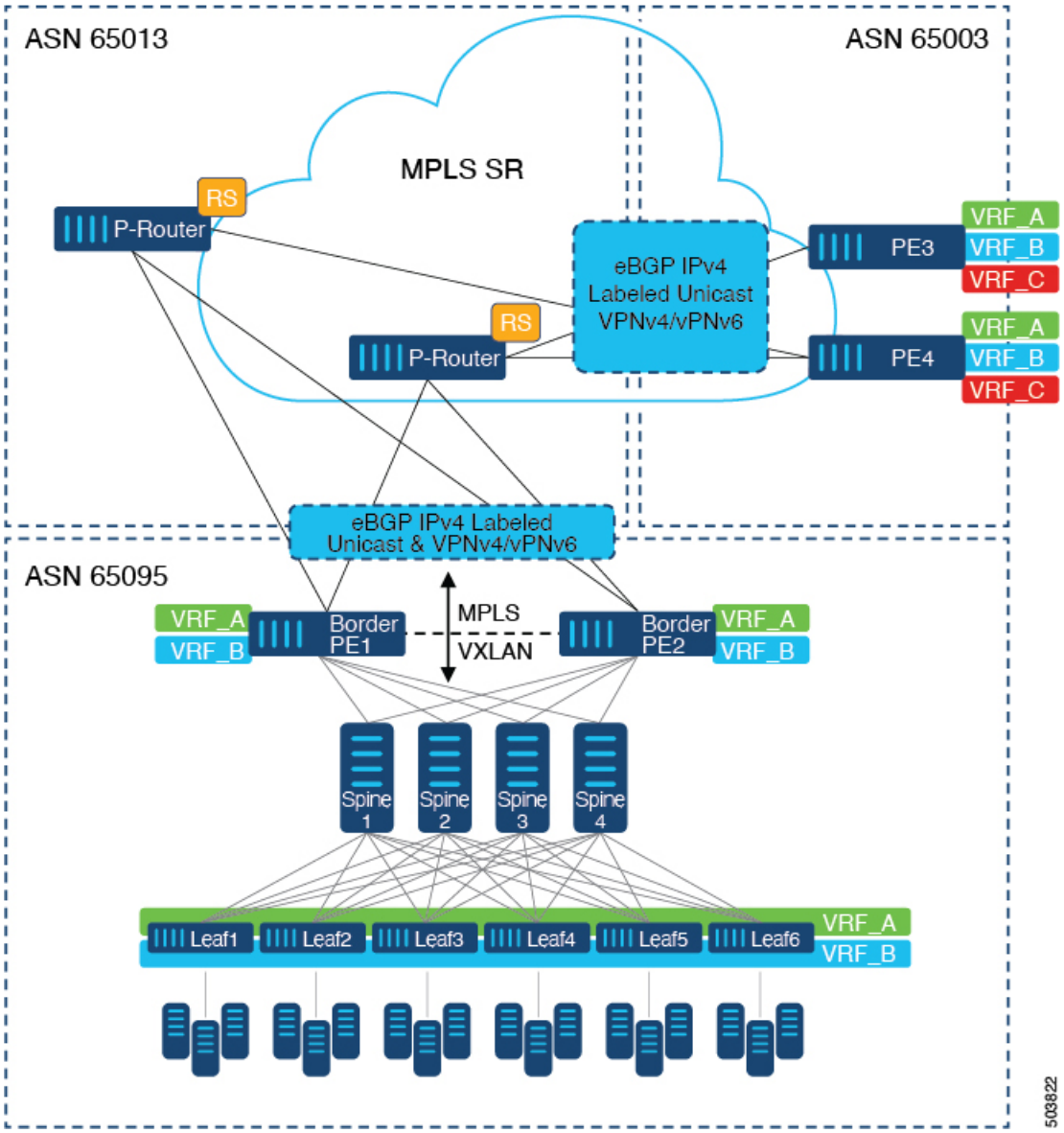
With the Data center (DC) established on VXLAN EVPN and the Core network requiring multi-tenant capable transport, there is a natural necessity to seamless integration. To provide this seamless integration between different control-plane protocols and encapsulations, in this case here from VXLAN to an MPLS-based Core network, the Cisco Nexus 9000 Series Switch provides the Border Provider Edge (Border PE) capability by interfacing the Data Center and the Core routers (Provider Routers or Provider Edge-Routers).

Figure 22: Topology with DC to Core Network Domain Separation



In the above figure, a single Data Center Fabric running VXLAN EVPN is depicted. The VRFs (VRF_A, VRF_B) present in the Data Center require to be extended over a WAN/Core running MPLS-based Segment Routing (MPLS-SR). The Data Center Fabrics Border switches acts as Border Provider Edge (Border PE1, Border PE2) interconnecting VXLAN BGP EVPN with MPLS-SR with L3VPN (VPNv4/VPNv6). The BPEs are interconnected with the Provider Router (P-Router) via eBGP using the IPv4 Labeled-Unicast as well as the VPNv4/VPNv6 Address-Family (AF). The P-Router act as BGP Route-Reflector for the mentioned AF and relays the necessary routes to the MPLS-SR Provider Edge (PE3, PE4) via iBGP. Beyond the usage of BGP as the control-plane, between the MPLS-SR nodes within the same Autonomous System (AS) uses a IGP (OSPF or ISIS) for label distribution. From the PEs shown in the above figure (PE3, PE4), Inter-AS Option A can be used to extend the Data Center or Core network VRFs to another external network. Even as this diagram shows only one Data Center, the MPLS-SR network can be used to interconnect multiple Data Center Fabrics.

Figure 23: Multiple Administrative Domains within the Core network



An alternative deployment scenario is when the Core network is separate into multiple Administrative Domains or Autonomous Systems (AS). In the above figure, a single Data Center Fabric running VXLAN EVPN is depicted. The VRFs (VRF_A, VRF_B) present in the Data Center requires to be extended over a WAN/Core running MPLS-based Segment Routing (MPLS-SR). The Data Center Fabrics Border switches acts as Border Provider Edge (Border PE1, Border PE2) interconnecting VXLAN BGP EVPN with MPLS-SR with L3VPN (VPNv4/VPNv6). The BPEs are interconnected with the Provider Router (P-Router) via eBGP using the IPv4 Labeled-Unicast as well as the VPNv4/VPNv6 Address-Family (AF). The P-Router act as BGP Route Server for the mentioned AF and relays the necessary routes to the MPLS-SR Provider Edge (PE3, PE4) via eBGP; no other control-plane protocol is used between the MPLS-SR nodes. Similar as in the previous scenario, the PEs (PE3, PE4) can operate with Inter-AS Option A to extend the Data Center or Core network VRFs to external network. Even as this diagram shows only one Data Center, the MPLS-SR network can be used to interconnect multiple Data Center Fabrics.

Beginning with Cisco NX-OS Release 10.3(1)F, DSCP Based SRTE Traffic Steering is supported on the border PE. For more information, see Configuring DSCP Based SRTE Traffic Steering. This scenario is supported only with L3VPN (MPLS SR). In the above diagram, which represents the border PE (border leaf) scenario, note the following:

1. The incoming VXLAN traffic is terminated and then sent into L3VPN (MPLS SR) so that it follows the standard routing best-path to PE3 or PE4.
2. Incoming VXLAN traffic entering PE1 is terminated, and the SRTE traffic steering policy applied on L3 VNI overrides the standard routing best-path and steer to choose an alternate path to PE3 or PE4 based on the SRTE flow steering policy.

For additional information on MPLS SR, see the *Cisco Nexus 9000 Series NX-OS Label Switching Configuration Guide*.

Guidelines and Limitations for Configuring Seamless Integration of EVPN with L3VPN (MPLS SR)

Feature	Cisco Nexus 9300-FX2, FX3, GX, GX2, H2R, H1 Platform Switches	Cisco Nexus 9504 and 9508 switches with -R Line Cards	Comments
VXLAN EVPN to SR-L3VPN	Yes	Yes	Extend Layer 3 connectivity between different DC pods Underlay IGP/BGP with SR extensions.
VXLAN EVPN to SR-L3VPN	Yes	Yes	Extend Layer 3 connectivity between DC POD running VXLAN and any domain (DC or CORE) running SR.
VXLAN EVPN to MPLS L3VPN (LDP)	No	Yes	Underlay is LDP.

The following Cisco Nexus platform switches support seamless integration of EVPN with L3VPN (MPLS SR):

- 9336C-FX2 switches
- 93240YC-FX2 switches
- 9300-FX3 platform switches
- 9300-GX platform switches
- 9504 and 9508 platform switches with 96136YC-R and 9636C-RX line cards (The 9636C-R and 9636Q-R line cards are not supported.)

Beginning with Cisco NX-OS Release 10.2(3)F, the seamless integration of EVPN with L3VPN (MPLS SR) is supported on Cisco Nexus 9300-GX2 platform switches.

Beginning with Cisco NX-OS Release 10.4(1)F, the seamless integration of EVPN with L3VPN (MPLS SR) is supported on Cisco Nexus 9332D-H2R switches.

Beginning with Cisco NX-OS Release 10.4(2)F, the seamless integration of EVPN with L3VPN (MPLS SR) is supported on Cisco Nexus 93400LD-H1 switches.

Beginning with Cisco NX-OS Release 10.4(3)F, the seamless integration of EVPN with L3VPN (MPLS SR) is supported on Cisco Nexus 9364C-H1, 9808/9804 switches with X9836DM-A and X98900CD-A line cards.

The following features are supported with seamless integration of EVPN with L3VPN (MPLS SR):

- Host Facing (Downlinks towards)
  - Individual Layer-3 interfaces (orphan ports)
  - Layer-3 Port-Channel
  - Layer-3 Sub-interfaces
  - Inter-AS Option A (often also called VRF-lite)
- Core Facing (Uplinks towards VXLAN)
  - Individual Layer-3 interfaces
  - Layer-3 Port-Channel
- Core Facing (Uplinks towards MPLS SR)
  - Individual Layer-3 interface
  - Per-VRF labels
  - VPN label statistics
- End-to-EndTime to Live (TTL) and Explicit Congestion Notification (ECN) with pipe-mode only.
- MPLS SegmentRouting and MPLS LDP cannot be configured at the same time on a Cisco Nexus 9504 and 9508 platform switches with Cisco Nexus 96136YC-R and Cisco Nexus 9636C-RX line cards.

The VXLAN-to-SR handoff QoS value is preserved during handoff and propagated from VXLAN tunnel packets to SR-tunneled packets for Cisco Nexus 9336C-FX2, 93240YC-FX2, 9300-FX3, and 9300-GX platform switches.

Beginning with Cisco NX-OS Release 10.2(3)F, the VXLAN-to-SR handoff QoS value is preserved during handoff and propagated from VXLAN tunnel packets to SR-tunneled packets on Cisco Nexus 9300-GX2 platform switches.

Beginning with Cisco NX-OS Release 10.4(1)F, the VXLAN-to-SR handoff QoS value is preserved during handoff and propagated from VXLAN tunnel packets to SR-tunneled packets on Cisco Nexus 9332D-H2R switches.

Beginning with Cisco NX-OS Release 10.4(2)F, the VXLAN-to-SR handoff QoS value is preserved during handoff and propagated from VXLAN tunnel packets to SR-tunneled packets on Cisco Nexus 93400LD-H1 switches.

Beginning with Cisco NX-OS Release 10.4(3)F, the VXLAN-to-SR handoff QoS value is preserved during handoff and propagated from VXLAN tunnel packets to SR-tunneled packets on Cisco Nexus 9364C-H1 switches.

The following features are not supported with seamless integration of EVPN with L3VPN (MPLS SR):

- Distributed Anycast Gateway or First-Hop Redundancy Protocol like HSRP, VRRP or GLBP.
- vPC for redundant Host or Network Service attachment.
- SVI/Sub-interfaces for Core facing uplinks (MPLS or VXLAN).
- SVI/Sub-interfaces with configured MAC addresses.
- Beginning with Cisco NX-OS Release 10.5(3)F, MPLS Segment Routing can be enabled on Border Gateway (BGW for VXLAN Multi-Site). In the earlier releases, MPLS Segment Routing and Border Gateway (BGW for VXLAN Multi-Site) cannot be configured at the same time.
- Layer-2 for stretched Subnet across the MPLS-SR domain
- No-drop for VXLAN/SR and SR/VXLAN handoff, for Cisco Nexus 9336C-FX2, 93240YC-FX2, and 9300-FX3 platform switches
- Statistics, for Cisco Nexus 9504 and 9508 platform switches with 96136YC-R and 9636C-RX line cards
- Priority flow control (PFC), for Cisco Nexus 9336C-FX2, 93240YC-FX2, 9300-FX3, and 9300-GX platform switches
- Beginning with Cisco NX-OS Release 10.3(1)F, the DSCP based SRTE traffic steering feature allows source routing of VXLAN packets that are matched using the DSCP fields in the IP header and steered into an SRTE path. Following are the guidelines and limitations for this feature:
  - This feature is supported only on Cisco Nexus 9300-FX2, 9300-FX3, 9300-GX, and 9300-GX2 ToR switches.
  - In case of border leaf or border PE, the ACL filters are applicable on the inner packets (IPv4 access list for IPv4 packets and IPv6 access list for IPv6 packets). This feature is only supported on L3VPN. MPLS EVPN is not supported with VXLAN.
- Beginning with Cisco NX-OS Release 10.3(2)F, seamless integration of EVPN with L3VPN (MPLS SR) is supported on Cisco Nexus 9300-FX platform switches and Cisco Nexus 9700-FX and 9700-GX line cards. Following are the guidelines and limitations for this feature:
  - When Cisco Nexus 9500 platform switch is in a hand-off mode and the MPLS encapsulated packets are forwarded on an L2 port, the dot1q header does not get added.
  - SVI/Sub-interfaces are not supported for core facing uplinks (MPLS or VXLAN) when Cisco Nexus 9500 platform switch is configured as EVPN to MPLS SR L3VPN hand off mode.
  - The DSCP to MPLS EXP promotion does not work on the FX TOR/line cards in DCI Mode. The copying of Inner DSCP value to MPLS EXP does not work on the FX TOR/line cards in this hand off mode. The MPLS EXP will be set to 0x7.
- Beginning with Cisco NX-OS Release 10.3(2)F, the DSCP based SR-TE flow steering feature is supported on Cisco Nexus 9300-FX platform switches and Cisco Nexus 9700-FX and 9700-GX line cards. Following are the guidelines and limitations for this feature:

- When Cisco Nexus 9500 platform switch is in a hand-off mode and the MPLS encapsulated packets are forwarded on an L2 port, the dot1q header does not get added.
- SVI/Sub-interfaces are not supported for core facing uplinks (MPLS or VXLAN) when Cisco Nexus 9500 platform switch is configured as EVPN to MPLS SR L3VPN hand off mode.
- The DSCP to MPLS EXP promotion does not work on the FX TOR/line cards in DCI Mode. The copying of Inner DSCP value to MPLS EXP does not work on the FX TOR/line cards in this hand off mode. The MPLS EXP will be set to 0x7.

- Beginning with Cisco NX-OS Release 10.4(3)F, Cisco Nexus 9808/9804 switches with X9836DM-A and X98900CD-A line cards support the MPLS SR QoS feature only on system level QoS and not at interface level QoS with the following limitations:
  - Default pipe mode is supported, so that the inner packet DSCP or precedence is preserved.
  - For setting MPLS experimental bits in system level QoS policy-map, the following match criteria are supported:
    - Match DSCP
    - Match precedence
  - At system level QoS, the following features are not supported:
    - Policing
    - Policy-map statistics
    - MPLS EXP to DSCP promotion
  - At interface level QoS, the policy with MPLS encapsulation is not supported.
  - Interface level QoS policy take priority over system level QoS policy. Traffic that does not match any criteria in the interface policy will be processed by the default profile in system level QoS.
  - Queuing statistics on MPLS interfaces may erroneously show 'UC ECN Mark pkts'."

Configuring Seamless Integration of EVPN with L3VPN (MPLS SR)

The following procedure for Border Provider Edge (Border PE) imports and reoriginates the routes from the VXLAN domain to the MPLS domain and in the other direction.

SUMMARY STEPS

1. **configure terminal**
2. **feature-set mpls**
3. **nv overlay evpn**
4. **feature bgp**
5. **feature mpls l3vpn**

6. **feature mpls segment-routing**
7. **feature interface-vlan**
8. **feature vn-segment-vlan-based**
9. **feature nv overlay**
10. **router bgp** *autonomous-system-number*
11. **address-family ipv4 unicast**
12. **network** *address*
13. **allocate-label all**
14. **exit**
15. **neighbor** *address* **remote-as** *number*
16. **update-source** *type/id*
17. **address-family l2vpn evpn**
18. **send-community both**
19. **import vpn unicast reoriginate**
20. **exit**
21. **neighbor** *address* **remote-as** *number*
22. **update-source** *type/id*
23. **address-family ipv4 labeled-unicast**
24. **send-community both**
25. **exit**
26. **neighbor** *address* **remote-as** *number*
27. **update-source** *type/id*
28. **ebgp-multihop** *number*
29. **address-family vpnv4 unicast**
30. **send-community both**
31. **import l2vpn evpn reoriginate**
32. **exit**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**feature-set mpls** **Example:** `switch(config)# feature-set mpls`	Enables the MPLS feature set.
Step 3	**nv overlay evpn** **Example:** `switch(config)# nv overlay evpn`	Enables VXLAN.

	Command or Action	Purpose
Step 4	**feature bgp** **Example:** `switch(config)# feature bgp`	Enables BGP.
Step 5	**feature mpls l3vpn** **Example:** `switch(config)# feature mpls l3vpn`	Enables Layer 3 VPN. **Note** Feature mpls l3vpn requires feature mpls segment-routing.
Step 6	**feature mpls segment-routing** **Example:** `switch(config)# feature mpls segment-routing`	Enables Segment Routing.
Step 7	**feature interface-vlan** **Example:** `switch(config)# feature interface-vlan`	Enables the interface VLAN.
Step 8	**feature vn-segment-vlan-based** **Example:** `switch(config)# feature vn-segment-vlan-based`	Enables the VLAN-based VN segment.
Step 9	**feature nv overlay** **Example:** `switch(config)# feature nv overlay`	Enables VXLAN.
Step 10	**router bgp** *autonomous-system-number* **Example:** `switch(config)# router bgp 65095`	Configures BGP. The value of *autonomous-system-number* is from 1 to 4294967295.
Step 11	**address-family ipv4 unicast** **Example:** `switch(config-router)# address-family ipv4 unicast`	Configures the address family for IPv4.
Step 12	**network** *address* **Example:** `switch(config-router-af)# network 10.51.0.51/32`	Injects prefixes into BGP for the MPLS-SR domain. **Note** All viable next-hops for MPLS-SR tunnel deposition on the Border PE must be advertised via the network statement (/32 only).
Step 13	**allocate-label all** **Example:** `switch(config-router-af)# allocate-label all`	Configures label allocation for every prefix injected via the network statement.
Step 14	**exit** **Example:**	Exits command mode.

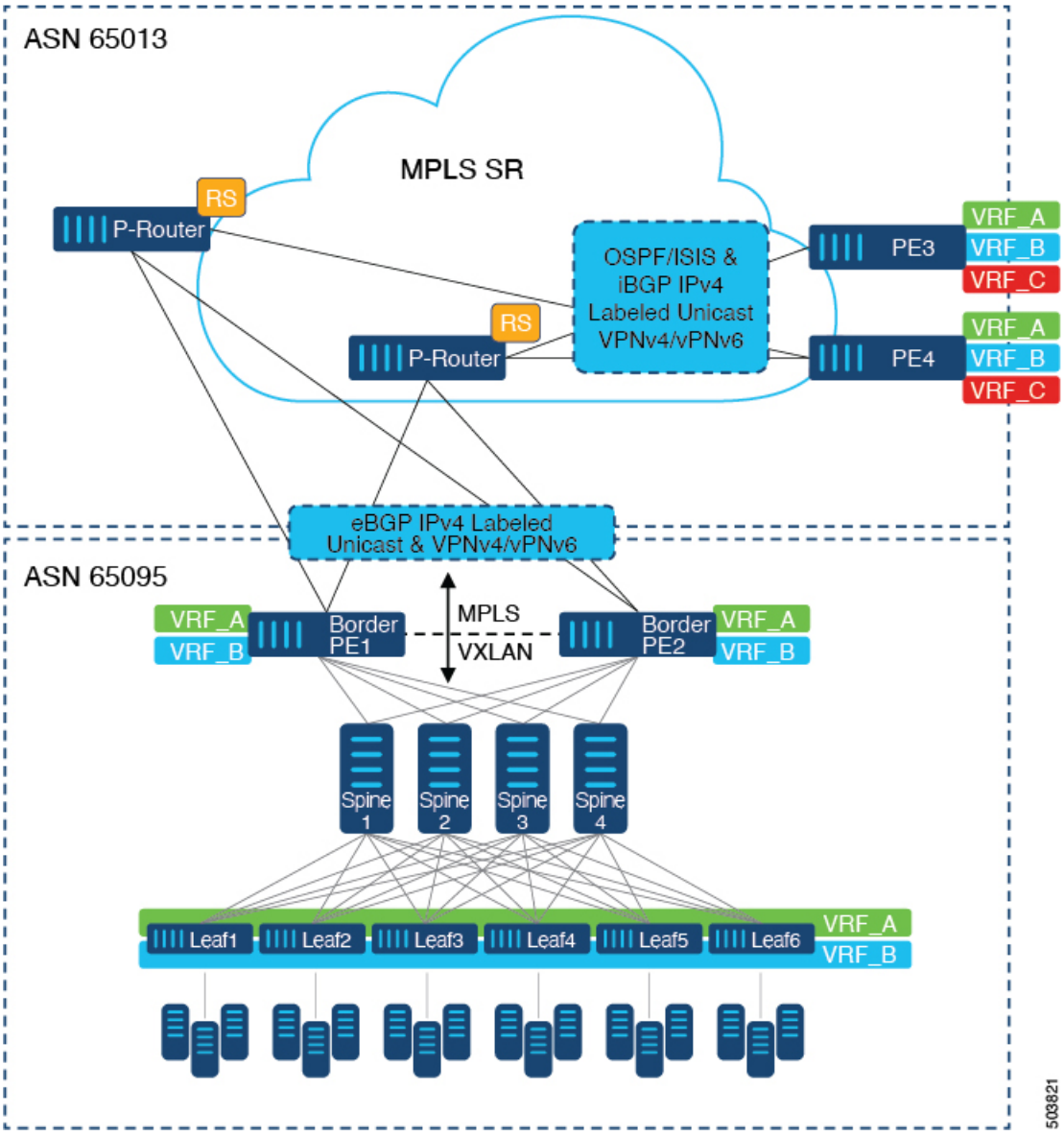
	Command or Action	Purpose
	`switch(config-router-af)# exit`	
Step 15	**neighbor** *address* **remote-as** *number* **Example:** `switch(config-router)# neighbor 10.95.0.95 remote-as 65095`	Defines the iBGP neighbor IPv4 address and remote Autonomous-System (AS) number towards the Route-Reflector.
Step 16	**update-source** *type/id* **Example:** `switch(config-router)# update-source loopback0`	Defines the interface for eBGP peering.
Step 17	**address-family l2vpn evpn** **Example:** `switch(config-router)# address-family l2vpn evpn`	Configures the L2VPN EVPN address family.
Step 18	**send-community both** **Example:** `switch(config-router-af)# send-community both`	Configures the community for BGP neighbors.
Step 19	**import vpn unicast reoriginate** **Example:** `switch(config-router-af)# import vpn unicast reoriginate`	Reoriginates the route with a new Route-Target. It can be extended to use an optional route-map.
Step 20	**exit** **Example:** `switch(config-router-af)# exit`	Exits command mode.
Step 21	**neighbor** *address* **remote-as** *number* **Example:** `switch(config-router)# neighbor 10.51.131.131 remote-as 65013`	Defines the eBGP neighbor IPv4 address and remote Autonomous-System (AS) number towards the P-Router.
Step 22	**update-source** *type/id* **Example:** `switch(config-router)# update-source Ethernet1/1`	Defines the interface for eBGP peering.
Step 23	**address-family ipv4 labeled-unicast** **Example:** `switch(config-router)# address-family ipv4 labeled-unicast`	Configures the address family for IPv4 labeled-unicast.
Step 24	**send-community both** **Example:** `switch(config-router-af)# send-community both`	Configures the community for BGP neighbors.

	Command or Action	Purpose
Step 25	**exit** **Example:** `switch(config-router-af)# exit`	Exits command mode.
Step 26	**neighbor** *address* **remote-as** *number* **Example:** `switch(config-router)# neighbor 10.131.0.131 remote-as 65013`	Defines the eBGP neighbor IPv4 address and remote Autonomous-System (AS) number.
Step 27	**update-source** *type/id* **Example:** `switch(config-router)# update-source loopback0`	Defines the interface for eBGP peering.
Step 28	**ebgp-multihop** *number* **Example:** `switch(config-router)# ebgp-multihop 5`	Specifies multihop TTL for the remote peer. The range of *number*is from 2 to 255.
Step 29	**address-family vpnv4 unicast** **Example:** `switch(config-router)# address-family vpnv4 unicast`	Configures the address family for VPNv4 or VPNv6.
Step 30	**send-community both** **Example:** `switch(config-router-af)# send-community both`	Configures the community for BGP neighbors.
Step 31	**import l2vpn evpn reoriginate** **Example:** `switch(config-router-af)# import l2vpn evpn reoriginate`	Reoriginates the route with a new Route-Target. It can be extended to use an optional route-map.
Step 32	**exit** **Example:** `switch(config-router-af)# exit`	Exits command mode.

Example Configuration for Configuring Seamless Integration of EVPN with L3VPN (MPLS SR)

Scenario - 1 with DC to Core Network Domain Separation and IGP within MPLS-SR network.

Figure 24: Topology with DC to Core Network Domain Separation



The following is a sample CLI configuration that is required to import and reoriginate the routes from the VXLAN domain to the MPLS domain and in the reverse direction. The sample CLI configuration represents only the necessary configuration for the respective roles.

Border PE

```
hostname BL51-N9336FX2
install feature-set mpls

feature-set mpls

feature bgp
feature mpls l3vpn
feature mpls segment-routing
```

```
feature ospf
feature interface-vlan
feature vn-segment-vlan-based
feature nv overlay

nv overlay evpn

mpls label range 16000 23999 static 6000 8000

segment-routing
  mpls
    connected-prefix-sid-map
      address-family ipv4
        10.51.0.51/32 index 51

vlan 2000
  vn-segment 50000

vrf context VRF_A
  vni 50000
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
    route-target import 50000:50000
    route-target export 50000:50000
  address-family ipv6 unicast
    route-target both auto
    route-target both auto evpn
    route-target import 50000:50000
    route-target export 50000:50000

interface Vlan2000
  no shutdown
  vrf member VRF_A
  no ip redirects
  ip forward
  ipv6 address use-link-local-only
  no ipv6 redirects

interface nve1
  no shutdown
  host-reachability protocol bgp
  source-interface loopback1
  member vni 50000 associate-vrf

interface Ethernet1/1
  description TO_P-ROUTER
  ip address 10.51.131.51/24
  mpls ip forwarding
  no shutdown

interface Ethernet1/36
  description TO_SPINE
  ip address 10.95.51.51/24
  ip router ospf 10 area 0.0.0.0
  no shutdown

interface loopback0
  description ROUTER-ID & SR-LOOPBACK
  ip address 10.51.0.51/32
  ip router ospf UNDERLAY area 0.0.0.0

interface loopback1
```

```
  description NVE-LOOPBACK
  ip address 10.51.1.51/32
  ip router ospf UNDERLAY area 0.0.0.0

router ospf UNDERLAY
  router-id 10.51.0.51

router bgp 65095
  address-family ipv4 unicast
    network 10.51.0.51/32
    allocate-label all
!
  neighbor 10.95.0.95
    remote-as 65095
    update-source loopback0
    address-family l2vpn evpn
      send-community
      send-community extended
      import vpn unicast reoriginate
!
  neighbor 10.51.131.131
    remote-as 65013
    update-source Ethernet1/1
    address-family ipv4 labeled-unicast
      send-community
      send-community extended
!
  neighbor 10.131.0.131
    remote-as 65013
    update-source loopback0
    ebgp-multihop 5
    address-family vpnv4 unicast
      send-community
      send-community extended
      import l2vpn evpn reoriginate
    address-family vpnv6 unicast
      send-community
      send-community extended
      import l2vpn evpn reoriginate

!
  vrf VRF_A
    address-family ipv4 unicast
      redistribute direct route-map fabric-rmap-redist-subnet
```

P-Router

```
hostname P131-N9336FX2
install feature-set mpls

feature-set mpls

feature bgp
feature isis
feature mpls l3vpn
feature mpls segment-routing

mpls label range 16000 23999 static 6000 8000

segment-routing
  mpls
    connected-prefix-sid-map
      address-family ipv4
        10.131.0.131/32 index 131
```

```
route-map RM_NH_UNCH permit 10
  set ip next-hop unchanged

interface Ethernet1/1
  description TO_BORDER-PE
  ip address 10.51.131.131/24
  ip router isis 10
  mpls ip forwarding
  no shutdown

interface Ethernet1/11
  description TO_PE
  ip address 10.52.131.131/24
  ip router isis 10
  mpls ip forwarding
  no shutdown

interface loopback0
  description ROUTER-ID & SR-LOOPBACK
  ip address 10.131.0.131/32
  ip router isis 10

router isis 10
  net 49.0000.0000.0131.00
  is-type level-2
  address-family ipv4 unicast
    segment-routing mpls

router bgp 65013
  event-history detail
  address-family ipv4 unicast
    allocate-label all
!
  neighbor 10.51.131.51
    remote-as 65095
    update-source Ethernet1/1
    address-family ipv4 labeled-unicast
      send-community
      send-community extended
!
  neighbor 10.51.0.51
    remote-as 65095
    update-source loopback0
    ebgp-multihop 5
    address-family vpnv4 unicast
      send-community
      send-community extended
      route-map RM_NH_UNCH out
    address-family vpnv6 unicast
      send-community
      send-community extended
      route-map RM_NH_UNCH out
!
  neighbor 10.52.131.52
    remote-as 65013
    update-source Ethernet1/11
    address-family ipv4 labeled-unicast
      send-community
      send-community extended
!
  neighbor 10.52.0.52
    remote-as 65013
    update-source loopback0
    address-family vpnv4 unicast
```

```
      send-community
      send-community extended
      route-reflector-client
      route-map RM_NH_UNCH out
    address-family vpnv6 unicast
      send-community
      send-community extended
      route-reflector-client
      route-map RM_NH_UNCH out
```

Provider Edge (PE)

```
hostname L52-N93240FX2
install feature-set mpls

feature-set mpls

feature bgp
feature isis
feature mpls l3vpn
feature mpls segment-routing

mpls label range 16000 23999 static 6000 8000

segment-routing
  mpls
    connected-prefix-sid-map
      address-family ipv4
        10.52.0.52/32 index 52

vrf context VRF_A
  rd auto
  address-family ipv4 unicast
    route-target import 50000:50000
    route-target export 50000:50000
  address-family ipv6 unicast
    route-target import 50000:50000
    route-target export 50000:50000

interface Ethernet1/49
  description TO_P-ROUTER
  ip address 10.52.131.52/24
  ip router isis 10
  mpls ip forwarding
  no shutdown

interface loopback0
  description ROUTER-ID & SR-LOOPBACK
  ip address 10.52.0.52/32
  ip router isis 10

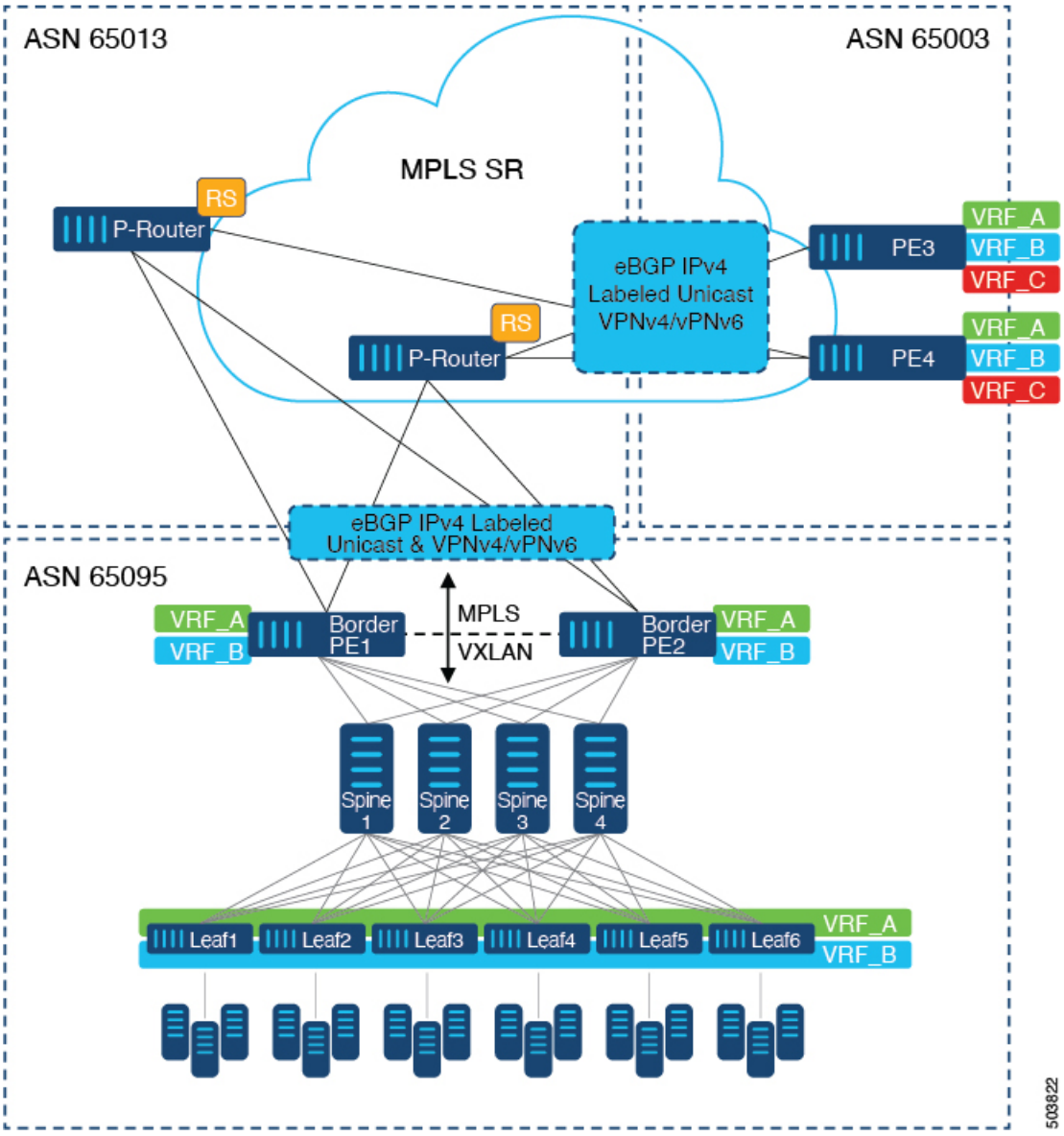
router isis 10
  net 49.0000.0000.0052.00
  is-type level-2
  address-family ipv4 unicast
    segment-routing mpls

router bgp 65013
  address-family ipv4 unicast
    network 10.52.0.52/32
    allocate-label all
!
  neighbor 10.52.131.131
    remote-as 65013
    update-source Ethernet1/49
```

```
    address-family ipv4 labeled-unicast
      send-community
      send-community extended
!
  neighbor 10.131.0.131
    remote-as 65013
    update-source loopback0
    address-family vpnv4 unicast
      send-community
      send-community extended
    address-family vpnv6 unicast
      send-community
      send-community extended
!
  vrf VRF_A
    address-family ipv4 unicast
      redistribute direct route-map fabric-rmap-redist-subnet
```

Scenario - 2 with DC to Core and within Core Network Domain Separation (eBGP within MPLS-SR network).

Figure 25: Multiple Administrative Domains within the Core network



The following is a sample CLI configuration that is required to import and reoriginate the routes from the VXLAN domain to the MPLS domain and in the reverse direction. The sample CLI configuration represents only the nodes that are different from Scenario #1, which are the P-Router and the Provider Edge (PE) roles. The Border PE remains the same for both scenarios.

P-Router

```
hostname P131-N9336FX2
install feature-set mpls

feature-set mpls

feature bgp
feature mpls l3vpn
feature mpls segment-routing
```

```
mpls label range 16000 23999 static 6000 8000

segment-routing
  mpls
    connected-prefix-sid-map
      address-family ipv4
        10.131.0.131/32 index 131

route-map RM_NH_UNCH permit 10
  set ip next-hop unchanged

interface Ethernet1/1
  description TO_BORDER-PE
  ip address 10.51.131.131/24
  mpls ip forwarding
  no shutdown

interface Ethernet1/11
  description TO_PE
  ip address 10.52.131.131/24
  mpls ip forwarding
  no shutdown

interface loopback0
  description ROUTER-ID & SR-LOOPBACK
  ip address 10.131.0.131/32
  ip router isis 10

router bgp 65013
  event-history detail
  address-family ipv4 unicast
    network 10.131.0.131/32
    allocate-label all
!
  address-family vpnv4 unicast
    retain route-target all
  address-family vpnv6 unicast
    retain route-target all
!
  neighbor 10.51.131.51
    remote-as 65095
    update-source Ethernet1/1
    address-family ipv4 labeled-unicast
      send-community
      send-community extended
!
  neighbor 10.51.0.51
    remote-as 65095
    update-source loopback0
    ebgp-multihop 5
    address-family vpnv4 unicast
      send-community
      send-community extended
      route-map RM_NH_UNCH out
    address-family vpnv6 unicast
      send-community
      send-community extended
      route-map RM_NH_UNCH out
!
  neighbor 10.52.131.52
    remote-as 65003
    update-source Ethernet1/11
    address-family ipv4 labeled-unicast
```

```
      send-community
      send-community extended
!
  neighbor 10.52.0.52
    remote-as 65003
    update-source loopback0
    ebgp-multihop 5
    address-family vpnv4 unicast
      send-community
      send-community extended
      route-map RM_NH_UNCH out
    address-family vpnv6 unicast
      send-community
      send-community extended
      route-map RM_NH_UNCH out
```

Provider Edge (PE)

```
hostname L52-N93240FX2
install feature-set mpls

feature-set mpls

feature bgp
feature mpls l3vpn
feature mpls segment-routing

mpls label range 16000 23999 static 6000 8000

segment-routing
  mpls
    connected-prefix-sid-map
      address-family ipv4
        10.52.0.52/32 index 52

vrf context VRF_A
  rd auto
  address-family ipv4 unicast
    route-target import 50000:50000
    route-target export 50000:50000
  address-family ipv6 unicast
    route-target import 50000:50000
    route-target export 50000:50000

interface Ethernet1/49
  description TO_P-ROUTER
  ip address 10.52.131.52/24
  mpls ip forwarding
  no shutdown

interface loopback0
  description ROUTER-ID & SR-LOOPBACK
  ip address 10.52.0.52/32
  ip router isis 10

router bgp 65003
  address-family ipv4 unicast
    network 10.52.0.52/32
    allocate-label all
!
  neighbor 10.52.131.131
    remote-as 65013
    update-source Ethernet1/49
    address-family ipv4 labeled-unicast
      send-community
```

```
      send-community extended
!
  neighbor 10.131.0.131
    remote-as 65013
    update-source loopback0
    ebgp-multihop 5
    address-family vpnv4 unicast
      send-community
      send-community extended
    address-family vpnv6 unicast
      send-community
      send-community extended
!
  vrf VRF_A
    address-family ipv4 unicast
      redistribute direct route-map fabric-rmap-redist-subnet
```

Configuring DSCP Based SR-TE Flow Steering

To configure DSCP based SR-TE flow steering, first configure the border PE or border leaf for seamless integration of EVPN with L3VPN; see Configuring Seamless Integration of EVPN with L3VPN (MPLS SR), on page 251. Then, to steer the traffic, perform the following configuration:

1. Configure SRTE policy. See *Configuration Process: SRTE Flow-based Traffic Steering* section under the *Configuring Segment Routing* chapter in the *Cisco Nexus 9000 Series NX-OS Label Switching Configuration Guide* on the Cisco portal.
2. Configure the L3 VNI interface. See Configuring New L3VNI Mode.
3. Apply the policy on the L3 VNI interface using the **ip/ipv6 policy route-map** *srte-policy* command.

Configuration Example for DSCP Based SR-TE Flow Steering

```
segment-routing
  traffic-engineering
    segment-list name PATH1
      index 50 mpls label 16100
    segment-list name PATH2
      index 50 mpls label 16500
      index 100 mpls label 16100

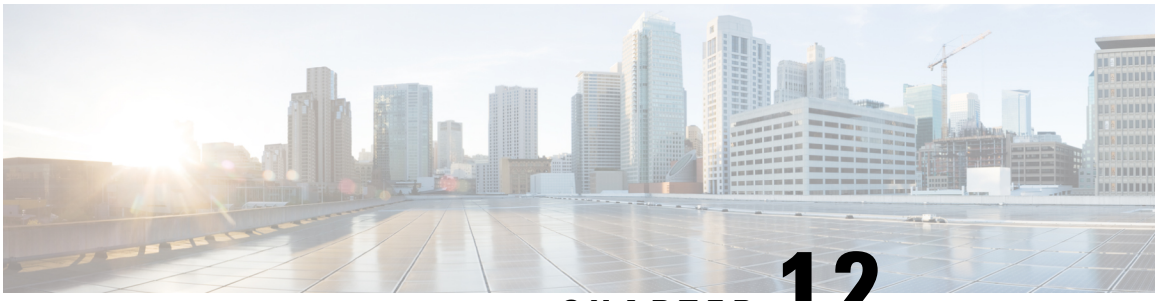
    policy blue
      color 202 endpoint 21.1.1.1
      candidate-paths
        preference 100
          explicit segment-list PATH2
    policy red
      color 201 endpoint 21.1.1.1
      candidate-paths
        preference 100
          explicit segment-list PATH1
ip access-list flow-1
  statistics per-entry
  5 permit ip any any dscp af11
ip access-list flow-2
  statistics per-entry
  5 permit ip any any dscp af12
```

```
route-map srte-flow1 permit 10
  match ip address flow-1
  set ip next-hop 61.1.1.1 srte-policy name red

route-map srte-flow1 permit 20
  match ip address flow-2
  set ip next-hop 61.1.1.1 srte-policy name blue

vrf context 501
  vni 90001 l3

interface vni90001
  ip policy route-map srte-flow1
```



CHAPTER 12

Configuring Seamless Integration of EVPN with L3VPN SRv6

This chapter contains the following sections:

- About Seamless Integration of EVPN with L3VPN SRv6 Handoff, on page 273
- Guidelines and Limitations for EVPN to L3VPN SRv6 Handoff, on page 274
- Importing L3VPN SRv6 Routes into EVPN VXLAN, on page 275
- Importing EVPN VXLAN Routes into L3VPN SRv6, on page 277
- Example Configuration for VXLAN EVPN to L3VPN SRv6 Handoff, on page 278

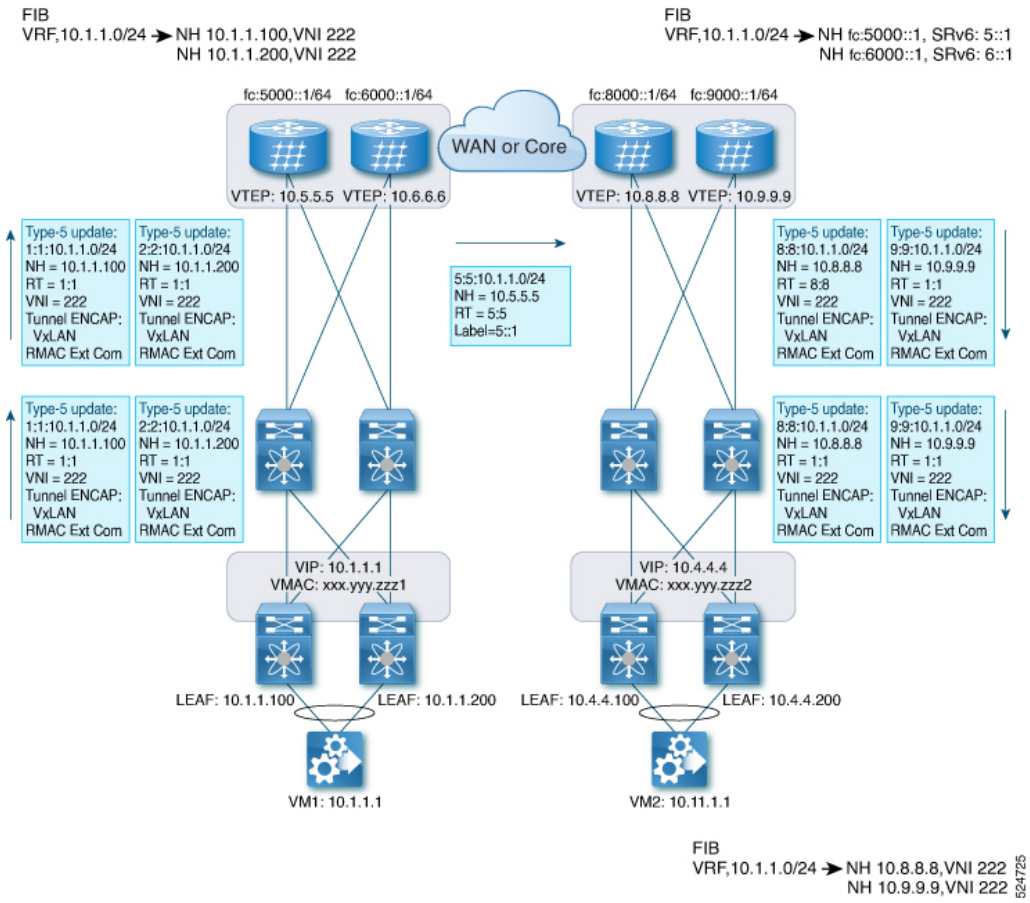
About Seamless Integration of EVPN with L3VPN SRv6 Handoff

Data Center (DC) deployments have adopted VXLAN EVPN for its benefits such as EVPN control-plane learning, multitenancy, seamless mobility, redundancy, and easier POD additions. Similarly, the CORE is either an IP-based L3VPN SRv6 network or transitioning from the IPv6-based L3VPN underlay to a more sophisticated solution like IPv6 Segment Routing (SRv6) for IPv6. SRv6 is adopted for its benefits such as:

- Simpler traffic engineering (TE) methods
- Easier configuration
- SDN adoption

With two different technologies, one within the data center (DC) and one in the Core, there is traffic handoff from VXLAN to an SRv6 core that becomes a necessity at the DCI nodes, which sit at the edge of the DC domain and interface with the Core edge router.

Figure 26: BGP EVPN VXLAN to L3VPN SRv6 Handoff



For traffic ingressing the EVPN-VxLAN fabric, the BGP EVPN routes get imported into a local VRF which contains the RD of the VRF. The bestpath is calculated and installed in the VRF's RIB, then inserted into the L3VPN SRv6 table. Along with the bestpath, the VRF's RD and per-VRF SRv6 SID are included. The L3VPN SRv6 route target is sent with the route, which is advertised to the L3VPN SRv6 peer.

For traffic egressing the EVPN VxLAN fabric, the BGP L3VPN SRv6 routes get imported into a local VRF which contains the RD of the VRF. The bestpath is calculated and installed in the VRF's RIB, then inserted into the EVPN table. Along with the bestpath, the VRF's RD and VNI are included. The EVPN-VXLAN route target is sent with the route, which is advertised to the EVPN-VxLAN peer.

Guidelines and Limitations for EVPN to L3VPN SRv6 Handoff

This feature has the following guidelines and limitations:

- The same RD import is supported for L3VPN SRV6 fabrics.
- The same RD import is not supported for EVPN VXLAN fabrics.
- On a handoff device, do not use the same RD import on the EVPN VXLAN side.
- Beginning with Cisco NX-OS Release 9.3(3), support is added for the following switches:

  - Cisco Nexus C93600CD-GX
  - Cisco Nexus C9364C-GX
  - Cisco Nexus C9316D-GX

- Beginning with Cisco NX-OS Release 10.2(1q)F, SRv6 DCI handoff is supported on Cisco Nexus 9332D-GX2B platform switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, SRv6 DCI handoff is supported on Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, SRv6 DCI handoff is supported on Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, SRv6 DCI handoff is supported on Cisco Nexus 9364C-H1 switches.
- Beginning with Cisco NX-OS Release 10.2(3)F, EVPN to L3VPN SRv6 Handoff is supported on Cisco Nexus 9300-GX2 platform switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, EVPN to L3VPN SRv6 Handoff is supported on Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, EVPN to L3VPN SRv6 Handoff is supported on Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, EVPN to L3VPN SRv6 Handoff is supported on Cisco Nexus 9364C-H1 switches.

Importing L3VPN SRv6 Routes into EVPN VXLAN

The process of handing off routes from the L3VPN SRv6 domain to the EVPN VXLAN fabric requires configuring the import condition for L3VPN SRv6 routes. Routes can be either IPv4 or IPv6. This task configures unidirectional route advertisement into the EVPN VXLAN fabric. For bidirectional advertisement, you must explicitly configure the import condition for the L3VPN SRv6 domain.

Before you begin

Make sure you have a fully configured L3VPN SRv6 fabric. For more information, see "Configuring Layer 3 VPN over SRv6" in the *Cisco Nexus 9000 Series NX-OS SRv6 Configuration Guide*.

SUMMARY STEPS

1. **config terminal**
2. **router bgp** *as-number*
3. **neighbor bgp** *ipv6-address* **remote-as** *as-number*
4. **address family vpnv4 unicast** or **address family vpnv6 unicast**
5. **import l2vpn evpn route-map** *name* **[reoriginate]**

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**config terminal**
	Example:
	`switch-1# config terminal`
	`Enter configuration commands, one per line. End with CNTL/Z.`
	`switch-1(config)#`
		Enter configuration mode.

Step 2	**router bgp** *as-number*
	Example:
	`switch-1(config)# router bgp 100`
	`switch-1(config-router)#`
		Enter BGP router configuration mode.

Step 3	**neighbor bgp** *ipv6-address* **remote-as** *as-number*
	Example:
	`switch-1(config-router)# neighbor fc:1234::1 remote-as 200`
	`switch-1(config-router-neighbor)#`
		Enter BGP router configuration mode.

Step 4	**address family vpnv4 unicast** or **address family vpnv6 unicast**
	Example:
	`switch-1(config-router-neighbor)# address-family vpnv4 unicast`
	`switch-1(config-router-neighbor-af)#`
	Example:
	`switch-1(config-router-neighbor)# address-family vpnv6 unicast`
	`switch-1(config-router-neighbor-af)#`
		Configure the IPv4 or IPv6 address family for unicast traffic that the EVPN VXLAN will handoff to L3VPN SRv6.

Step 5	**import l2vpn evpn route-map** *name* **[reoriginate]**
	Example:
	`switch-1(config-router-neighbor-af)# import l2vpn evpn route-map test reoriginate`
	`switch-1(config-router-neighbor-af)#`
		Configure the IPv4 or IPv6 address family for unicast traffic that EVPN VXLAN will handoff to L3VPN SRv6. This command enables routes learned from L3VPN SRv6 domain to be advertised to the EVPN VXLAN domain. Using the optional **reoriginate** keyword advertises only domain-specific RTs.

What to do next

For bidirectional route advertisement, configure importing EVPN VXLAN routes into the L3VPN SRv6 domain.

Importing EVPN VXLAN Routes into L3VPN SRv6

The process of handing off routes from the EVPN VXLAN fabric to the L3VPN SRv6 domain requires configuring the import condition for EVPN VXLAN routes. Routes can be either IPv4 or IPv6. This task configures unidirectional route advertisement into the L3VPN SRv6 fabric. For bidirectional advertisement, you must explicitly configure the import condition for the EVPN VXLAN domain.

Before you begin

Make sure you have a fully configured L3VPN SRv6 fabric. For more information, see "Configuring Layer 3 VPN over SRv6" in the *Cisco Nexus 9000 Series NX-OS SRv6 Configuration Guide*.

SUMMARY STEPS

1. **config terminal**
2. **router bgp** *as-number*
3. **neighbor** *ipv6-address* **remote-as** *as-number*
4. **address-family l2vpn evpn**
5. **import vpn unicast route-map** *name* **[reoriginate]**

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**config terminal**
	Example:
	`switch-1# config terminal`
	`Enter configuration commands, one per line. End with CNTL/Z.`
	`switch-1(config)#`
		Enter configuration mode.

Step 2	**router bgp** *as-number*
	Example:
	`switch-1(config)# router bgp 200`
	`switch-1(config-router)#`
		Enter BGP router configuration mode.

Step 3	**neighbor** *ipv6-address* **remote-as** *as-number*
	Example:
	`switch-1(config-router)# neighbor fc:1234::1 remote-as 100`
	`switch-1(config-router-neighbor)#`
		Enter BGP router configuration mode.

Step 4	**address-family l2vpn evpn**
	Example:
	`switch(config-router-neighbor)# address-family l2vpn evpn`
	`switch(config-router-neighbor-af)#`
		Configure the address family for unicast traffic that EVPN VXLAN will handoff to L3VPN SRv6.

	Command or Action	Purpose

Step 5	**import vpn unicast route-map** *name* **[reoriginate]**
	Example:
	`switch-1(config-router-neighbor-af)# import vpn unicast route-map test reoriginate`
	`switch-1(config-router-neighbor-af)#`
		Configure the IPv4 or IPv6 address family for unicast traffic that EVPN VXLAN will handoff to L3VPN SRv6. This command enables routes learned from the EVPN VXLAN domain to be advertised to the L3VPN SRv6 domain. Using the optional **reoriginate** keyword advertises only domain-specific RTs.

What to do next

For bidirectional route advertisement, configure importing L3VPN SRv6 routes into the EVPN VXLAN fabric.

Example Configuration for VXLAN EVPN to L3VPN SRv6 Handoff

```
feature vn-segment-vlan-based
feature nv overlay
feature interface-vlan
nv overlay evpn
feature srv6

vrf context customer1
    vni 10000
    rd auto
    address-family ipv4 unicast
     route-target both 1:1
     route-target both auto evpn
    address-family ipv6 unicast
     route-target both 1:1
     route-target both auto evpn

segment-routing
  srv6
    encapsulation
     source-address loopback1
    locators
     locator DCI_1
      prefix café:1234::/64

interface loopback0
  ip address 10.1.1.0/32

interface loopback1
  ip address 10.1.1.1/32
  ipv6 address fc:4567::1/128

interface nve1
  source-interface loopback0
  member vni 10000 associate-vrf
  host-reachability protocol bgp

vlan 100
  vn-segment 10000

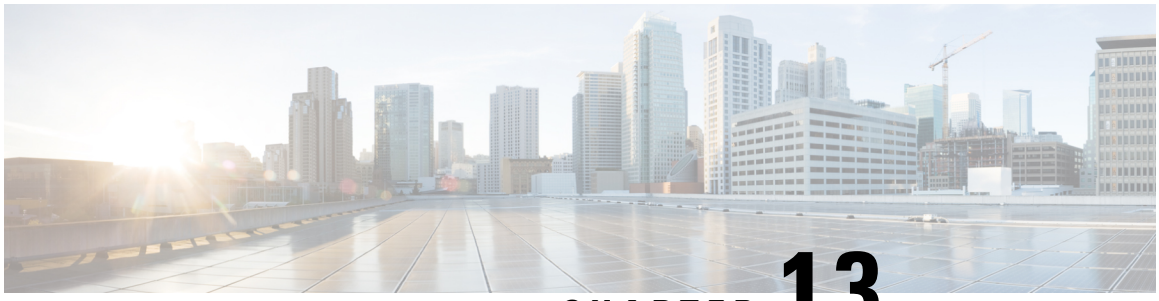
interface vlan 100
 ip forward
```

```
 ipv6 address use-link-local-only
 vrf member customer1

router bgp 65000
   segment-routing srv6
    locator DCI_1
   neighbor 10.2.2.2 remote-as 200
      remote-as 75000
      address-family l2vpn evpn
      import vpn route-map | reoriginate
   neighbor fc:1234::1 remote-as 100
      remote-as 65000
      address-family vpnv4 unicast
      import l2vpn evpn route-map | reoriginate
      address-family vpnv6 unicast
      import l2vpn evpn route-map | reoriginate

vrf customer
  segment-routing srv6
  alloc-mode per-vrf
  address-family ipv4 unicast
  address-family ipv6 unicast
```

Note In the **vni number** command, do not use the **L3** keyword during configuration of VNI under VRF, as the new L3 VNI configuration is not supported on VLAN-BD for VNIs which are dynamically allocated.



CHAPTER 13

Configuring Seamless Integration of EVPN (TRM) with MVPN

This chapter contains the following sections:

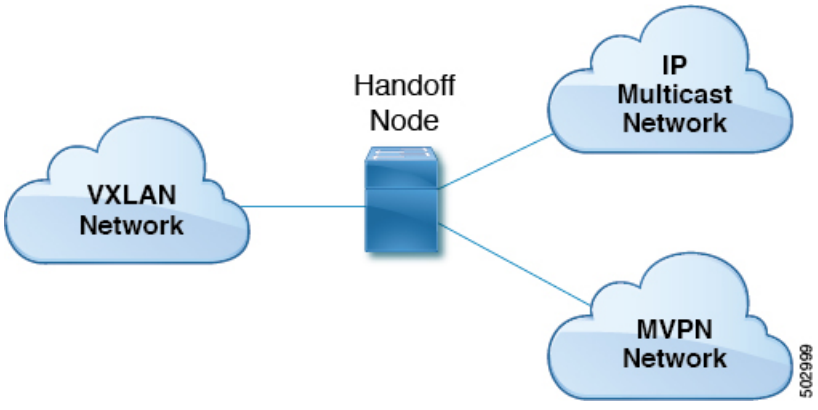
- About Seamless Integration of EVPN (TRM) with MVPN (Draft Rosen), on page 281
- Guidelines and Limitations for Seamless Integration of EVPN (TRM) with MVPN , on page 282
- Configuring the Handoff Node for Seamless Integration of EVPN (TRM) with MVPN, on page 283
- Configuration Example for Seamless Integration of EVPN (TRM) with MVPN , on page 288

About Seamless Integration of EVPN (TRM) with MVPN (Draft Rosen)

Seamless integration of EVPN (TRM) with MVPN (draft rosen) enables packets to be handed off between a VXLAN network (TRM or TRM Multi-Site) and an MVPN network. To support this feature, VXLAN TRM and MVPN must be supported on a Cisco Nexus device node, the handoff node.

The handoff node is the PE for the MVPN network and the VTEP for the VXLAN network. It connects to the VXLAN, MVPN, and IP multicast networks, as shown in the following figure.

Figure 27: VXLAN - MVPN Handoff Network



Sources and receivers can be in any of the three networks (VXLAN, MVPN, or IP multicast).

All multicast traffic (that is, the tenant traffic from the VXLAN, MVPN, or multicast network) is routed from one domain to another domain. The handoff node acts as the central node. It performs the necessary packet forwarding, encapsulation, and decapsulation to send the traffic to the respective receivers.

Supported RP Positions

The rendezvous point (RP) for the customer (overlay) network can be in any of the three networks (VXLAN, MVPN, or IP multicast).

Table 6: Supported RP Locations

RP Locations	Description
RP in IP network	• The RP can be connected only to the MVPN PE and not to the handoff nodes. • The RP can be connected only to the VXLAN handoff nodes. • The RP can be connected to both the MVPN PE and VXLAN.
RP internal to VXLAN fabric	All VTEPs are RPs inside the VXLAN fabric. All MVPN PEs use the RP configured on the VXLAN fabric.
RP on VXLAN MVPN handoff node	The RP is the VXLAN MVPN handoff node.
RP in MVPN network	The RP is external to the VXLAN network. It's configured on one of the nodes in the MPLS cloud, other than the handoff node.
RP Everywhere (PIM Anycast RP or MSDP-based Anycast RP)	The Anycast RP can be configured on the VXLAN leaf. The RP set can be configured on the handoff node or any MVPN PE.

Guidelines and Limitations for Seamless Integration of EVPN (TRM) with MVPN

This feature has the following guidelines and limitations:

- Only Cisco Nexus 9504 and 9508 platform switches with the N9K-X9636C-RX line card support seamless integration of EVPN (TRM) with MVPN. Other -R Series line cards can't function as the handoff node.
- The handoff node can have local (directly connected) multicast sources or receivers for the customer network.
- Any existing underlay properties, such as ASM/SSM for MVPN or ASM for TRM, are supported on the handoff node.
- The handoff node supports PIM SSM and ASM for the overlay.

- Inter-AS option A is supported on the handoff node toward the IP multicast network.
- If the number of MDT source loopback IP addresses and NVE loopback IP addresses exceeds the maximum limit, traffic drops might occur.
- The following functionality isn't supported for seamless integration of EVPN (TRM) with MVPN:
  - vPC on the handoff node
  - VXLAN ingress replication
  - SVIs and subinterfaces as core-facing interfaces for MVPN
  - Inter-AS options B and C on MVPN nodes
  - PIM SSM as a VXLAN underlay
  - Bidirectional PIM as an underlay or overlay
  - ECMP with a mix of MPLS and IP paths
- Any existing limitations for VXLAN, TRM, and MVPN also apply to seamless integration of EVPN (TRM) with MVPN.

Configuring the Handoff Node for Seamless Integration of EVPN (TRM) with MVPN

This section documents the configurations that are required on the handoff node. Configurations for other nodes (such as VXLAN leafs and spines, MVPN PE, and RS/RR) are the same as in previous releases.

PIM/IGMP Configuration for the Handoff Node

Follow these guidelines when configuring PIM/IGMP for the handoff node:

- Make sure that the Rendezvous Point (RP) is different for TRM and the MVPN underlay, as shown in the following example.

```
ip pim rp-address 90.1.1.100 group-list 225.0.0.0/8 --- TRM Underlay
ip pim rp-address 91.1.1.100 group-list 233.0.0.0/8 --- MVPN Underlay
```

- Use a common RP for overlay multicast traffic.
- The RP can be in static, PIM Anycast, or PIM MSDP mode. The following example shows the RP configuration inside the VRF:

```
vrf context vrfVxLAN5001
  vni 5001
  ip pim rp-address 111.1.1.1 group-list 226.0.0.0/8
  ip pim rp-address 112.2.1.1 group-list 227.0.0.0/8
```

- Enable IGMP snooping for VXLAN traffic using the **ip igmp snooping vxlan** command.
- Enable PIM sparse mode on all source interfaces and interfaces required to carry PIM traffic.

BGP Configuration for the Handoff Node

Follow these guidelines when configuring BGP for the handoff node:

- Add all VXLAN leafs as L2EVPN and TRM neighbors; include the redundant handoff node. If a route reflector is used, add only RR as a neighbor.
- Add all MVPN PEs as VPN neighbors. In MDT mode, add the MVPN PEs as MDT neighbors.
- Import configuration to advertise unicast routes from L2EVPN neighbors to VPN neighbors and vice versa.
- The BGP source identifier can be different or the same as the source interfaces used for the VTEP identifier (configured under the NVE interface)/MVPN PE identifier.

```
feature bgp
address-family ipv4 mdt
address-family ipv4 mvpn

neighbor 2.1.1.1
  address-family ipv4 mvpn
    send-community extended
  address-family l2vpn evpn
    send-community extended
    import vpn unicast reoriginate

neighbor 30.30.30.30
  address-family vpnv4 unicast
    send-community
    send-community extended
    next-hop-self
    import l2vpn evpn reoriginate
  address-family ipv4 mdt
    send-community extended
    no next-hop-third-party
```

- Never use Inter-AS option B between MVPN peers. Instead, configure the **no allocate-label option-b** command under the VPNv4 unicast address family.

```
address-family vpnv4 unicast
    no allocate-label option-b
```

- Set maximum paths should be set in EBGP mode.

```
address-family l2vpn evpn
    maximum-paths 8
vrf vrfVxLAN5001
    address-family ipv4 unicast
        maximum-paths 8
```

- If handoff nodes are deployed in dual mode, use the **route-map** command to avoid advertising prefixes associated with orphan hosts under the VPN address family.

```
ip prefix-list ROUTES_CONNECTED_NON_LOCAL seq 2 premit 15.14.0.15/32

route-map ROUTES_CONNECTED_NON_LOCAL deny
    match ip address prefix-list ROUTES_CONNECTED_NON_LOCAL

neighbor 8.8.8.8
    remote-as 100
```

```
    update-source loopback1
    address-family vpnv4 unicast
      send-community
      send-community extended
      route-map ROUTES_CONNECTED_NON_LOCAL out
```

VXLAN Configuration for the Handoff Node

Follow these guidelines when configuring VXLAN for the handoff node:

- Enable the following features:

```
feature nv overlay
feature ngmvpn
feature interface-vlan
feature vn-segment-vlan-based
```

- Configure the required L3 VNI:

```
L3VNIs are mapped to tenant VRF.
vlan 2501
  vn-segment 5001 <-- Associate VNI to a VLAN.
```

- Configure the NVE interface:

```
interface nve1
  no shutdown
  host-reachability protocol bgp
  source-interface loopback1 <-- This interface should not be the same as the MVPN
source interface.
  global suppress-arp
member vni 5001 associate-vrf <-- L3VNI
  mcast-group 233.1.1.1 <-- The underlay multicast group for VXLAN should be different
 from the MVPN default/data MDT.
```

- Configure the tenant VRF:

```
vrf context vrfVxLAN5001
  vni 5001 <-- Associate VNI to VRF.
  rd auto
address-family ipv4 unicast
    route-target both auto
    route-target both auto mvpn
    route-target both auto evpn

interface Vlan2501 <-- SVI interface associated with the L3VNI
  no shutdown
  mtu 9216 <-- The overlay header requires 58 byes, so the max tenant traffic is
(Configured MTU - 58).
  vrf member vrfVxLAN5001
  no ip redirects
  ip forward
  ipv6 forward
  no ipv6 redirects
  ip pim sparse-mode <-- PIM is enabled.

interface Vlan2 <-- SVI interface associated with L2 VNI
  no shutdown
  vrf member vrfVxLAN5001
```

```
 no ip redirects
 ip address 100.1.1.1/16
 no ipv6 redirects
 ip pim sparse-mode <-- PIM enabled on L2VNI
fabric forwarding mode anycast-gateway
```

MVPN Configuration for the Handoff Node

Follow these guidelines when configuring MVPN for the handoff node:

- Enable the following features:

```
install feature-set mpls
allow feature-set mpls
feature-set mpls
feature mpls l3vpn
feature mvpn
feature mpls ldp
```

- MPLS LDP Configuration:
  - Enable MPLS LDP (**mpls ip**) on all interfaces that are MPLS links.
  - Do not advertise loopback interfaces used for VXLAN as MPLS prefixes.
    - Configure a prefix list that contains IP addresses that identify the MVPN PE node.

```
ip prefix-list LDP-LOOPBACK seq 51 permit 9.1.1.10/32
ip prefix-list LDP-LOOPBACK seq 52 permit 9.1.2.10/32
```

    - Configure label allocation only for MVPN PE identifiers.

```
mpls ldp configuration
  explicit-null
  advertise-labels for LDP-LOOPBACK
  label allocate global prefix-list LDP-LOOPBACK
```

- Tenant VRF Configuration:
  - For the default MDT mode, make the underlay multicast group the same for all tenant multicast traffic under the VRF.

```
vrf context vrfVxLAN5001
  vni 5001
  mdt default 225.1.100.1
  mdt source loopback100 <-- If the source interface is not configured, the BGP
identifier is used as the source interface.
  mdt asm-use-shared-tree <-- If the underlay is configured in ASM mode
  no mdt enforce-bgp-mdt-safi <-- Enabled by befault but should be negated if BGP
MDT should not be used for discovery.
  mdt mtu <mtu-value> <-- Overlay ENCAP Max MTU value
```

  - For the data MDT mode, configure a unique multicast group-set for a subset of or all tenant multicast traffic.

```
mdt data 229.1.100.2/32 immediate-switch
mdt data 232.1.10.4/24 immediate-switch
```

```
route-map DATA_MDT_MAP permit 10
  match ip multicast group 237.1.1.1/32
mdt data 235.1.1.1/32 immediate-switch route-map DATA_MDT_MAP
```

- Enable MVPN tunnel statistics.

```
hardware profile mvpn-stats module all
```

CoPP Configuration for the Handoff Node

Both TRM and MVPN are heavily dependent on the control plane. Make sure to set the CoPP policy bandwidth as per the topology.

The following CoPP classes are used for TRM and MVPN traffic:

- **copp-system-p-class-multicast-router** (The default bandwidth is 3000 pps.)
- **copp-system-p-class-l3mc-data** (The default bandwidth is 3000 pps.)
- **copp-system-p-class-l2-default** (The default bandwidth is 50 pps.)
- **copp-class-normal-igmp** (The default bandwidth is 6000 pps.)

The following configuration example shows CoPP policies that can be configured to avoid control packet drops with multicast route scale.



Note The policer values in this example are approximations and might not be optimal for all topologies or traffic patterns. Configure the CoPP policies according to the MVPN/TRM traffic pattern.

```
copp copy profile strict prefix custom
  policy-map type control-plane custom-copp-policy-strict
    class custom-copp-class-normal-igmp
      police cir 6000 pps bc 512 packets conform transmit violate drop
  control-plane
    service-policy input custom-copp-policy-strict

copp copy profile strict prefix custom
  policy-map type control-plane custom-copp-policy-strict
    class custom-copp-class-multicast-router
      police cir 6000 pps bc 512 packets conform transmit violate drop
  control-plane
    service-policy input custom-copp-policy-strict

copp copy profile strict prefix custom
  policy-map type control-plane custom-copp-policy-strict
    class copp-system-p-class-l3mc-data
      police cir 3000 pps bc 512 packets conform transmit violate drop
  control-plane
    service-policy input custom-copp-policy-strict

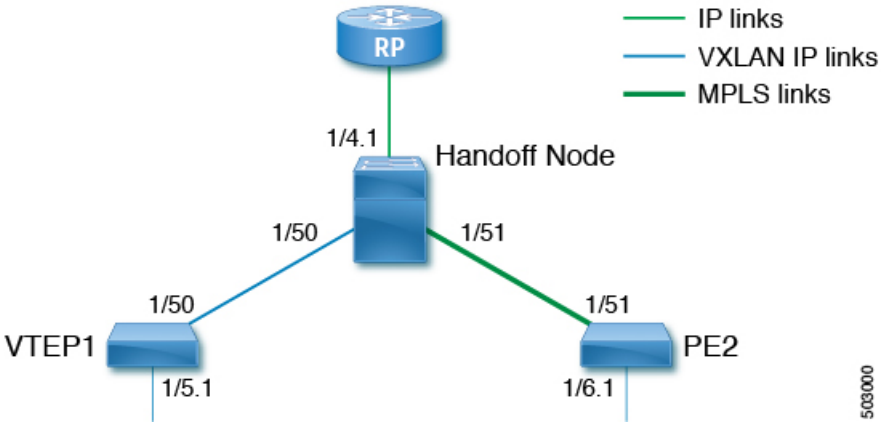
copp copy profile strict prefix custom
  policy-map type control-plane custom-copp-policy-strict
    class custom-copp-class-l2-default
      police cir 9000 pps bc 512 packets conform transmit violate drop
  control-plane
```

```
service-policy input custom-copp-policy-strict
```

Configuration Example for Seamless Integration of EVPN (TRM) with MVPN

The following figure shows a sample topology with a VXLAN network on the left, an MVPN network on the right, and a centralized handoff node.

Figure 28: Sample Topology for Seamless Integration of EVPN (TRM) with MVPN



The following example show sample configurations for the VTEP, handoff node, and PE in this topology.

Configuration on VTEP1:

```
feature ngmvpn
feature interface-vlan
feature vn-segment-vlan-based
feature nv overlay
feature pim
nv overlay evpn
ip pim rp-address 90.1.1.100 group-list 225.0.0.0/8
ip pim ssm range 232.0.0.0/8

vlan 555
  vn-segment 55500

route-map ALL_ROUTES permit 10
interface nve1
  no shutdown
  host-reachability protocol bgp
  source-interface loopback2
  member vni 55500 associate-vrf
    mcast-group 225.3.3.3

interface loopack1
  ip address 196.196.196.196/32

interface loopback2
  ip address 197.197.197.197/32
  ip pim sparse-mode
```

```
feature bgp
router bgp 1
    address-family l2vpn evpn
        maximum-paths 8
        maximum-paths ibgp 8
    neighbor 2.1.1.2
        remote-as 1
        update-source loopback 1
        address-family ipv4 unicast
          send-community extended
        address-family ipv6 unicast
          send-community extended
        address-family ipv4 mvpn
          send-community extended
        address-family l2vpn evpn
          send-community extended
    vrf vrfVxLAN5023
        address-family ipv4 unicast
          advertise l2vpn evpn
          redistribute direct route-map ALL_ROUTES
          maximum-paths 8
          maximum-paths ibgp 8

vrf context vpn1
  vni 55500
  ip pim rp-address 27.27.27.27 group-list 224.0.0.0/4
  ip pim ssm range 232.0.0.0/8
  ip multicast multipath s-g-hash next-hop-based
rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto mvpn
    route-target both auto evpn

interface Vlan555
  no shutdown
  vrf member vpn1
  ip forward
  ip pim sparse-mode

interface Ethernet 1/50
  ip pim sparse-mode

interface Ethernet1/5.1
  encapsulation dot1q 90
  vrf member vpn1
  ip address 10.11.12.13/24
  ip pim sparse-mode
  no shutdown
```

Configuration on the handoff node:

```
install feature-set mpls
  allow feature-set mpls
feature-set mpls
feature ngmvpn
feature bgp
feature pim
feature mpls l3vpn
feature mvpn
feature mpls ldp
feature interface-vlan
feature vn-segment-vlan-based
feature nv overlay
```

```
nv overlay evpn

ip pim rp-address 90.1.1.100 group-list 225.0.0.0/8
ip pim rp-address 91.1.1.100 group-list 232.0.0.0/8

interface loopback1
  ip address 90.1.1.100 /32
  ip pim sparse-mode

interface loopback2
  ip address 91.1.1.100 /32
  ip pim sparse-mode

ip prefix-list LDP-LOOPBACK seq 2 permit 20.20.20.20/32
ip prefix-list LDP-LOOPBACK seq 3 permit 30.30.30.30/32
mpls ldp configuration
    advertise-labels for LDP-LOOPBACK
    label allocate label global prefix-list LDP-LOOPBACK

interface Ethernet 1/50
    ip pim sparse-mode

interface Ethernet 1/51
    ip pim sparse-mode
    mpls ip

interface Ethernet1/4.1
  encapsulation dot1q 50
  vrf member vpn1
  ip pim sparse-mode
  no shutdown

interface loopback0
  ip address 20.20.20.20/32
  ip pim sparse-mode

vlan 555
  vn-segment 55500

route-map ALL_ROUTES permit 10

interface nve1
  no shutdown
  host-reachability protocol bgp
  source-interface loopback3
  member vni 55500 associate-vrf
    mcast-group 225.3.3.3

interface loopback3
  ip address 198.198.198.198/32
  ip pim sparse-mode

vrf context vpn1
  vni 55500
  ip pim rp-address 27.27.27.27 group-list 224.0.0.0/4
  ip pim ssm range 232.0.0.0/8
  ip multicast multipath s-g-hash next-hop-based
  mdt default 232.1.1.1
  mdt source loopback 0
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto mvpn
    route-target both auto evpn
```

```
interface Vlan555
  no shutdown
  vrf member vpn1
  ip forward
  ip pim sparse-mode

router bgp 1
    address-family l2vpn evpn
        maximum-paths 8
        maximum-paths ibgp 8
    address-family vpnv4 unicast
        no allocate-label option-b
    address-family ipv4 mdt
    address-family ipv4 mvpn
        maximum-paths 8
        maximum-paths ibgp 8
    neighbor 196.196.196.196
        remote-as 1
        address-family ipv4 unicast
          send-community extended
        address-family ipv6 unicast
          send-community extended
        address-family ipv4 mvpn
          send-community extended
        address-family l2vpn evpn
          send-community extended
          import vpn unicast reoriginate

router bgp 1
    neighbor 30.30.30.30
        remote-as 100
        update-source loopback0
        ebgp-multihop 255
        address-family ipv4 unicast
          send-community extended
        address-family vpnv4 unicast
          send-community
          send-community extended
          next-hop-self
          import l2vpn evpn reoriginate
        address-family ipv4 mdt
          send-community extended
          no next-hop-third-party
```

Configuration on PE2:

```
install feature-set mpls
  allow feature-set mpls
feature-set mpls
feature bgp
feature pim
feature mpls l3vpn
feature mpls ldp
feature interface-vlan

ip pim rp-address 91.1.1.100 group-list 232.0.0.0/8
ip prefix-list LDP-LOOPBACK seq 2 permit 20.20.20.20/32
ip prefix-list LDP-LOOPBACK seq 3 permit 30.30.30.30/32
mpls ldp configuration
    advertise-labels for LDP-LOOPBACK
    label allocate label global prefix-list LDP-LOOPBACK

interface Ethernet 1/51
```

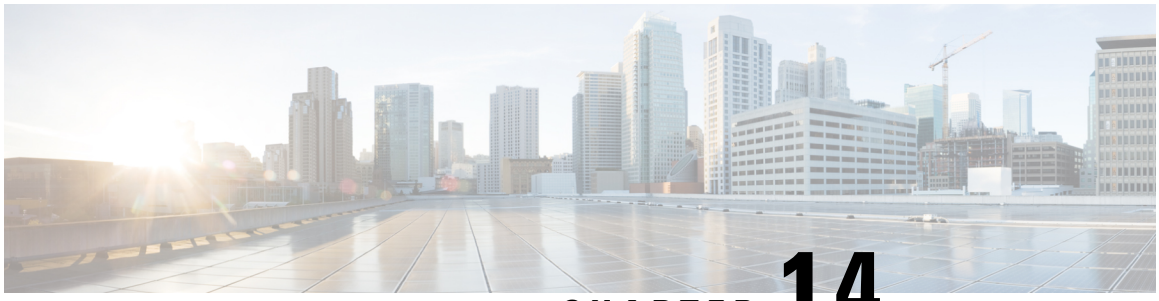
```
    ip pim sparse-mode
    mpls ip

interface Ethernet1/6.1
  encapsulation dot1q 50
  vrf member vpn1
  ip pim sparse-mode
  no shutdown

interface loopback0
  ip address 30.30.30.30/32
  ip pim sparse-mode

vrf context vpn1
  ip pim rp-address 27.27.27.27 group-list 224.0.0.0/4
  ip pim ssm range 232.0.0.0/8
  ip multicast multipath s-g-hash next-hop-based
  mdt default 232.1.1.1
  mdt source loopback 0
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto mvpn
    route-target both auto evpn

router bgp 100
      router-id 30.30.30.30
      address-family vpnv4 unicast
            additional-paths send
            additional-paths receive
            no allocate-label option-b
      neighbor 20.20.20.20
            remote-as 1
            update-source loopback0
            address-family vpnv4 unicast
                send-community
                send-community extended
            address-family ipv4 mdt
                send-community extended
                no next-hop-third-party
```



CHAPTER 14

Configuring VXLAN EVPN Multi-Site

This chapter contains these sections:

- About VXLAN EVPN Multi-Site, on page 293
- About VXLAN EVPN Multi-Site with IPv6 Underlay, on page 294
- Dual RD Support for Multi-Site, on page 295
- RP Placement in the DCI Core, on page 296
- Interoperability with EVPN Multi-Homing Using ESI for Multi-Site Anycast BGW , on page 296
- Guidelines and Limitations for VXLAN EVPN Multi-Site, on page 297
- Guidelines and Limitations for VXLAN EVPN Multi-Site with IPv6 Underlay, on page 302
- Enabling VXLAN EVPN Multi-Site, on page 302
- Enabling VXLAN EVPN Multi-Site with IPv6 Multicast Underlay, on page 304
- Configuring Dual RD Support for Multi-Site, on page 306
- Configuring VNI Dual Mode, on page 308
- Configuring Fabric/DCI Link Tracking, on page 309
- Configuring Fabric External Neighbors, on page 310
- Configuring VXLAN EVPN Multi-Site Storm Control, on page 311
- Verifying VXLAN EVPN Multi-Site Storm Control, on page 312
- Multi-Site with vPC Support, on page 312
- Configuration Example for Multi-Site with Asymmetric VNIs, on page 318
- Advertise Using PIP Towards Fabric, on page 319
- TRM with Multi-Site, on page 321
- VXLAN EVPN Multi-Site with RFC 5549 Underlay, on page 333

About VXLAN EVPN Multi-Site

The VXLAN EVPN Multi-Site solution interconnects two or more BGP-based Ethernet VPN (EVPN) sites/fabrics (overlay domains) in a scalable fashion over an IP-only network. This solution uses border gateways (BGWs) in anycast or vPC mode to terminate and interconnect two sites. The BGWs provide the network control boundary that is necessary for traffic enforcement and failure containment functionality.

In the BGP control plane for releases prior to Cisco NX-OS Release 9.3(5), BGP sessions between the BGWs rewrite the next hop information of EVPN routes and reoriginate them. Beginning with Cisco NX-OS Release 9.3(5), reorigination is always enabled (with either single or dual route distinguishers), and rewrite is not performed. For more information, see Dual RD Support for Multi-Site, on page 295.

VXLAN Tunnel Endpoints (VTEPs) are only aware of their overlay domain internal neighbors, including the BGWs. All routes external to the fabric have a next hop on the BGWs for Layer 2 and Layer 3 traffic.

The BGW is the node that interacts with nodes within a site and with nodes that are external to the site. For example, in a leaf-spine data center fabric, it can be a leaf, a spine, or a separate device acting as a gateway to interconnect the sites.

The VXLAN EVPN Multi-Site feature can be conceptualized as multiple site-local EVPN control planes and IP forwarding domains interconnected via a single common EVPN control and IP forwarding domain. Every EVPN node is identified with a unique site-scope identifier. A site-local EVPN domain consists of EVPN nodes with the same site identifier. BGWs on one hand are also part of the site-specific EVPN domain and on the other hand a part of a common EVPN domain to interconnect with BGWs from other sites. For a given site, these BGWs facilitate site-specific nodes to visualize all other sites to be reachable only via them. This means:

- Site-local bridging domains are interconnected only via BGWs with bridging domains from other sites.
- Site-local routing domains are interconnected only via BGWs with routing domains from other sites.
- Site-local flood domains are interconnected only via BGWs with flood domains from other sites.

Selective Advertisement is defined as the configuration of the per-tenant information on the BGW. Specifically, this means IP VRF or MAC VRF (EVPN instance). In cases where external connectivity (VRF-lite) and EVPN Multi-Site coexist on the same BGW, the advertisements are always enabled.

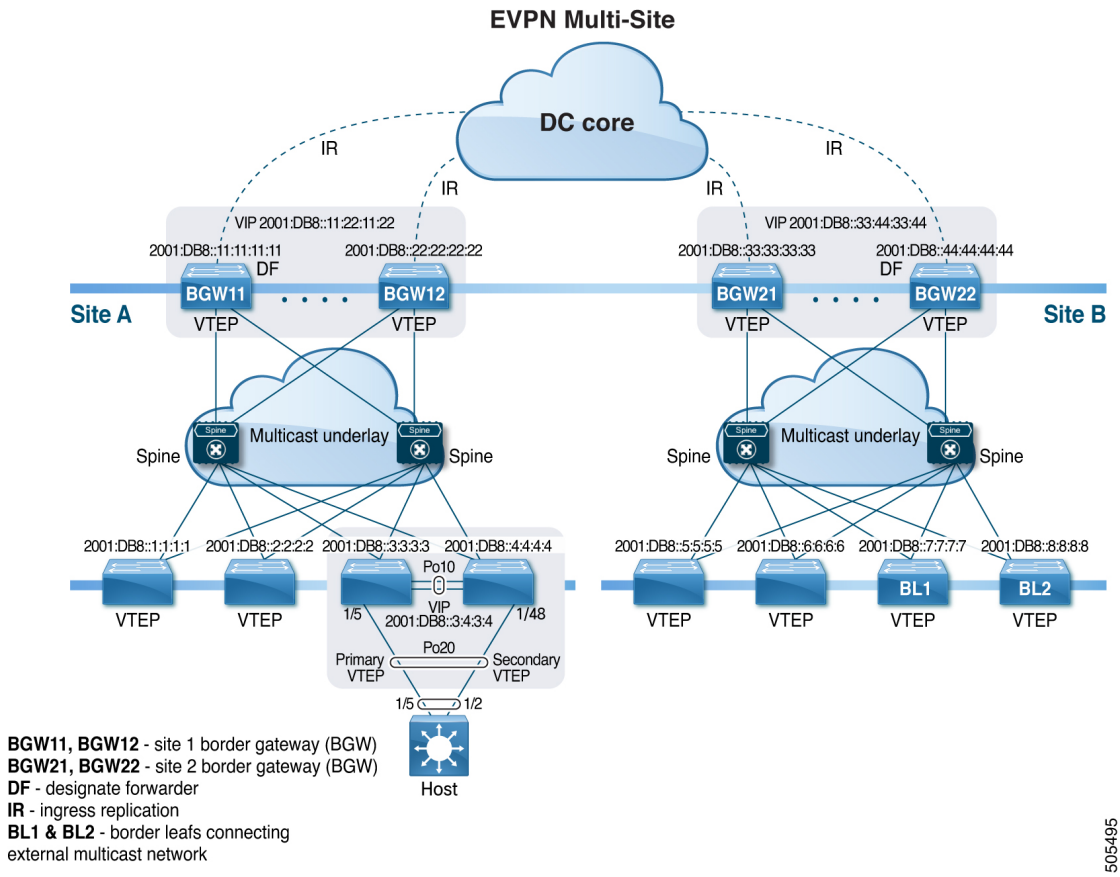


Note The MVPN VRI ID must be configured for TRM on anycast BGW if the site ID is greater than two bytes. The same VRI ID needs to be configured in all anycast BGWs that are part of the same site. However, the VRI ID must be unique within the network. That is, other anycast BGWs or vPC leaves must use different VRI IDs.

About VXLAN EVPN Multi-Site with IPv6 Underlay

Beginning with Cisco NX-OS Release 10.4(3)F, the support is provided for VXLAN EVPN Multi-Site with IPv6 Underlay.

Figure 29: Topology - VXLAN EVPN Multi-Site with IPv6 Underlay



The above topology shows four leafs and two spines in the VXLAN EVPN fabric and two Anycast BGWs. Inside the fabric, the underlay is an IPv6 Multicast running PIMv6. RP is positioned in the spine with anycast RP. BGWs support VXLAN with IPv6 Protocol-Independent Multicast (PIMv6) Any-Source Multicast (ASM) on the fabric side and Ingress Replication (IPv6) on the DCI side.

Beginning with Cisco NX-OS Release 10.5(1)F, the underlay network supports the following combinations for VXLAN EVPN Multi-Site:

- In the data center fabric, both Multicast Underlay (PIMv6) Any-Source Multicast (ASM) and Ingress Replication (IPv6) are supported.
- In the Data Center Interconnect (DCI), only Ingress Replication (IPv6) is supported.

Dual RD Support for Multi-Site

Beginning with Cisco NX-OS Release 9.3(5), VXLAN EVPN Multi-Site supports route reorigination with dual route distinguishers (RDs). This behavior is enabled automatically.

Each VRF or L2VNI tracks two RDs: a primary RD (which is unique) and a secondary RD (which is the same across BGWs). Reoriginated routes are advertised with the secondary type-0 RD (site-id:VNI). All other

routes are advertised with the primary RD. The secondary RD is allocated automatically once the router is in Multi-Site BGW mode.

If the site ID is greater than 2 bytes, the secondary RD can't be generated automatically on the Multi-Site BGW, and the following message appears:

```
%BGP-4-DUAL_RD_GENERATION_FAILED: bgp- [12564] Unable to generate dual RD on EVPN multisite
 border gateway. This may increase memory consumption on other BGP routers receiving
re-originated EVPN routes. Configure router bgp <asn> ; rd dual id <id> to avoid it.
```

In this case, you can either manually configure the secondary RD value or disable dual RDs. For more information, see Configuring Dual RD Support for Multi-Site, on page 306.

RP Placement in the DCI Core

- PIM RP placement for DCI multicast underlay
  - PIM RP and multicast groups for fabric underlay and DCI underlay must be different.
  - DCI underlay group must not overlap with fabric underlay group.
  - Multicast groups and RPs in DCI and fabric networks should be distinct and configured based on specific ranges.
  - RP can be placed on any node in the DCI core, there can be one or more RPs.
- Direct BGW to BGW Peering Deployment
  - PIM RP should be set on BGWs with anycast PIM RP for redundancy.
- BGW to Cloud Model Deployment
  - PIM RP could be placed on the DCI underlay layer, such as core routers or superspines.

Interoperability with EVPN Multi-Homing Using ESI for Multi-Site Anycast BGW

Beginning Cisco NX-OS Release 10.2(2)F, EVPN MAC/IP routes (Type 2) with non-reserved as well as with reserved ESI (0 or MAX-ESI) values are evaluated for forwarding (ESI RX). The definition of the EVPN MAC/IP route resolution is defined in RFC 7432 Section 9.2.2.

EVPN MAC/IP routes (Type 2) -

- with reserved ESI value (0 or MAX-ESI) are resolved solely by the MAC/IP route alone (BGP next-hop within Type 2).
- with non-reserved ESI value are resolved only if an accompanied per-ES Ethernet Auto-Discovery route (Type 1, per-ES EAD) is present.

In addition to the MAC/IP route resolution as mentioned above, the Multi-Site BGW supports the forward, rewrite and re-originate of MAC/IP routes with reserved and non-reserved ESI values. In all these cases, the per-ES EAD route is re-originated by the Multi-Site BGW.

The EVPN MAC/IP route resolution with the different ESI values is supported on Cisco Nexus 9300-FX, -FX2, -FX3, and -GX Platform Switches in Anycast and vPC Border Gateway mode.

Guidelines and Limitations for VXLAN EVPN Multi-Site

VXLAN EVPN Multi-Site has the following configuration guidelines and limitations:

- The following switches support VXLAN EVPN Multi-Site:
  - Cisco Nexus 9300-FX platform switches
  - Cisco Nexus 9300-FX2 platform switches
  - Cisco Nexus 9300-FX3 platform switches
  - Cisco Nexus 9300-GX platform switches
  - Cisco Nexus 9300-GX2 platform switches
  - Cisco Nexus 9332D-H2R switches
  - Cisco Nexus 93400LD-H1 switches
  - Cisco Nexus 9364C-H1 switches
  - Cisco Nexus 9800 platform switches with X9836DM-A and X98900CD-A line cards
  - Cisco Nexus 9500 platform switches with -FX or -GX or -FX3 line cards



Note Cisco Nexus 9500 platform switches with -R/RX line cards don't support VXLAN EVPN Multi-Site.

- Beginning with Cisco NX-OS Release 10.2(3)F, the VXLAN EVPN Multi-Site is supported on the Cisco Nexus 9300-GX2 platform switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, the VXLAN EVPN Multi-Site is supported on the Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, the VXLAN EVPN Multi-Site is supported on the Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, the VXLAN EVPN Multi-Site is supported on the Cisco Nexus 9364C-H1 switches.
- Beginning with Cisco NX-OS Release 10.5(2)F, the following features are supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.
  - Multi-Hop BFD
  - VXLAN and iVXLAN stripping
  - DCI advertise PIP (without cloudsec) on vPC and Anycast BGW

Switch or Port restrictions

- The **evpn multisite dci-tracking** is mandatory for anycast BGWs and vPC BGW DCI links.
- EVPN multisite DCI-tracking and EVPN multisite fabric-tracking are only supported on physical interfaces. Use on SVIs is not supported.

Deployment restrictions

- In a VXLAN EVPN Multi-Site deployment, when you use the ttag feature, make sure that the ttag is stripped (**ttag-strip**) on BGW's DCI interfaces attached to Non-NXOS gear.
- VXLAN EVPN Multi-Site and Tenant Routed Multicast (TRM) are supported between sources and receivers deployed across different sites.
- The Multi-Site BGW allows the coexistence of Multi-Site extensions (Layer 2 unicast/multicast and Layer 3 unicast) as well as Layer 3 unicast and multicast external connectivity.
- In TRM with multi-site deployments, all BGWs receive traffic from fabric. However, only the designated forwarder (DF) BGW forwards the traffic. All other BGWs drop the traffic through a default drop ACL. This ACL is programmed in all DCI tracking ports. Don't remove the **evpn multisite dci-tracking** configuration from the DCI uplink ports. If you do, you remove the ACL, which creates a nondeterministic traffic flow in which packets can be dropped or duplicated instead of deterministically forwarded by only one BGW, the DF.
- Prior to NX-OS 10.2(2)F only ingress replication was supported between DCI peers across the core. Beginning with Cisco NX-OS Release 10.2(2)F both ingress replication and multicast are supported between DCI peers across the core.
- The DCI underlay group and the fabric underlay group must be distinct, ensuring no overlap between DCI multicast and fabric multicast underlay groups.
- Bind NVE to a loopback address that is separate from loopback addresses that are required by Layer 3 protocols. A best practice is to use a dedicated loopback address for the NVE source interface (PIP VTEP) and multi-site source interface (anycast and virtual IP VTEP).
- Beginning with Cisco NX-OS Release 9.3(5), if you disable the **host-reachability protocol bgp** command under the NVE interface in a VXLAN EVPN Multi-Site topology, the NVE interface stays operationally down.
- Beginning with Cisco NX-OS Release 9.3(5), Multi-Site Border Gateways re-originate incoming remote routes when advertising to the site's local spine/leaf switches. These re-originated routes modify the following fields:
  - RD value changes to [Multisite Site ID:L3 VNID].
  - It is mandatory that Route-Targets are defined on all VTEP that are participating in a given VRF, this includes and is explicitly required for the BGW to extend the given VRF. Prior to Cisco NX-OS Release 9.3(5), Route-Targets from intra-site VTEPs were inadvertently kept across the site boundary, even if not defined on the BGW. Starting from Cisco NX-OS Release 9.3(5) the mandatory behavior is enforced. By adding the necessary Route-Targets to the BGW, the change from inadvertent Route-Target advertisement to explicit Route-Target advertisement can be performed.
  - Path type changes from external to local.

    - For SVI-related triggers (such as shut/unshut or PIM enable/disable), a 30-second delay was added, allowing the Multicast FIB (MFIB) Distribution module (MFDM) to clear the hardware table before toggling between L2 and L3 modes or vice versa.

- Ensure that the **ip pim sparse-mode** is enabled on the Multi-Site VIP loopback interface.
- To improve the convergence in case of fabric link failure and avoid issues in case of fabric link flapping, ensure to configure multi-hop BFD between loopbacks of spines and BGWs.

  In the specific scenario where a BGW node becomes completely isolated from the fabric due to all its fabric links failing, the use of multi-hop BFD ensures that the BGP sessions between the spines and the isolated BGW can be immediately brought down, without relying on the configured BGP hold-time value.
- In a VXLAN Multi-Site environment, a border gateway device that uses ECMP for routing through both a VXLAN overlay and an L3 prefix to access remote site subnets might encounter adjacency resolution failure for one of these routes. If the switch attempts to use this unresolved prefix, it will result in traffic being dropped.
- To improve convergence during the reload of anycast BGW routers in multi-plane topology, ensure to configure the **multi-hop BFD** and **nexthop trigger-delay** commands.
- Following guidelines and limitations are applied when a multisite Border Gateway is put into Maintenance Mode:
    - BUM Traffic from remote Fabrics will still be attracted to the Border gateway that is in maintenance mode
    - Border Gateway in maintenance mode still participates in Designated Forwarder Election
    - Default Maintenance mode profile applies the command "ip pim isolate" and so the Border gateway is isolated from S,G tree towards the fabric direction. This leads to BUM traffic loss and hence an appropriate maintenance mode profile should be used for Border Gateways than the default.

vPC BGW restrictions

- BGWs in a vPC topology are supported.
- vPC mode can support only two BGWs.
- vPC mode can support both Layer 2 hosts and Layer 3 services on local interfaces.
- In vPC mode, BUM is replicated to either of the BGWs for traffic coming from the external site. Hence, both BGWs are forwarders for site external to site internal (DCI to fabric) direction.
- In vPC mode, BUM is replicated to either of the BGWs for traffic coming from the local site leaf for a VLAN using Ingress Replication (IR) underlay. Both BGWs are forwarders for site internal to site external (fabric to DCI) direction for VLANs using the IR underlay.
- In vPC mode, BUM is replicated to both BGWs for traffic coming from the local site leaf for a VLAN using the multicast underlay. Therefore, a decapper/forwarder election happens, and the decapsulation winner/forwarder only forwards the site-local traffic to external site BGWs for VLANs using the multicast underlay.
- In vPC mode, all Layer 3 services/attachments are advertised in BGP via EVPN Type-5 routes with their virtual IP as next hop. If the VIP/PIP feature is configured, they are advertised with PIP as the next hop.

Unsupported features

- Multicast Flood Domain between inter-site/fabric BGWs isn't supported.
- iBGP EVPN Peering between BGWs of different fabrics/sites isn't supported.
- PIM BiDir is not supported for fabric underlay multicast replication with VXLAN Multi-Site.
- FEX is not supported on a vPC BGW and Anycast BGW.

Anycast BGW restrictions

- Anycast mode can support up to six BGWs per site.
- Anycast mode can support only Layer 3 services that are attached to local interfaces.
- In Anycast mode, BUM is replicated to each border leaf. DF election between the border leafs for a particular site determines which border leaf forwards the inter-site traffic (fabric to DCI and conversely) for that site.
- In Anycast mode, all Layer 3 services are advertised in BGP via EVPN Type-5 routes with their physical IP as the next hop.
- If different Anycast Gateway MAC addresses are configured across sites, enable ARP suppression and ND suppression for all VLANs that have been extended.

Supported features

- Beginning with Cisco NX-OS Release 9.3(5), VTEPs support VXLAN-encapsulated traffic over parent interfaces if subinterfaces are configured. This feature is supported for VXLAN EVPN Multi-Site and DCI. DCI tracking can be enabled only on the parent interface.
- Beginning with Cisco NX-OS Release 9.3(5), VXLAN EVPN Multi-Site supports asymmetric VNIs. For more information, see Multi-Site with Asymmetric VNIs and Configuration Example for Multi-Site with Asymmetric VNIs, on page 318.
- **Dual RD**

  The following guidelines and limitations apply to dual RD support for Multi-Site:

  - Dual RD are supported beginning with Cisco NX-OS Release 9.3(5).
  - Dual RD is enabled automatically for Cisco Nexus 9300-FX/FX2 platform switches and Cisco Nexus 9500 platform switches with -FX/FX3 line cards that have VXLAN EVPN Multi-Site enabled.
  - Beginning with Cisco NX-OS Release 10.2(3)F, the dual RD support for Multi-Site is supported on the Cisco Nexus 9300-FX3 platform switches.
  - To use CloudSec or other features that require PIP advertisement for multi-site reoriginated routes, configure BGP additional paths on the route server if dual RD are enabled on the BGW, or disable dual RD.
  - Sending secondary RD additional paths at the BGW node isn't supported.
  - During an ISSU, the number of paths for the leaf nodes might double temporarily while all BGWs are being upgraded.

Guidelines and Limitations for VXLAN Multi-Site Anycast BGW Support on Cisco Nexus 9800 Series Switches

- Beginning with Cisco NX-OS Release 10.4(3)F, the VXLAN Multi-Site Anycast BGW is supported on the Cisco Nexus 9808/9804 switches with X9836DM-A and X98900CD-A line cards.
  - VXLAN Multi-Site Anycast BGW supports the following features:
    - VXLAN BGP EVPN fabric and multi-site interconnect
    - VXLAN Layer-2 VNI and new Layer-3 VNI which is not VLAN based
    - IPv4 underlay
    - IPv6 underlay
    - Ingress Replication on fabric and DCI side
    - Multicast underlay in Fabric
    - Bud node
    - TRMv4
    - TRMv6
    - NGOAM
    - VXLAN Counters
      - Per VXLAN peer based total packet/byte counters are supported.
      - Per VNI based total packet/byte counters are supported
  - VXLAN Multi-Site Anycast BGW does not support the following features:
    - Downstream VNI and route leak
    - L3 Port channel as a fabric or DCI link
    - Multicast underlay on DCI side
    - VXLAN access features
    - IGMP snooping
    - Separate VXLAN counters for broadcast, multicast, and unicast traffic
    - Data MDT
    - EVPN storm control

Guidelines and limitations for VXLAN Multi-Site Anycast BGW support on Cisco Nexus 9336C-SE1 Switches

- Beginning with Cisco NX-OS Release 10.6(1)F, the VXLAN Multi-Site Anycast BGW is supported on the Cisco Nexus 9336C-SE1 switches. For information about supported and unsupported features, see the Guidelines and limitations for VXLAN support on Cisco Nexus 9336C-SE1 switches, on page 55 section.

Guidelines and Limitations for VXLAN EVPN Multi-Site with IPv6 Underlay

VXLAN EVPN Multi-Site with IPv6 Underlay has the following configuration guidelines and limitations:

- Cisco Nexus 9300-FX, FX2, FX3, GX, GX2, H2R and H1 ToR switches are supported as the leaf VTEP or BGW.
- Cisco Nexus N9K-X9716D-GX and N9K-X9736C-FX, N9K-X9736C-FX3 line cards are supported only on the spine (EoR).
- When an EoR is deployed as a spine node with Multicast Underlay (PIMv6) Any-Source Multicast (ASM), it is mandatory to configure non-default template using one of the following commands in global configuration mode:
  - **system routing template-multicast-heavy**
  - **system routing template-multicast-ext-heavy**
- vPC BGWs are not supported with IPv6 multicast underlay.
- Dual stack configuration is not supported for NVE source interface loopback and multi-site interface loopback.
- Beginning with Cisco NX-OS Release 10.5(1)F, VXLAN EVPN Multi-Site in the data center fabric supports both Multicast Underlay (PIMv6) Any-Source Multicast (ASM) and Ingress Replication (IPv6) in the underlay. This support is available on the following switches and line cards:
  - Cisco Nexus 9300-FX, FX2, FX3, GX, GX2, H2R, and H1 ToR switches as the leaf VTEPs.
  - Cisco Nexus N9K-X9716D-GX and N9K-X9736C-FX line cards as spines if the underlay is configured for Multicast Underlay (PIMv6) Any-Source Multicast (ASM).
  - Cisco Nexus N9K-X9716D-GX and N9K-X9736C-FX line cards as VTEPs if the underlay uses Ingress Replication (IPv6).
- Beginning with Cisco NX-OS Release 10.5(2)F, VXLAN EVPN Multi-Site with IPv6 Underlay support is extended on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line cards as VTEPs if the underlay uses Ingress Replication (IPv6).

Enabling VXLAN EVPN Multi-Site

This procedure enables the VXLAN EVPN Multi-Site feature. Multi-Site is enabled on the BGWs only. The site-id must be the same on all BGWs in the fabric/site.

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**evpn multisite border-gateway** *ms-id* **Example:** `switch(config)# evpn multisite border-gateway 100`	Configures the site ID for a site/fabric. The range of values for *ms-id* is 1 to 2,814,749,767,110,655. The *ms-id* must be the same in all BGWs within the same fabric/site.
Step 3	**split-horizon per-site** **Example:** `switch(config-evpn-msite-bgw)# split-horizon per-site`	Enables to receive packets encapsulated with DCI group from another border gateway on the same site and avoids duplication of packets. **Note** Use this command when DCI multicast underlay is configured on a site with anycast border gateway.
Step 4	**interface nve 1** **Example:** `switch(config-evpn-msite-bgw)# interface nve 1`	Creates a VXLAN overlay interface that terminates VXLAN tunnels. **Note** Only one NVE interface is allowed on the switch.
Step 5	**source-interface loopback** *src-if* **Example:** `switch(config-if-nve)# source-interface loopback 0`	The source interface must be a loopback interface that is configured on the switch with a valid /32 IP address. This /32 IP address must be known by the transient devices in the transport network and the remote VTEPs. This requirement is accomplished by advertising it through a dynamic routing protocol in the transport network.
Step 6	**host-reachability protocol bgp** **Example:** `switch(config-if-nve)# host-reachability protocol bgp`	Defines BGP as the mechanism for host reachability advertisement.
Step 7	**multisite border-gateway interface loopback** *vi-num* **Example:** `switch(config-if-nve)# multisite border-gateway interface loopback 100`	Defines the loopback interface used for the BGW virtual IP address (VIP). The border-gateway interface must be a loopback interface that is configured on the switch with a valid /32 IP address. This /32 IP address must be known by the transient devices in the transport network and the remote VTEPs. This requirement is accomplished by advertising it through a dynamic routing protocol in the transport network. This loopback must be different than the source interface loopback. The range of *vi-num* is from 0 to 1023.

	Command or Action	Purpose
Step 8	**no shutdown** **Example:** `switch(config-if-nve)# no shutdown`	Negates the **shutdown** command.
Step 9	**exit** **Example:** `switch(config-if-nve)# exit`	Exits the NVE configuration mode.
Step 10	**interface loopback** *loopback-number* **Example:** `switch(config)# interface loopback 0`	Configures the loopback interface.
Step 11	**ip address** *ip-address* **Example:** `switch(config-if)# ip address 198.0.2.0/32`	Configures the IP address for the loopback interface.

Enabling VXLAN EVPN Multi-Site with IPv6 Multicast Underlay

This procedure enables the VXLAN EVPN Multi-Site feature with IPv6 multicast underlay. Multi-Site is enabled on the BGWs only. The site-id must be the same on all BGWs in the fabric/site.

SUMMARY STEPS

1. **configure terminal**
2. **evpn multisite border-gateway** *ms-id*
3. **interface nve 1**
4. **source-interface loopback** *src-if*
5. **host-reachability protocol bgp**
6. **multisite border-gateway interface loopback** *vi-num*
7. (Optional) **multisite virtual-rmac** *mac-address*
8. **member vni** *vni-range*
9. **multisite ingress-replication**
10. **mcast-group** *ipv6-address*
11. **no shutdown**
12. **exit**
13. **interface loopback** *loopback-number*
14. **ipv6 address** *ipv6-address*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**evpn multisite border-gateway** *ms-id* **Example:** `switch(config)# evpn multisite border-gateway 100`	Configures the site ID for a site/fabric. The range of values for *ms-id* is 1 to 2,814,749,767,110,655. The *ms-id* must be the same in all BGWs within the same fabric/site. **Note** The **mvpn vri id** *id* command is required on BGWs if site-id value is greater than 2 bytes, and this value has to be same across all same site BGWs and unique in TRM domain. Also this value must not collide with any site-id value.
Step 3	**interface nve 1** **Example:** `switch(config-evpn-msite-bgw)# interface nve 1`	Creates a VXLAN overlay interface that terminates VXLAN tunnels. **Note** Only one NVE interface is allowed on the switch.
Step 4	**source-interface loopback** *src-if* **Example:** `switch(config-if-nve)# source-interface loopback 0`	The source interface must be a loopback interface that is configured on the switch with a valid /128 IPv6 address. This /128 IPv6 address must be known by the transient devices in the transport network and the remote VTEPs. This requirement is accomplished by advertising it through a dynamic routing protocol in the transport network.
Step 5	**host-reachability protocol bgp** **Example:** `switch(config-if-nve)# host-reachability protocol bgp`	Defines BGP as the mechanism for host reachability advertisement.
Step 6	**multisite border-gateway interface loopback** *vi-num* **Example:** `switch(config-if-nve)# multisite border-gateway interface loopback 100`	Defines the loopback interface used for the BGW virtual IPv6 address (VIP). The border-gateway interface must be a loopback interface that is configured on the switch with a valid /128 IPv6 address. This /128 IPv6 address must be known by the transient devices in the transport network and the remote VTEPs. This requirement is accomplished by advertising it through a dynamic routing protocol in the transport network. This loopback must be different than the source interface loopback. The range of *vi-num* is from 0 to 1023.

	Command or Action	Purpose
Step 7	(Optional) **multisite virtual-rmac** *mac-address* **Example:** `switch(config-if-nve)# multisite virtual-rmac 0600.0000.abcd`	For interoperability with other switches, user have to manually configure vMAC on Nexus 9000 switches to override the auto generated vMAC. The default behavior is to auto generate. If manual VMAC is configured, manual vMAC will take precedence. **Note** Only unicast MAC address range is supported for vMAC address configuration.
Step 8	**member vni** *vni-range* **Example:** `switch(config-if-nve)# member vni 50101`	Configures the virtual network identifier (VNI). The range for vni-range is from 1 to 16,777,214. The value of vni-range can be a single value like 5000 or a range like 5001-5008.
Step 9	**multisite ingress-replication** **Example:** `switch(config-if-nve-vni)# multisite ingress-replication`	Defines the Multi-Site replication method for extending TRM functionality across sites.
Step 10	**mcast-group** *ipv6-address* **Example:** `switch(config-if-nve-vni)# mcast-group ff03::101`	Configures the IPv6 Multicast group within the fabric
Step 11	**no shutdown** **Example:** `switch(config-if-nve)# no shutdown`	Negates the **shutdown** command.
Step 12	**exit** **Example:** `switch(config-if-nve)# exit`	Exits the NVE configuration mode.
Step 13	**interface loopback** *loopback-number* **Example:** `switch(config)# interface loopback 0`	Configures the loopback interface.
Step 14	**ipv6 address** *ipv6-address* **Example:** `switch(config-if)# ipv6 address 2001:DB8::11:11:11:11/128`	Configures the IPv6 address for the loopback interface.

Configuring Dual RD Support for Multi-Site

Follow these steps if you need to manually configure the secondary RD value or disable dual RDs.

Before you begin

Enable VXLAN EVPN Multi-Site.

Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# configure terminal`
	`switch(config)#`
		Enters global configuration mode.

Step 2	**router bgp** *as-num*
	Example:
	`switch(config)# router bgp 100`
	`switch(config-router)#`
		Configures the autonomous system number. The range for *as-num* is from 1 to 4,294,967,295.

Step 3	[**no**] **rd dual id** [*2-bytes*] **Example:** `switch(config-router)# rd dual id 1`	Defines the first 2 bytes of the secondary RD. The ID must be the same across the Multi-Site BGWs. The range is from 1 to 65535. **Note** If necessary, you can use the **no rd dual** command to disable dual RDs and fall back to a single RD.
Step 4	(Optional) **show bgp evi** *evi-id* **Example:** `switch(config-router)# show bgp evi 100`	Displays the secondary RD configured as part of the **rd dual id** [*2-bytes*] command for the specified EVI.

Example

The following example shows sample output for the **show bgp evi** *evi-id* command:

```
switch# show bgp evi 100
------------------------------------------------
  L2VNI ID                    : 100 (L2-100)
  RD                          : 3.3.3.3:32867
  Secondary RD                : 1:100
  Prefixes (local/total)      : 1/6
  Created                     : Jun 23 22:35:13.368170
  Last Oper Up/Down           : Jun 23 22:35:13.369005 / never
  Enabled                     : Yes

  Active Export RT list       :
        100:100
  Active Import RT list       :
        100:100
```

Configuring VNI Dual Mode

This procedure describes the configuration of the BUM traffic domain for a given VLAN. Support exists for using multicast or ingress replication inside the fabric/site and ingress replication across different fabrics/sites.

Note If you have multiple VRFs and only one is extended to ALL leaf switches, you can add a dummy loopback to that one extended VRF and advertise through BGP. Otherwise, you'll need to check how many VRFs are extended and to which switches, and then add a dummy loopback to the respective VRFs and advertise them as well. Therefore, use the **advertise-pip** command to prevent potential user errors in the future.

For more information about configuring multicast or ingress replication for a large number of VNIs, see Example of VXLAN BGP EVPN (eBGP), on page 162.

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**interface nve 1** **Example:** `switch(config)# interface nve 1`	Creates a VXLAN overlay interface that terminates VXLAN tunnels. **Note** Only one NVE interface is allowed on the switch.
Step 3	**member vni** *vni-range* **Example:** `switch(config-if-nve)# member vni 200`	Configures the virtual network identifier (VNI). The range for *vni-range* is from 1 to 16,777,214. The value of *vni-range* can be a single value like 5000 or a range like 5001-5008. **Note** Enter one of the Step 4 or Step 5 commands.
Step 4	**mcast-group** *ip-addr* **Example:** `switch(config-if-nve-vni)# mcast-group 255.0.4.1`	Configures the NVE Multicast group IP prefix within the fabric.
Step 5	**ingress-replication protocol bgp** **Example:** `switch(config-if-nve-vni)# ingress-replication protocol bgp`	Enables BGP EVPN with ingress replication for the VNI within the fabric.
Step 6	**multisite ingress-replication** **Example:**	Defines the Multi-Site BUM replication method for extending the Layer 2 VNI.

	Command or Action	Purpose
	`switch(config-if-nve-vni)# multisite ingress-replication`	

Configuring Fabric/DCI Link Tracking

This procedure describes the configuration to track all DCI-facing interfaces and site internal/fabric facing interfaces. Tracking is mandatory and is used to disable reorigination of EVPN routes either from or to a site if all the DCI/fabric links go down.

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**interface ethernet** *port* **Example:** `switch(config)# interface ethernet1/1`	Enters interface configuration mode for the DCI or fabric interface. **Note** Enter one of the following commands in Step 3 or Step 4.
Step 3	**evpn multisite dci-tracking** **Example:** `switch(config-if)# evpn multisite dci-tracking`	Configures DCI interface tracking.
Step 4	(Optional) **evpn multisite fabric-tracking** **Example:** `switch(config-if)# evpn multisite fabric-tracking`	Configures EVPN Multi-Site fabric tracking. The **evpn multisite fabric-tracking** is mandatory for anycast BGWs and vPC BGW fabric links.

Step 5	**ip address** *ip-addr* | **ipv6 address** *ipv6-addr*
	Example:
	For IPv4
	`switch(config-if)# ip address 192.1.1.1`
	Example:
	For IPv6
	`switch(config-if)# ipv6 address 2001:DB8::192:1:1:1`
		Configures the IP or IPv6 address.

Step 6	**no shutdown** **Example:** `switch(config-if)# no shutdown`	Negates the **shutdown** command.

Configuring Fabric External Neighbors

This procedure describes the configuration of fabric external/DCI neighbors for communication to other site/fabric BGWs.

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**router bgp** *as-num* **Example:** `switch(config)# router bgp 100`	Configures the autonomous system number. The range for *as-num* is from 1 to 4,294,967,295.

Step 3	**neighbor** [*ip-addr* | *ipv6-addr*]
	Example:
	For IPv4
	`switch(config-router)# neighbor 100.0.0.1`
	Example:
	For IPv6
	`switch(config-router)# neighbor 2001:DB8::100:0:0:1`
		Configures a BGP neighbor.

Step 4	**remote-as** *value* **Example:** `switch(config-router-neighbor)# remote-as 69000`	Configures remote peer's autonomous system number.
Step 5	**peer-type fabric-external** **Example:** `switch(config-router-neighbor)# peer-type fabric-external`	Enables the next hop rewrite for Multi-Site. Defines site external BGP neighbors for EVPN exchange. The default for **peer-type** is **fabric-internal**. **Note** The **peer-type fabric-external** command is required only for VXLAN Multi-Site BGWs. It is not required for pseudo BGWs.
Step 6	**address-family l2vpn evpn** **Example:** `switch(config-router-neighbor)# address-family l2vpn evpn`	Configures the address family Layer 2 VPN EVPN under the BGP neighbor.
Step 7	**rewrite-evpn-rt-asn** **Example:**	Rewrites the route target (RT) information to simplify the MAC-VRF and IP-VRF configuration. BGP receives a route, and as it processes the RT attributes, it checks if the

	Command or Action	Purpose
	`switch(config-router-neighbor)# rewrite-evpn-rt-asn`	AS value matches the peer AS that is sending that route and replaces it. Specifically, this command changes the incoming route target's AS number to match the BGP-configured neighbor's remote AS number. You can see the modified RT value in the receiver router.

Configuring VXLAN EVPN Multi-Site Storm Control

VXLAN EVPN Multi-Site Storm Control allows rate limiting of multidestination (BUM) traffic on Multi-Site BGWs. You can control BUM traffic sent over the DCI link using a policer on fabric links in the ingress direction.

Remote peer reachability must be only through DCI links. Appropriate routing configuration must ensure that remote site routes are not advertised over Fabric links.

Multicast traffic is policed only on DCI interfaces, while unknown unicast and broadcast traffic is policed on both DCI and fabric interfaces.

Cisco NX-OS Release 9.3(6) and later releases optimize rate granularity and accuracy. Bandwidth is calculated based on the accumulated DCI uplink bandwidth, and only interfaces tagged with DCI tracking are considered. (Prior releases also include fabric-tagged interfaces.) In addition, granularity is enhanced by supporting two digits after the decimal point. These enhancements apply to the Cisco Nexus 9300-FX/FX2/FX3, and 9300-GX platform switches.

Beginning with Cisco NX-OS Release 10.5(2)F, VXLAN EVPN Multi-Site Storm Control is supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.

Note For information on access port storm control, see the Cisco Nexus 9000 Series NX-OS Layer 2 Configuration Guide.

SUMMARY STEPS

1. **configure terminal**
2. [**no**] **evpn storm-control** {**broadcast** | **multicast** | **unicast**} {**level** *level*}

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# configure terminal`
	`switch(config)#`
		Enters global configuration mode.

	Command or Action	Purpose

Step 2	[**no**] **evpn storm-control** {**broadcast** | **multicast** | **unicast**} {**level** *level*}
	Example:
	`switch(config)# evpn storm-control unicast level 10`
	Example:
	`switch(config)# evpn storm-control unicast level 10.20`
		Configures the storm suppression level as a number from 0–100.
		0 means that all traffic is dropped, and 100 means that all traffic is allowed. For any value in between, the unknown unicast traffic rate is restricted to a percentage of available bandwidth. For example, a value of 10 means that the traffic rate is restricted to 10% of the available bandwidth, and anything above that rate is dropped.
		Beginning with Cisco NX-OS Release 9.3(6), you can configure the level as a fractional value by adding two digits after the decimal point. For example, you can enter a value of 10.20.

Verifying VXLAN EVPN Multi-Site Storm Control

To display EVPN storm control setting information, enter the following command:

Command	Purpose
slot 1 show hardware vxlan storm-control	Displays the status of EVPN storm control setting.

Note Once the Storm control hits the threshold, a message is logged as stated below:

```
BGWY-1 %ETHPORT-5-STORM_CONTROL_ABOVE_THRESHOLD: Traffic in port Ethernet1/32 exceeds the
configured threshold , action - Trap (message repeated 38 times)
```

Multi-Site with vPC Support

About Multi-Site with vPC Support

The BGWs can be in a vPC complex. In this case, it is possible to support dually-attached directly-connected hosts that might be bridged or routed as well as dually-attached firewalls or service attachments. The vPC BGWs have vPC-specific multihoming techniques and do not rely on EVPN Type 4 routes for DF election or split horizon.

Guidelines and Limitations for Multi-Site with vPC Support

Multi-Site with vPC support has the following configuration guidelines and limitations:

- 4000 VNIs for vPC are not supported.
- For BUM with continued VIP use, the MCT link is used as transport upon core isolation or fabric isolation, and for unicast traffic in fabric isolation.

- Beginning with Cisco NX-OS Release 10.1(2), TRM Multisite with vPC BGW is supported.
- The routes to remote Multisite BGW loopback addresses must always prioritize the DCI link path over the iBGP protocol between vPC Border Gateway switches configured using the backup SVI. The backup SVI should be used strictly in the event of a DCI link failure.
- vPC BGWs are not supported with IPv6 multicast underlay.

Configuring Multi-Site with vPC Support

This procedure describes the configuration of Multi-Site with vPC support:

- Configure vPC domain.
- Configure port channels.
- Configuring vPC Peer Link.

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**feature vpc** **Example:** `switch(config)# feature vpc`	Enables vPCs on the device.
Step 3	**feature interface-vlan** **Example:** `switch(config)# feature interface-vlan`	Enables the interface VLAN feature on the device.
Step 4	**feature lacp** **Example:** `switch(config)# feature lacp`	Enables the LACP feature on the device.
Step 5	**feature pim** **Example:** `switch(config)# feature pim`	Enables the PIM feature on the device.
Step 6	**feature ospf** **Example:** `switch(config)# feature ospf`	Enables the OSPF feature on the device.
Step 7	**ip pim rp-address** *address* **group-list** *range* **Example:**	Defines a PIM RP address for the underlay multicast group range.

	Command or Action	Purpose
	`switch(config)# ip pim rp-address 100.100.100.1 group-list 224.0.0/4`	
Step 8	**vpc domain** *domain-id* **Example:** `switch(config)# vpc domain 1`	Creates a vPC domain on the device and enters vpn-domain configuration mode for configuration purposes. There is no default. The range is from 1 to 1000.
Step 9	**peer switch** **Example:** `switch(config-vpc-domain)# peer switch`	Defines the peer switch.
Step 10	**peer gateway** **Example:** `switch(config-vpc-domain)# peer gateway`	Enables Layer 3 forwarding for packets destined to the gateway MAC address of the vPC.
Step 11	**peer-keepalive destination** *ip-address* **Example:** `switch(config-vpc-domain)# peer-keepalive destination 172.28.230.85`	Configures the IPv4 address for the remote end of the vPC peer-keepalive link. **Note** The system does not form the vPC peer link until you configure a vPC peer-keepalive link. The management ports and VRF are the defaults.
Step 12	**ip arp synchronize** **Example:** `switch(config-vpc-domain)# ip arp synchronize`	Enables IP ARP synchronize under the vPC domain to facilitate faster ARP table population following device reload.
Step 13	**ipv6 nd synchronize** **Example:** `switch(config-vpc-domain)# ipv6 nd synchronize`	Enables IPv6 ND synchronization under the vPC domain to facilitate faster ND table population following device reload.

Step 14	Create the vPC peer-link.
	Example:
	`switch(config)# interface port-channel 1`
	`switch(config)# switchport`
	`switch(config)# switchport mode trunk`
	`switch(config)# switchport trunk allowed vlan 1,10,100-200`
	`switch(config)# mtu 9216`
	`switch(config)# vpc peer-link`
	`switch(config)# no shut`
	
	`switch(config)# interface Ethernet 1/1, 1/21`
	`switch(config)# switchport`
	`switch(config)# mtu 9216`
	`switch(config)# channel-group 1 mode active`
	`switch(config)# no shutdown`
		Creates the vPC peer-link port-channel interface and adds two member interfaces to it.

	Command or Action	Purpose
Step 15	**system nve infra-vlans** *range* **Example:** `switch(config)# system nve infra-vlans 10`	Defines a non-VXLAN-enabled VLAN as a backup routed path.
Step 16	**vlan** *number* **Example:** `switch(config)# vlan 10`	Creates the VLAN to be used as an infra-VLAN.

Step 17	Create the SVI.
	Example:
	`switch(config)# interface vlan 10`
	`switch(config)# ip address 10.10.10.1/30`
	`switch(config)# ip router ospf process UNDERLAY area 0`
	`switch(config)# ip pim sparse-mode`
	`switch(config)# no ip redirects`
	`switch(config)# mtu 9216`
	`switch(config)# no shutdown`
		Creates the SVI used for the backup routed path over the vPC peer-link.

Step 18	(Optional) **delay restore interface-vlan** *seconds* **Example:** `switch(config-vpc-domain)# delay restore interface-vlan 45`	Enables the delay restore timer for SVIs. We recommend tuning this value when the SVI/VNI scale is high. For example, when the SCI count is 1000, we recommend that you set the delay restore to 45 seconds.
Step 19	**evpn multisite border-gateway** *ms-id* **Example:** `switch(config)# evpn multisite border-gateway 100`	Configures the site ID for a site/fabric. The range of values for *ms-id* is 1 to 281474976710655. The *ms-id* must be the same in all BGWs within the same fabric/site.
Step 20	**interface nve 1** **Example:** `switch(config-evpn-msite-bgw)# interface nve 1`	Creates a VXLAN overlay interface that terminates VXLAN tunnels. **Note** Only one NVE interface is allowed on the switch.
Step 21	**source-interface loopback** *src-if* **Example:** `switch(config-if-nve)# source-interface loopback 0`	Defines the source interface, which must be a loopback interface with a valid /32 IP address. This /32 IP address must be known by the transient devices in the transport network and the remote VTEPs. This requirement is accomplished by advertising the address through a dynamic routing protocol in the transport network.
Step 22	**host-reachability protocol bgp** **Example:** `switch(config-if-nve)# host-reachability protocol bgp`	Defines BGP as the mechanism for host reachability advertisement.

	Command or Action	Purpose
Step 23	**multisite border-gateway interface loopback** *vi-num* **Example:** `switch(config-if-nve)# multisite border-gateway interface loopback 100`	Defines the loopback interface used for the BGW virtual IP address (VIP). The BGW interface must be a loopback interface that is configured on the switch with a valid /32 IP address. This /32 IP address must be known by the transient devices in the transport network and the remote VTEPs. This requirement is accomplished by advertising the address through a dynamic routing protocol in the transport network. This loopback must be different than the source interface loopback. The range of *vi-num* is from 0 to 1023.
Step 24	**no shutdown** **Example:** `switch(config-if-nve)# no shutdown`	Negates the **shutdown** command.
Step 25	**exit** **Example:** `switch(config-if-nve)# exit`	Exits the NVE configuration mode.
Step 26	**interface loopback** *loopback-number* **Example:** `switch(config)# interface loopback 0`	Configures the loopback interface.
Step 27	**ip address** *ip-address* **Example:** `switch(config-if)# ip address 198.0.2.0/32`	Configures the primary IP address for the loopback interface.
Step 28	**ip address** *ip-address* **secondary** **Example:** `switch(config-if)# ip address 198.0.2.1/32 secondary`	Configures the secondary IP address for the loopback interface.
Step 29	**ip pim sparse-mode** **Example:** `switch(config-if)# ip pim sparse-mode`	Configures PIM sparse mode on the loopback interface.

Verifying the Multi-Site with vPC Support Configuration

To display Multi-Site with vPC support information, enter one of the following commands:

show vpc brief	Displays general vPC and CC status.
show vpc consistency-parameters global	Displays the status of those parameters that must be consistent across all vPC interfaces.

show vpc consistency-parameters vni	Displays configuration information for VNIs under the NVE interface that must be consistent across both vPC peers.

Output example for the **show vpc brief** command:

```
switch# show vpc brief
Legend:
                (*) - local vPC is down, forwarding via vPC peer-link

vPC domain id                     : 1
Peer status                       : peer adjacency formed ok     (<--- peer up)
vPC keep-alive status             : peer is alive
Configuration consistency status  : success (<----- CC passed)
Per-vlan consistency status       : success                       (<---- per-VNI CCpassed)
Type-2 consistency status         : success
vPC role                          : secondary
Number of vPCs configured         : 1
Peer Gateway                      : Enabled
Dual-active excluded VLANs        : -
Graceful Consistency Check        : Enabled
Auto-recovery status              : Enabled, timer is off.(timeout = 240s)
Delay-restore status              : Timer is off.(timeout = 30s)
Delay-restore SVI status          : Timer is off.(timeout = 10s)
Operational Layer3 Peer-router    : Disabled
[...]
```

Output example for the **show vpc consistency-parameters global** command:

```
switch# show vpc consistency-parameters global

    Legend:
        Type 1 : vPC will be suspended in case of mismatch

Name                        Type  Local Value            Peer Value
-------------               ----  ---------------------- -----------------------
[...]
Nve1 Adm St, Src Adm St,    1     Up, Up, 2.1.44.5, CP,  Up, Up, 2.1.44.5, CP,
Sec IP, Host Reach, VMAC          TRUE, Disabled,        TRUE, Disabled,
Adv, SA,mcast l2, mcast           0.0.0.0, 0.0.0.0,      0.0.0.0, 0.0.0.0,
l3, IR BGP,MS Adm St, Reo         Disabled, Up,          Disabled, Up,
                                  200.200.200.200        200.200.200.200
[...]
```

Output example for the **show vpc consistency-parameters vni** command:

```
switch(config-if-nve-vni)# show vpc consistency-parameters vni

    Legend:
        Type 1 : vPC will be suspended in case of mismatch

Name                        Type  Local Value            Peer Value
-------------               ----  ---------------------- -----------------------
Nve1 Vni, Mcast, Mode,      1     11577, 234.1.1.1,      11577, 234.1.1.1,
Type, Flags                       Mcast, L2, MS IR       Mcast, L2, MS IR
Nve1 Vni, Mcast, Mode,      1     11576, 234.1.1.1,      11576, 234.1.1.1,
Type, Flags                       Mcast, L2, MS IR       Mcast, L2, MS IR
[...]
```

Configuration Example for Multi-Site with Asymmetric VNIs

The following example shows how two sites with different sets of VNIs can connect to the same MAC VRF or IP VRF. One site uses VNI 200 internally, and the other site uses VNI 300 internally. Route-target auto no longer matches because the VNI values are different. Therefore, the route-target values must be manually configured. In this example, the value 222:333 stitches together the two VNIs from different sites.

The BGW of site 1 has L2VNI 200 and L3VNI 201.

The BGW of site 2 has L2VNI 300 and L3VNI 301.



Note This configuration example assumes that basic Multi-Site configurations are already in place.



Note You must have VLAN-to-VRF mapping on the BGW. This requirement is necessary to maintain L2VNI-to-L3VNI mapping, which is needed for reorigination of MAC-IP routes at BGWs.

Layer 3 Configuration

In the BGW node of site 1, configure the common RT 201:301 for stitching the two sites using L3VNI 201 and L3VNI 301:

```
vrf context vni201
  vni 201
  address-family ipv4 unicast
    route-target both auto evpn
    route-target import 201:301 evpn
    route-target export 201:301 evpn
```

In the BGW node of site 2, configure the common RT 201:301 for stitching the two sites using L3VNI 201 and L3VNI 301:

```
vrf context vni301
  vni 301
  address-family ipv4 unicast
    route-target both auto evpn
    route-target import 201:301 evpn
    route-target export 201:301 evpn
```

Layer 2 Configuration

In the BGW node of site 1, configure the common RT 222:333 for stitching the two sites using L2VNI 200 and L2VNI 300:

```
evpn
  vni 200 l2
    rd auto
    route-target import auto
    route-target import 222:333
    route-target export auto
    route-target export 222:333
```

For proper reorigination of L3 labels of MAC-IP routes, associate the VRF (L3VNI) to the L2VNI:

```
interface Vlan 200
  vrf member vni201
```

In the BGW node of site 2, configure the common RT 222:333 for stitching the two sites using L2VNI 200 and L2VNI 300:

```
evpn
  vni 300 l2
    rd auto
    route-target import auto
    route-target import 222:333
    route-target export auto
    route-target export 222:333
```

For proper reorigination of L3 labels of MAC-IP routes, associate the VRF (L3VNI) to the L2VNI:

```
interface vlan 300
  vrf member vni301
```

Advertise Using PIP Towards Fabric

Beginning with Cisco NX-OS Release 10.5(1)F, you can configure BGW to advertise external EVPN type-5 routes with PIP as next-hop and PIP's RMAC towards fabric side. With this configuration, BGW uses PIP instead of VIP for route advertisement.

Guidelines and Limitations

Following guidelines and limitations are applicable when you advertise route using PIP:

- Only L3 support is added in Cisco NX-OS Release 10.5(1)F.
- This feature is not applicable on VPC BGW.
- With this solution, traffic loss is expected until routes are updated on BGW and Leaf after remote BGW goes down.
- User must configure **maximum-paths** under EVPN and VRF's address-family on the leaf. This enables BGP to select all the paths as best-path or multi-paths and download all next-hops to forwarding plane to achieve load-balancing.
- In topologies with separate BGW and Spine, user must do one of the following 2 options:
  - Disable dual RD on BGWs.
  - configure add-path command in Spine to advertise all the EVPN paths to Leaf switches if dual RD is enabled on BGWs.
- The **fabric-advertise-pip l3** command has to be configured on all the BGWs on the same site.
- This solution is applicable to multiplane topology with only one BGW per plane per site. If more than one BGW per site is connected to a single plane, this solution is not required.
- When **fabric-advertise-pip l3** is enabled, BGWs accept remote type-5 routes from other BGWs on the same site with their PIP addresses. This causes increase in the number of paths per route on BGWs and is directly proportional to the number of BGWs on same site.

Configuring BGW to Advertise using PIP Towards Fabric

This section provides the configuration procedure for enabling advertise of remote routes with PIP as next-hop towards fabric on an Anycast BGW.

SUMMARY STEPS

1. **configure terminal**
2. **evpn multisite border-gateway** *ms-id*
3. **fabric-advertise-pip l3**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**evpn multisite border-gateway** *ms-id* **Example:** `switch(config)# evpn multisite border-gateway 100`	Configures the site ID for a site/fabric. The range of values for ms-id is 1 to 2,814,749,767,110,655. The ms-id must be the same in all BGWs within the same fabric/site.
Step 3	**fabric-advertise-pip l3** **Example:** `switch(config-evpn-msite-bgw)# fabric-advertise-pip l3`	Enables advertisement of remote EVPN type-5 routes with PIP next-hop towards fabric.

Verifying the Configuration

Use the **show nve interface nve 1 detail** command to verify the configuration:

```
switch(config)# show nve interface nve 1 detail
Interface: nve1, State: Up, encapsulation: VXLAN
VPC Capability: VPC-VIP-Only [not-notified]
Local Router MAC: 4464.3c31.802f
Host Learning Mode: Control-Plane
Source-Interface: loopback1 (primary: 20:1::21, secondary: 0.0.0.0)
Source Interface State: Up
Virtual RMAC Advertisement: No
NVE Flags:
Interface Handle: 0x49000001
Source Interface hold-down-time: 180
Source Interface hold-up-time: 30
Remaining hold-down time: 0 seconds
Virtual Router MAC: N/A
Virtual Router MAC Re-origination: 0022.3344.5566
Interface state: nve-intf-add-complete
Fabric convergence time: 37 seconds
```

```
Fabric convergence time left: 0 seconds
Multisite delay-restore time: 50 seconds
Multisite delay-restore time left: 0 seconds
Multisite dci-advertise-pip configured: False
Multisite fabric-advertise-pip l3 configured: True
```

TRM with Multi-Site

This section contains the following topics:

- Information About Configuring TRM with Multi-Site, on page 321
- Guidelines and Limitations for TRM with Multi-Site, on page 325
- Configuring TRM with Multi-Site, on page 329
- Verifying TRM with Multi-Site Configuration, on page 333

Information About Configuring TRM with Multi-Site

Tenant Routed Multicast (TRM) with Multi-Site enables multicast forwarding across multiple VXLAN EVPN fabrics that are connected via Multi-Site. This feature provides Layer 3 multicast services across sites for sources and receivers across different sites. It addresses the requirement of East-West multicast traffic between sites.

Each TRM site is operating independently. Border gateways on each site allow stitching across the sites. There can be multiple border gateways for each site. Multicast source and receiver information across sites is propagated by BGP on the border gateways that are configured with TRM. The border gateway on each site receives the multicast packet and re-encapsulates the packet before sending it to the local site. Beginning with Cisco NX-OS Release 10.1(2), TRM with Multi-Site supports both Anycast Border Gateway and vPC Border Gateway.

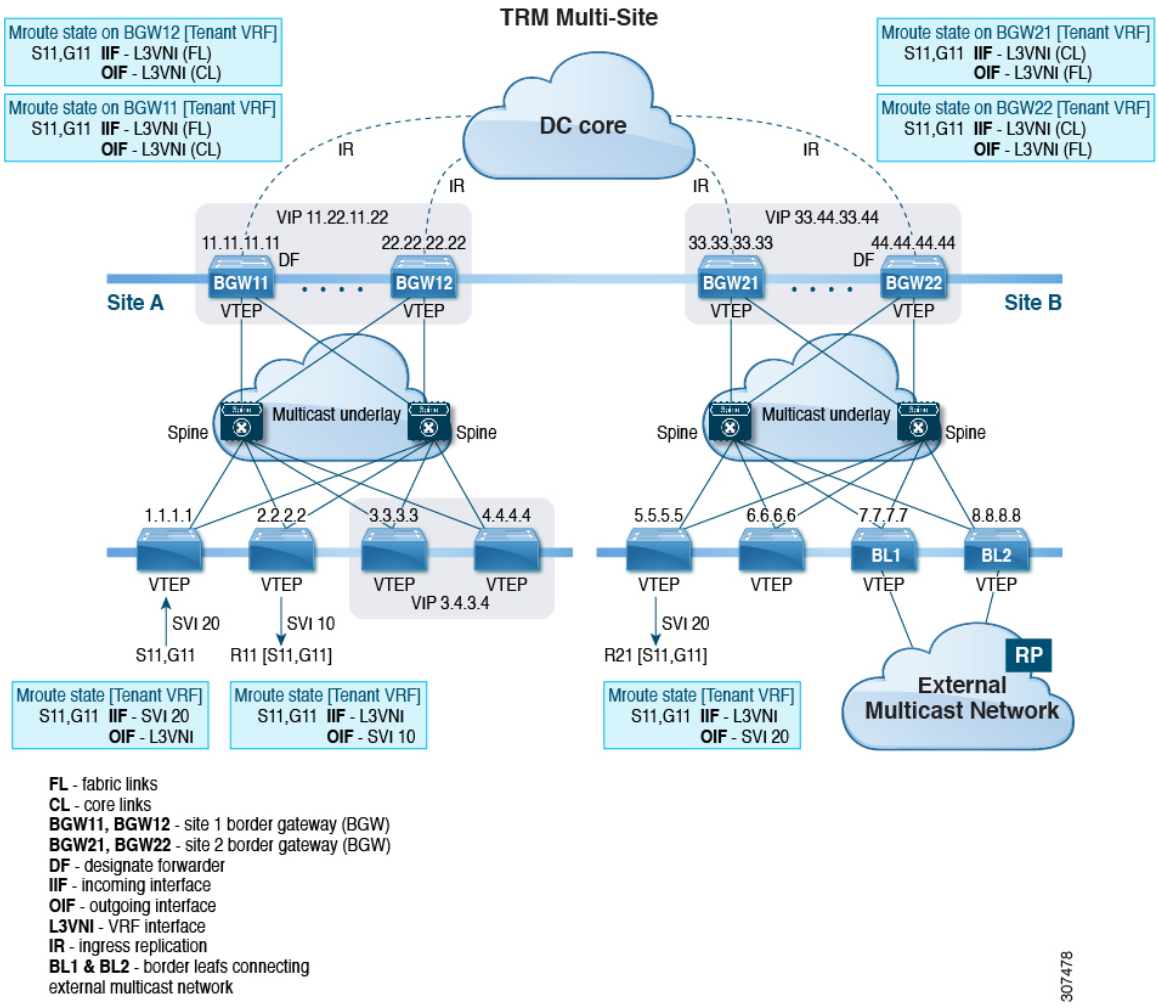
The border gateway that is elected as Designated Forwarder (DF) for the L3VNI forwards the traffic from fabric toward the core side. In the TRM Multicast-Anycast Gateway model, we use the VIP-R based model to send traffic toward remote sites. The IR destination IP is the VIP-R of the remote site. Each site that has the receiver gets only one copy from the source site. DF forwarding is applicable only on Anycast Border Gateways.



Note Only the DF sends the traffic toward remote sites.

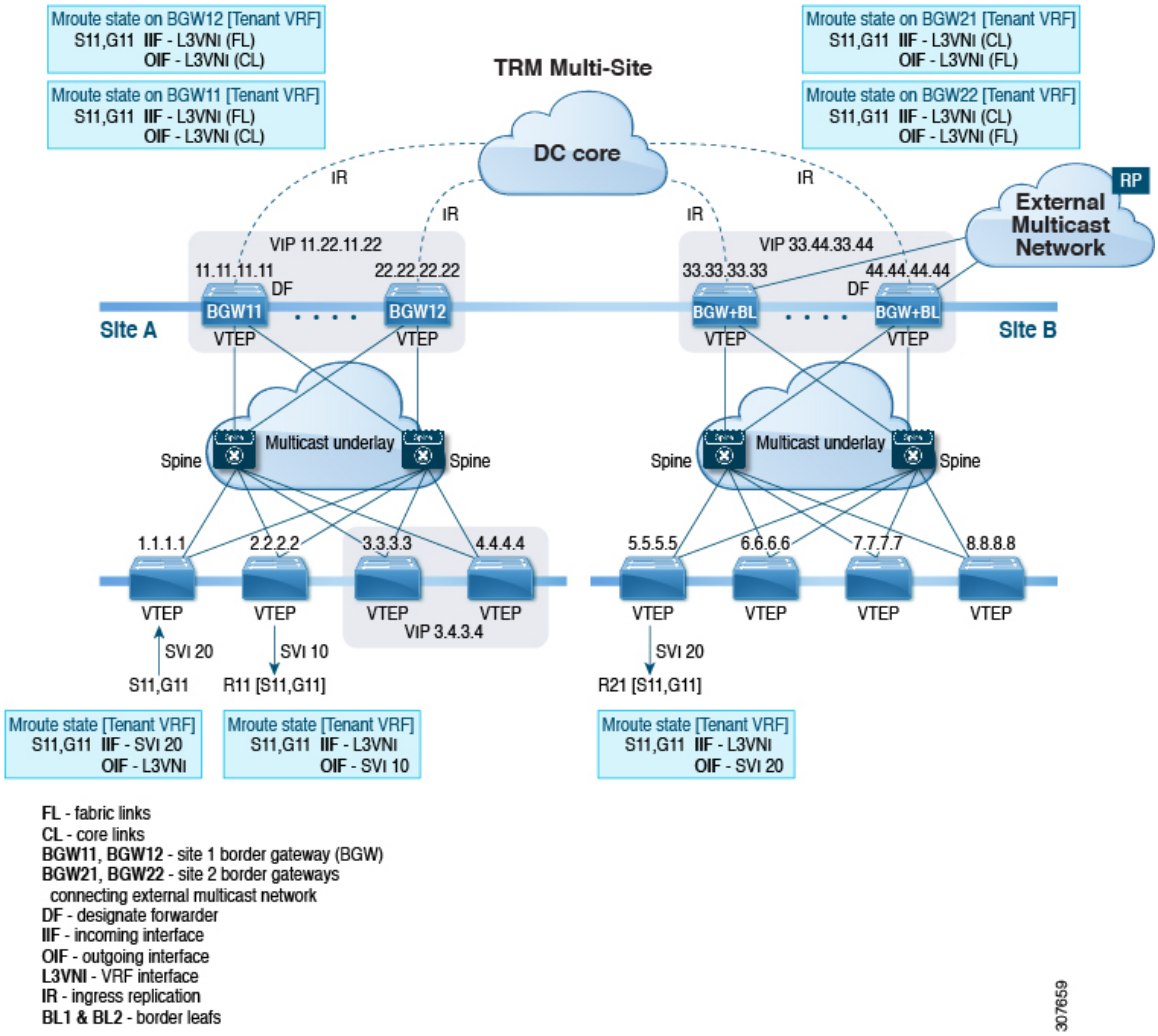
On the remote site, the BGW that receives the inter-site multicast traffic from the core forwards the traffic toward the fabric side. The DF check is not done from the core to fabric direction because non-DF can also receive the VIP-R copy from the source site.

Figure 30: TRM with Multi-Site Topology, BL External Multicast Connectivity



Beginning with Cisco NX-OS Release 9.3(3), TRM with Multi-Site supports BGW connections to the external multicast network in addition to the BL connectivity, which is supported in previous releases. Forwarding occurs as documented in the previous example, except the exit point to the external multicast network can optionally be provided through the BGW.

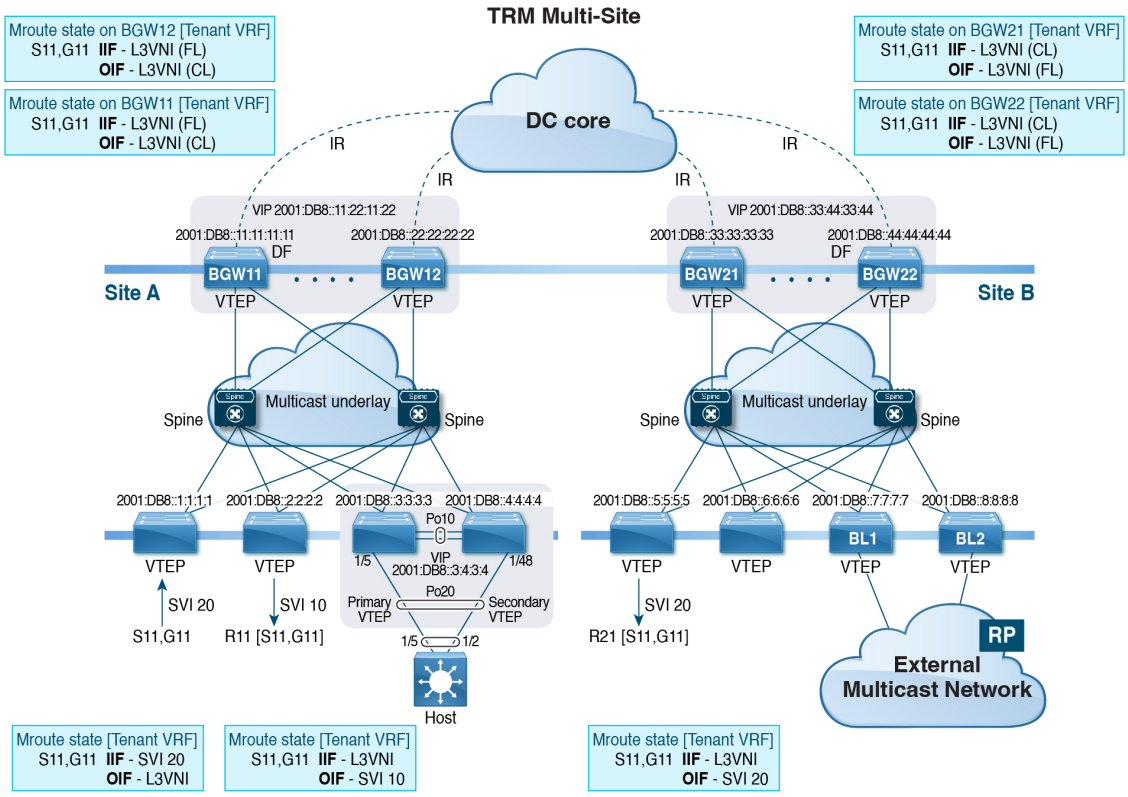
Figure 31: TRM with Multi-Site Topology, BGW External Multicast Connectivity



Information About Configuring TRM Multi-Site with IPv6 Underlay

Beginning with Cisco NX-OS Release 10.4(3)F, the support is provided for TRM Multi-Site with IPv6 Underlay.

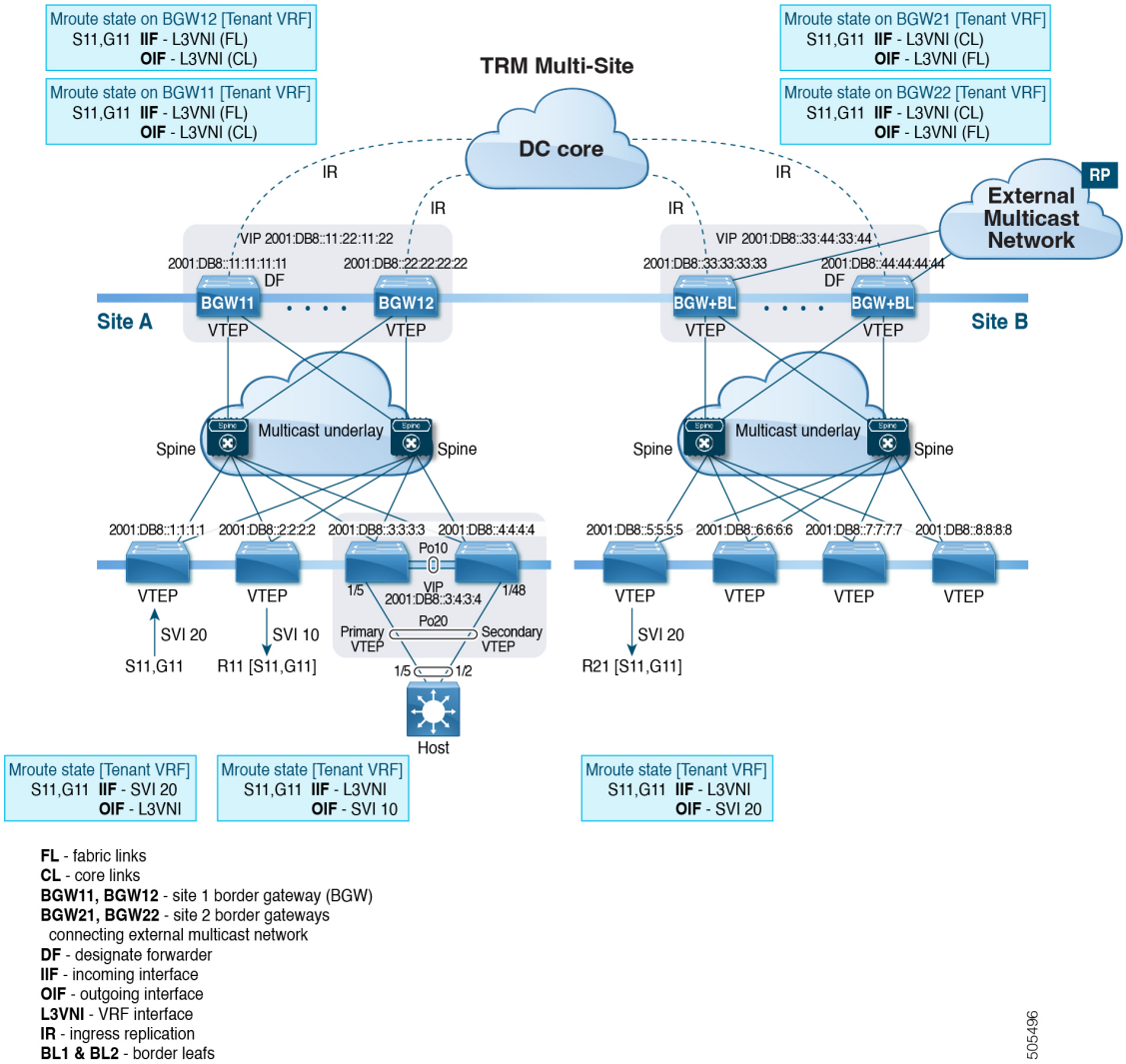
Figure 32: TRM Multi-Site with IPv6 Underlay Topology, BL External Multicast Connectivity



FL - fabric links
CL - core links
BGW11, BGW12 - site 1 border gateway (BGW)
BGW21, BGW22 - site 2 border gateway (BGW)
DF - designate forwarder
IIF - incoming interface
OIF - outgoing interface
L3VNI - VRF interface
IR - ingress replication
BL1 & BL2 - border leafs connecting external multicast network

505494

Figure 33: TRM Multi-Site with IPv6 Underlay Topology, BGW External Multicast Connectivity



FL - fabric links
CL - core links
BGW11, BGW12 - site 1 border gateway (BGW)
BGW21, BGW22 - site 2 border gateways connecting external multicast network
DF - designate forwarder
IIF - incoming interface
OIF - outgoing interface
L3VNI - VRF interface
IR - ingress replication
BL1 & BL2 - border leafs

The above topology shows four leafs and two spines in the VXLAN EVPN fabric and two Anycast BGWs. Inside the fabric, the underlay is an IPv6 Multicast running PIMv6. RP is positioned in the spine with anycast RP. BGWs support VXLAN with IPv6 Protocol-Independent Multicast (PIMv6) Any-Source Multicast (ASM) on the fabric side and Ingress Replication (IPv6) on the DCI side.

Beginning with Cisco NX-OS Release 10.5(1)F, the underlay network supports the following combinations for TRM Multi-Site:

- In the data center fabric, both Multicast Underlay (PIMv6) Any-Source Multicast (ASM) and Ingress Replication (IPv6) are supported.
- In the Data Center Interconnect (DCI), only Ingress Replication (IPv6) is supported.

Guidelines and Limitations for TRM with Multi-Site

TRM with Multi-Site has the following guidelines and limitations:

- The following platforms support TRM with Multi-Site:
  - Cisco Nexus 9300-FX/FX2/FX3 platform switches
  - Cisco Nexus 9300-GX platform switches
  - Cisco Nexus 9300-GX2 platform switches
  - Cisco Nexus 9332D-H2R switches
  - Cisco Nexus 93400LD-H1 switches
  - Cisco Nexus 9364C-H1 switches
  - Cisco Nexus 9500 platform switches with Nexus X97160YC-EX and FX/FX3 line cards
- Beginning with Cisco NX-OS Release 9.3(3), a border leaf and Multi-Site border gateway can coexist on the same node for multicast traffic.
- Beginning with Cisco NX-OS Release 9.3(3), all border gateways for a given site must run the same Cisco NX-OS 9.3(x) image.
- Cisco NX-OS Release 10.1(2) has the following guidelines and limitations:
  - You need to add a VRF lite link (per Tenant VRF) between the vPC peers in order to support the L3 hosts attached to the vPC primary and secondary peers.
  - Backup SVI is needed between the two vPC peers.
  - Orphan ports attached with L2 and L3 are supported with vPC BGW.
  - TRM multi-site with vPC BGW is not supported with vMCT.

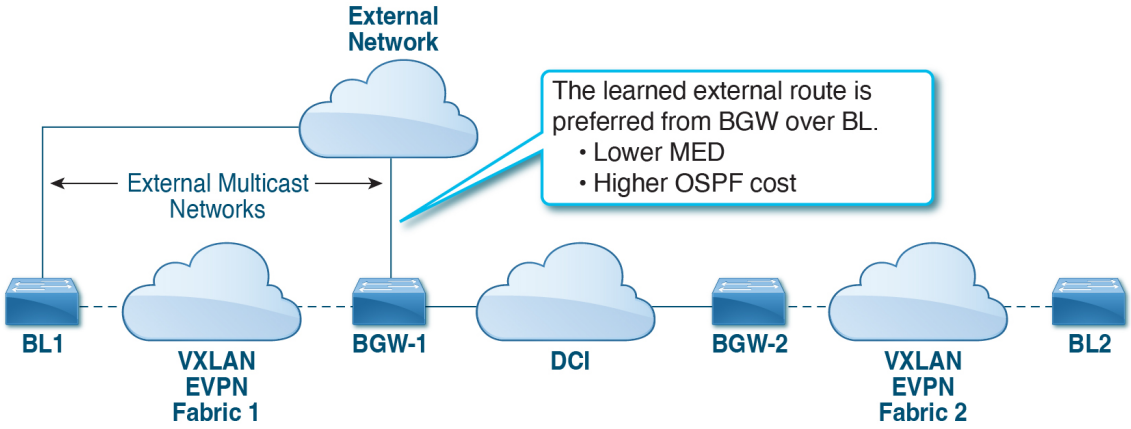
  For details on TRM and Configuring TRM with vPC Support, see Configuring Tenant Routed Multicast.
- TRM multi-site with vPC BGW and with Anycast BGW are supported on Cisco Nexus 9300FX, FX2, and FX3 family switches. Beginning with Cisco NX-OS Release 10.2(1)F, TRM with vPC BGW and with Anycast BGW are supported on Cisco Nexus 9300-GX family switches.
- Beginning with Cisco NX-OS Release 10.2(1q)F, TRM with Multi-Site is supported on the Cisco Nexus N9K-C9332D-GX2B platform switches.
- Beginning with Cisco NX-OS Release 10.2(1q)F, the TRM multi-site with vPC BGW and with Anycast BGW are supported on the Cisco Nexus C9332D-GX2B platform switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, the TRM multi-site with vPC BGW and with Anycast BGW are supported on the Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, the TRM multi-site with vPC BGW and with Anycast BGW are supported on the Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, the TRM multi-site with vPC BGW and with Anycast BGW are supported on the Cisco Nexus 9364C-H1 switches.
- Beginning with Cisco NX-OS Release 10.5(2)F, the TRM multi-site with vPC BGW and with Anycast BGW are supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.
- Beginning with Cisco NX-OS Release 10.2(2)F, multicast group configuration is used to encapsulate TRM and L2 BUM packets in the DCI core using the **multisite mcast-group** *dci-core-group* command.

- Beginning with Cisco NX-OS Release 10.2(3)F, the TRM multi-site is supported on the Cisco Nexus 9364D-GX2A and 9348D-GX2A switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, the TRM multi-site is supported on the Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, the TRM multi-site is supported on the Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, the TRM multi-site is supported on the Cisco Nexus 9364C-H1 switches.
- TRM with Multi-Site supports the following features:
  - TRM Multi-Site with vPC Border Gateway.
  - PIM ASM multicast underlay in the VXLAN fabric
  - TRM with Multi-Site Layer 3 mode only
  - TRM with Multi-Site with Anycast Gateway
  - Terminating VRF-lite at the border leaf
  - The following RP models with TRM Multi-Site:
    - External RP
    - RP Everywhere
    - Internal RP
- Only one pair of vPC BGW can be configured on one site.
- A pair of vPC BGW and Anycast BGW cannot co-exist on the same site.
- Prior to NX-OS 10.2(2)F only ingress replication was supported between DCI peers across the core. Beginning with Cisco NX-OS Release 10.2(2)F both ingress replication and multicast are supported between DCI peers across the core.
- Border routers reoriginate MVPN routes from fabric to core and from core to fabric.
- Only eBGP peering between border gateways of different sites is supported.
- Each site must have a local RP for the TRM underlay.
- Keep each site's underlay unicast routing isolated from another site's underlay unicast routing. This requirement also applies to Multi-Site.
- MVPN address family must be enabled between BGWs.
- When configuring BGW connections to the external multicast fabric, be aware of the following:
  - The multicast underlay must be configured between all BGWs on the fabric side even if the site doesn’t have any leafs in the fabric site.
  - Sources and receivers that are Layer-3 attached through VRF-Lite links to the BGW of a single site acting therefore also as Border Leaf (BL) node need to have reachability through the external Layer-3 network. If there's a Layer-3 attached source on BGW BL Node-1 and a Layer-3 attached receiver

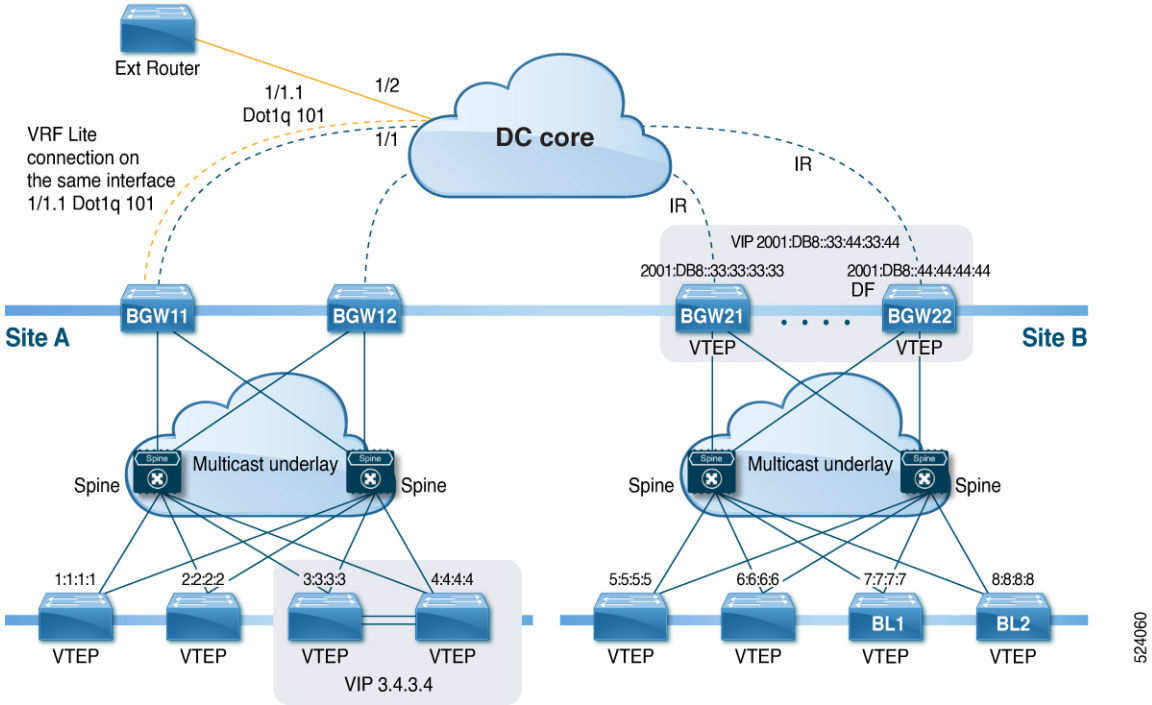
on BGW BL Node-2 for the same site, the traffic between these two endpoints flows through the external Layer-3 network and not through the fabric.

- External multicast networks should be connected only through the BGW or BL. If a deployment requires external multicast network connectivity from both the BGW and BL at the same site, make sure that external routes that are learned from the BGW are preferred over the BL. To do so, the BGW must have a lower MED and a higher OSPF cost (on the external links) than the BL.

  The following figure shows a site with external network connectivity through BGW-BLs and an internal leaf (BL1). The path to the external source should be through BGW-1 (rather than through BL1) to avoid duplication on the remote site receiver.



- The BGW supports VRF-lite hand-off and Multi-site configuration on the same physical interface as shown in the diagram.



- MED is supported for iBGP only.

Guidelines and Limitations for TRM Multi-Site with IPv6 Underlay

TRM Multi-Site with IPv6 Underlay has the following configuration guidelines and limitations:

- BGWs will support VXLAN with Protocol-Independent Multicast (PIMv6) Any-Source Multicast (ASM) on the fabric side and Ingress Replication (IPv6) on the DCI side.
- Cisco Nexus 9300-FX, FX2, FX3, GX, GX2, H2R and H1 ToR switches are supported as the leaf VTEP.
- Cisco Nexus N9K-X9716D-GX and N9K-X9736C-FX, N9K-X9736C-FX3 line cards are supported only on the spine (EoR).
- When an EoR is deployed as a spine node with Multicast Underlay (PIMv6) Any-Source Multicast (ASM), it is mandatory to configure non-default template using one of the following commands in global configuration mode:
  - **system routing template-multicast-heavy**
  - **system routing template-multicast-ext-heavy**
- Beginning with Cisco NX-OS Release 10.5(1)F, TRM Multi-Site in the data center fabric supports both Multicast Underlay (PIMv6) Any-Source Multicast (ASM) and Ingress Replication (IPv6) in the underlay. This support is available on the following switches and line cards:
  - Cisco Nexus 9300-FX, FX2, FX3, GX, GX2, H2R, and H1 ToR switches as the leaf VTEPs.
  - Cisco Nexus N9K-X9716D-GX and N9K-X9736C-FX line cards as spines if the underlay is configured for Multicast Underlay (PIMv6) Any-Source Multicast (ASM).
  - Cisco Nexus N9K-X9716D-GX and N9K-X9736C-FX line cards as VTEPs if the underlay uses Ingress Replication (IPv6).
- Beginning with Cisco NX-OS Release 10.5(2)F, TRM Multi-Site with IPv6 Underlay support is extended on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line cards as VTEPs if the underlay uses Ingress Replication (IPv6).

Configuring TRM with Multi-Site

Before you begin

The following must be configured:

- VXLAN TRM
- VXLAN Multi-Site

This section provides the configuration procedure for Anycast BGW with TRM. For vPC BGW with TRM, vPC must be configured along with VxLAN TRM and VxLAN Multi-site.

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**interface nve1** **Example:** `switch(config)# interface nve1`	Configures the NVE interface.
Step 3	**no shutdown** **Example:** `switch(config-if-nve)# no shutdown`	Brings up the NVE interface.
Step 4	**host-reachability protocol bgp** **Example:** `switch(config-if-nve)# host-reachability protocol bgp`	Defines BGP as the mechanism for host reachability advertisement.
Step 5	**source-interface loopback** *src-if* **Example:** `switch(config-if-nve)# source-interface loopback 0`	Defines the source interface, which must be a loopback interface with a valid /32 IP address. This /32 IP address must be known by the transient devices in the transport network and the remote VTEPs. This requirement is accomplished by advertising the address through a dynamic routing protocol in the transport network.
Step 6	**multisite border-gateway interface loopback** *vi-num* **Example:** `switch(config-if-nve)# multisite border-gateway interface loopback 1`	Defines the loopback interface used for the border gateway virtual IP address (VIP). The border-gateway interface must be a loopback interface that is configured on the switch with a valid /32 IP address. This /32 IP address must be known by the transient devices in the transport network and the remote VTEPs. This requirement is accomplished by advertising the address through a dynamic routing protocol in the transport network. This loopback must be different than the source interface loopback. The range of *vi-num* is from 0 to 1023.
Step 7	**member vni** *vni-range* **associate-vrf** **Example:** `switch(config-if-nve)# member vni 10010 associate-vrf`	Configures the virtual network identifier (VNI). The range for *vni-range* is from 1 to 16,777,214 The value of *vni-range* can be a single value like 5000 or a range like 5001-5008.
Step 8	**mcast-group** *ip-addr* **Example:** `switch(config-if-nve-vni)# mcast-group 225.0.0.1`	Configures the NVE multicast group IP prefix within the fabric.

	Command or Action	Purpose
Step 9	**multisite mcast-group** *dci-core-group address* **Example:** `switch(config-if-nve-vni)# multisite mcast-group 226.1.1.1`	Configures the multicast group which is used to encapsulate TRM and L2 BUM packets in the DCI core.
Step 10	**multisite ingress-replication optimized** **Example:** `switch(config-if-nve-vni)# multisite ingress-replication optimized`	Defines the Multi-Site BUM replication method for extending the Layer 2 VNI.

Configuring TRM Multi-Site with IPv6 Underlay

This section provides the configuration procedure on Anycast BGW for TRM with IPv6 Multicast Underlay with Protocol-Independent Multicast (PIMv6) Any-Source Multicast (ASM) on the fabric side and Ingress Replication (IPv6) on the DCI side.

Before you begin

The following must be configured:

- VXLAN TRM
- VXLAN Multi-Site

SUMMARY STEPS

1. **configure terminal**
2. **interface nve1**
3. **no shutdown**
4. **host-reachability protocol bgp**
5. **source-interface loopback** *src-if*
6. **multisite border-gateway interface loopback** *vi-num*
7. **member vni** *vni-range* **associate-vrf**
8. **mcast-group** *ipv6-addr*
9. **multisite ingress-replication optimized**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.

	Command or Action	Purpose
Step 2	**interface nve1** **Example:** `switch(config)# interface nve1`	Configures the NVE interface.
Step 3	**no shutdown** **Example:** `switch(config-if-nve)# no shutdown`	Brings up the NVE interface.
Step 4	**host-reachability protocol bgp** **Example:** `switch(config-if-nve)# host-reachability protocol bgp`	Defines BGP as the mechanism for host reachability advertisement.
Step 5	**source-interface loopback** *src-if* **Example:** `switch(config-if-nve)# source-interface loopback 0`	Defines the source interface, which must be a loopback interface with a valid /128 IPv6 address. This /128 IPv6 address must be known by the transient devices in the transport network and the remote VTEPs. This requirement is accomplished by advertising the address through a dynamic routing protocol in the transport network.
Step 6	**multisite border-gateway interface loopback** *vi-num* **Example:** `switch(config-if-nve)# multisite border-gateway interface loopback 1`	Defines the loopback interface used for the border gateway virtual IP address (VIP). The border-gateway interface must be a loopback interface that is configured on the switch with a valid /128 IPv6 address. This /128 IPv6 address must be known by the transient devices in the transport network and the remote VTEPs. This requirement is accomplished by advertising the address through a dynamic routing protocol in the transport network. This loopback must be different than the source interface loopback. The range of *vi-num* is from 0 to 1023.
Step 7	**member vni** *vni-range* **associate-vrf** **Example:** `switch(config-if-nve)# member vni 90001 associate-vrf`	Configures the virtual network identifier (VNI). The range for *vni-range* is from 1 to 16,777,214. The value of *vni-range* can be a single value like 5000 or a range like 5001-5008.
Step 8	**mcast-group** *ipv6-addr* **Example:** `switch(config-if-nve-vni)# mcast-group ff03:ff03::101:1`	Configures the NVE multicast group IPv6 prefix within the fabric.
Step 9	**multisite ingress-replication optimized** **Example:** `switch(config-if-nve-vni)# multisite ingress-replication optimized`	Defines the Multi-Site replication method for extending TRM functionality across sites.

Verifying TRM with Multi-Site Configuration

To display the status for the TRM with Multi-Site configuration, enter the following command:

Command	Purpose
show nve vni *virtual-network-identifier*	Displays the L3VNI. **Note** For this feature, optimized IR is the default setting for the Multi-Site extended L3VNI. MS-IR flag inherently means that it's MS-IR optimized.

Example of the **show nve vni** command:

For IPv4

```
switch(config)# show nve vni 51001
Codes: CP - Control Plane        DP - Data Plane
       UC - Unconfigured         SA - Suppress ARP
       SU - Suppress Unknown Unicast
       Xconn - Crossconnect
       MS-IR - Multisite Ingress Replication

Interface VNI      Multicast-group   State Mode Type [BD/VRF]      Flags
--------- -------- ----------------- ----- ---- ------------------ -----
nve1      51001    226.0.0.1         Up    CP   L3 [cust_1]        MS-IR
```

For IPv6

```
switch(config)# show nve vni 90001
Codes: CP - Control Plane        DP - Data Plane
       UC - Unconfigured         SA - Suppress ARP
       S-ND - Suppress ND
       SU - Suppress Unknown Unicast
       Xconn - Crossconnect
       MS-IR - Multisite Ingress Replication
       HYB - Hybrid IRB mode

Interface VNI      Multicast-group   State Mode Type [BD/VRF]      Flags
--------- -------- ----------------- ----- ---- ------------------ -----
nve1      90001    ff03:ff03::101:1  Up    CP   L3 [v1]            MS-IR

switch(config)#
```

VXLAN EVPN Multi-Site with RFC 5549 Underlay

VXLAN EVPN Multi-Site with RFC 5549 Underlay Overview

The VXLAN EVPN Multi-Site with RFC 5549 underlay feature is introduced for deployments where you already have a VXLAN IPv4 network infrastructure and must integrate IPv6 functionality.

The RFC 5549 enables the advertisement of underlay IPv4 prefixes using BGP with an IPv6 address as the next hop, effectively allowing IPv4 connectivity to be established over an IPv6 underlay network.

Benefits of VXLAN EVPN Multi-Site with RFC 5549 Underlay

- Helps your network against the exhaustion of IPv4 addresses.
- Enhances scalability and efficiency by integrating IPv6 into the VXLAN IPv4 network infrastructure.
- Offers simpler, more direct communication between devices.
- Provides sub-second convergence.
- Supports the following topologies:
  - Border spine gateway to leaf.
  - vPC BGWs to leaf.
  - Anycast BGWs to leaf.

Supported Platform and Release for VXLAN EVPN Multi-Site with RFC 5549 Underlay

Feature	Release	Platform
VXLAN BGP EVPN with RFC 5549 Underlay on multi-site	10.5(1)F	Cisco Nexus 9000 Series switches

Functionalities of VXLAN EVPN Multi-Site with RFC 5549 Underlay

This section describes how the VXLAN BGP EVPN with RFC 5549 underlay functions on a multi-site environments.

VTEP IPv4 Address Advertisement

The VTEP addresses are advertised in BGP with RFC 5549 underlay such that the IPv4 VTEP addresses are routed using the following IPv6 enabled interfaces.

- **Link-local address (LLA)**: In this type of BGP peering configuration, BGP uses the interface's link-local address to establish IPv6 sessions, allowing for the advertisement of IPv4 addresses through them. This approach eliminates the necessity for configuring global addresses on interfaces.
- **Global IPv6 address**: In this type of BGP peering configuration, BGP creates a standard IPv6 neighbor relationship using the global IPv6 address of the directly connected interface. Therefore, it is required to configure a global IPv6 address on the directly connected interface for the peering to be established.

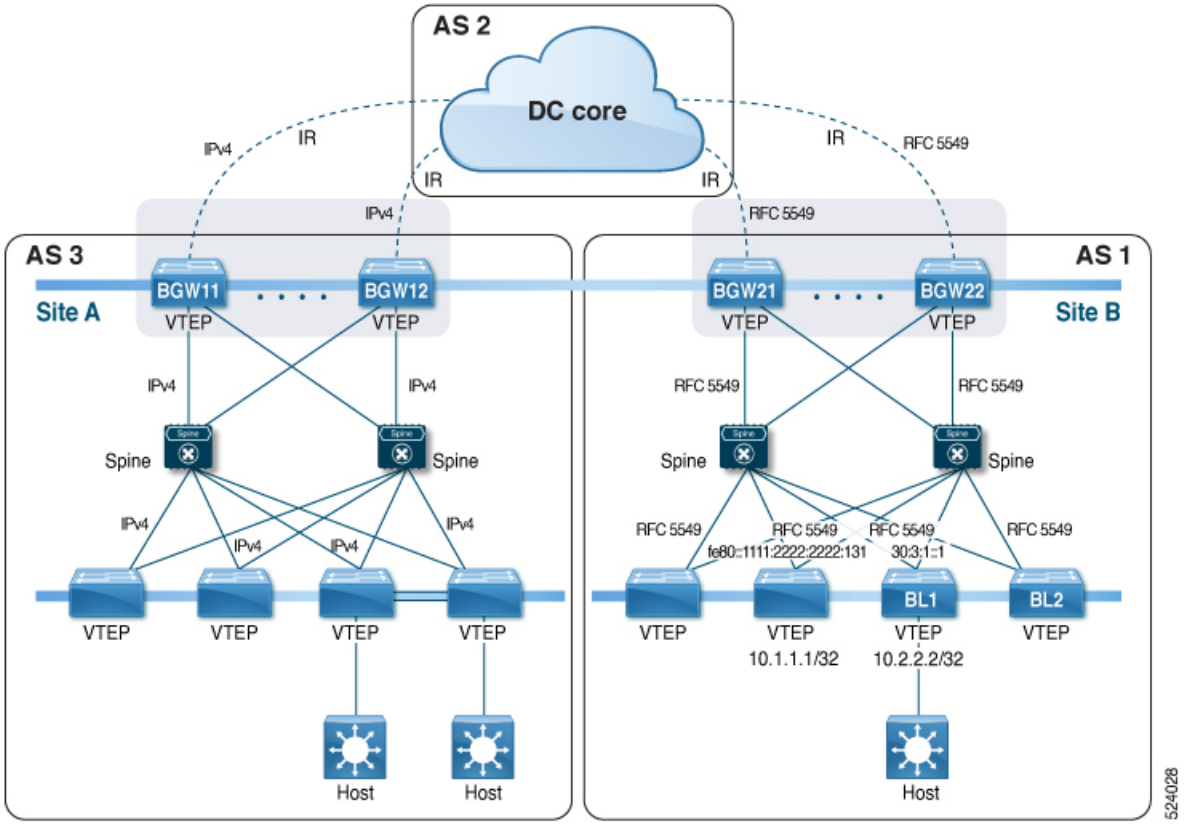
VXLAN EVPN Services Advertisement over IPv4 Session

When VTEPs IPv4 reachability is advertised with RFC 5549 underlay, an EVPN IPv4 BGP session must be created to advertise VXLANv4 services on a multi-site environment.

VXLAN BGP EVPN with RFC 5549 Underlay on Multi-Site

On Multi-Site, the VXLAN BGP EVPN with RFC 5549 Underlay supports node to node communication as shown in the diagram.

The BGW's VTEP IPv4 address is advertised using IPv6 address (LLA or Global address) as the next-hop address in RFC 5549 underlay between the sites.



Guidelines for VXLAN EVPN Multi-Site with RFC 5549 Underlay

- VTEP must be from the same address family of IPv4.
- VTEP loop-back address must be an IPv4 address.
- During upgrading, ensure that the supported release versions are used for nodes as mentioned below:

Nodes	Supported Release
Core	10.2(3)F or higher
Spine	10.2(3)F or higher
Leaf	10.2(3)F or higher
Border Gateway	10.5(1)F or higher

- The routes to remote multi-site BGW loopback addresses must always prioritize the DCI link path over the iBGP protocol between vPC Border Gateway switches configured using the backup SVI. The backup SVI must be used strictly in the event of a DCI link failure.
- Configuring BGP peering on multiple interfaces with the same Link-Local Address (LLA) is not supported. If such a configuration is implemented, the resulting behavior will be unpredictable.

- Beginning with Cisco NX-OS Release 10.6(1)F, RFC5549 underlay is supported on Cisco Nexus 9364E-SG2-Q and 9364E-SG2-O switches.

Supported and Unsupported Features of VXLAN EVPN Multi-Site with RFC 5549 Underlay

- Following are the supported and unsupported features for VXLAN BGP EVPN with RFC 5549 Underlay on multi-site environment:

Features	Supported Release	Supported or Not Supported
Ingress Replication	10.5(1)F	Supported
Overlay TCP Session	10.5(1)F	Supported
DSVNI	10.5(1)F	Supported
Anycast BGW	10.5(1)F	Supported
vPC BGW	10.5(1)F	Supported
vMCT	10.5(1)F	Supported
Route Leaking: Host addresses and between non default customer VRFs	10.5(1)F	Supported
IPv4 over RFC 5549 with default VRF	10.5(1)F	Supported
IPv4 over RFC 5549 with tenant VRF	-	Not Supported
Policy Based Routing	-	Not Supported
NGOAM	-	Not Supported
ND Suppression	-	Not Supported
ARP Suppression	-	Not Supported
IPv4 VNF: IPv4 services over IPv4 PE-CE BGP session	-	Not Supported
RFC 5549 VNF: IPv4 services over IPv6 PE-CE BGP session	-	Not Supported
IPv6 VNF: IPv6 services over IPv6 PE-CE BGP session	-	Not Supported
vPC	-	Not Supported
BFD	-	Not Supported

Features	Supported Release	Supported or Not Supported
Underlay multicast	-	Not Supported
MultiCast/TRM	-	Not Supported
Firewall Cluster	-	Not Supported
New L3VNI config	-	Not Supported
Group Policy Option	-	Not Supported
First Hop Security	-	Not Supported
PVLAN over VXLAN	-	Not Supported
VXLAN-TE	-	Not Supported
CloudSec	-	Not Supported

Configuring VXLAN EVPN Multi-Site with RFC 5549 Underlay on Multi-Site

You can perform the configuration of VXLAN EVPN Multi-Site with RFC 5549 underlay using LLA or global IPv6 address on a multi-site environment.

To configure BGW VXLAN EVPN Multi-Site with RFC 5549 underlay, follow these steps:

Procedure

Step 1 **configure terminal**

Example:

```
leaf# config terminal
leaf(config)#
```

Enters global configuration mode.

Step 2 **interface ethernet** *port*

Example:

```
leaf(config)# interface ethernet1/2
leaf(config-if)#
```

Enters Link local interface configuration mode for RFC 5549.

Step 3 **ip forward**

Example:

```
leaf(config-if)# ip forward
```

Enables IPv4 based lookup even when the interface VLAN has no IP address defined.

Step 4 To generate link-local address, use any of the following options:

- Link-local IPv6 address

  OR
- Global IPv6 address

a) To configure LLA, use any of the following options:

- **ipv6 link-local** *LL_ipv6_address* (Manual option)

  OR
- **ipv6 address use-link-local-only** (Auto option)

  OR
- **ipv6 link-local use-bia** (Auto option)

Example:

```
leaf(config-if)# ipv6 link-local fe80::1111:2222:2222:3101
```

OR

```
leaf(config-if)# ipv6 address use-link-local-only
```

OR

```
leaf(config-if)# ipv6 link-local use-bia
```

Configures the link-local IPv6 address for the interface based on the specified choice.

OR

a) Global IPv6 address: By default generates the LLA

Example:

ipv6 address *ipv6_address*

```
leaf(config-if)# ipv6 address 2000:1:1::1/64
```

Configures the Global IPv6 address for the interface.

Step 5 **exit**

Example:

```
leaf(config-if)# exit
leaf(config)#
```

Exits the interface configuration mode.

Step 6 **router bgp** *as-number*

Example:

```
leaf(config)# router bgp 1
leaf(config-router)#
```

Enters BGP router configuration mode.

Step 7 **neighbor** [*ipv6_address* | *ethernet port*]

Example:

```
leaf(config-router)# neighbor ethernet1/2
```

Configures a BGP neighbor for an interface.

Step 8 **remote-as** *value*

Example:

```
leaf(config-router)# remote-as 2
leaf(config-router-neighbor)#
```

Configures remote peer's autonomous system number for BGP neighbor.

Step 9 **peer-type fabric-external**

Example:

```
leaf(config-router-neighbor)# peer-type fabric-external
```

Enables the next hop rewrite for Multi-Site. Defines site external BGP neighbors for EVPN exchange. The default for **peer-type** is **fabric-internal**.

Note
The **peer-type fabric-external** command is required only for VXLAN Multi-Site BGWs.

Step 10 **address-family ipv4 unicast**

Example:

```
leaf(config-router-neighbor)# address-family ipv4 unicast
leaf(config-router-af)#
```

Configures the IPv4 unicast address family.

Step 11 **disable-peer-as-check**

Example:

```
leaf(config-router-af)# disable-peer-as-check
```

Disables checking the peer AS number during route advertisement. Configure this parameter on the spine for eBGP when all leafs are using the same AS but the spines have a different AS than leafs.

Note
This command is required for eBGP. For more information on eBGP configuration, see eBGP Underlay IP Network , on page 25.

Examples of VXLAN EVPN Multi-Site with RFC 5549 Underlay for Multi-Site

- The following example shows how the IPv4 address is advertised over the IPv6 interface from spine to leaf using LLA address:

```
spine# show ip route 10.1.1.1
IP Route Table for VRF "default"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

fe80::1111:2222:2222:131, ubest/mbest: 1/0
    *via 10.1.1.1/32%default, Eth1/2, [200/0], 6d09h, bgp-2, external, tag 2
```

- The following example shows how the IPv4 address is advertised over the IPv6 interface from spine to leaf using global IPv6 address:

```
spine# show ip route 10.2.2.2
IP Route Table for VRF "default"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

10.2.2.2/32, ubest/mbest: 1/0
    *via 30:3:1::1%default, Eth1/2, [200/0], 6d09h, bgp-2, external, tag 2
```



CHAPTER 15

Configure VXLAN EVPN Traffic Engineering - Multi-Site Egress Load Balancing

This chapter describes how to configure the VXLAN EVPN Traffic Engineering (TE) - Multi-Site Egress Load-Balancing feature on Cisco NX-OS devices.

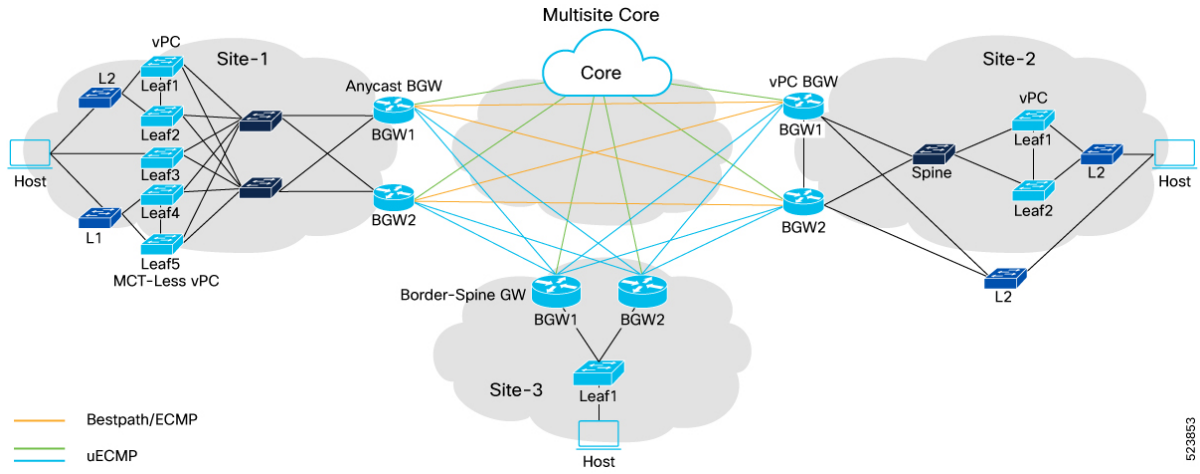
- VXLAN EVPN TE - Multi-Site egress load balancing, on page 341
- Guidelines and Limitations for VXLAN EVPN TE - Multi-Site Egress Load-Balancing , on page 342
- VXLAN EVPN TE - Multi-Site egress load balancing configurations, on page 344
- Verifying VXLAN EVPN TE - Multi-Site Egress Load-Balancing Configuration, on page 354
- Configuration examples for VXLAN EVPN TE - Multi-Site egress load balancing, on page 356

VXLAN EVPN TE - Multi-Site egress load balancing

The VXLAN EVPN TE - Multi-Site egress load balancing feature is a technology in the networking category that

- facilitates traffic steering,
- enables load balancing for data sent between different fabrics across multi-site links, and
- operates across an IP-based underlay network, also known as the Inter-site Network (ISN) to provide IP traffic engineering for overlay-encapsulated traffic over the underlay.

Topology of VXLAN EVPN TE - Multi-Site egress load balancing



The VXLAN EVPN TE - Multi-Site egress load balancing enhances the use of inter-data center links.

- In this topology, three VXLAN EVPN fabrics, that belong to the same Multi-Site domain, connect to the network through a local tier of Border Gateway devices (BGWs) at remote sites.
- Deployment of various BGW types, such as Anycast BGWs, vPC BGWs, and Border Gateway Spine devices is supported.
- Different connectivity options are allowed for interconnection such as
  - through direct BGW-to-BGW links, or
  - through a generic core infrastructure (ISN).
- Typically, the paths designated in orange serve as the best path or multipath for communication between endpoints at site-1 and site-2.
- Enabling the VXLAN EVPN TE - Multi-Site egress load balancing feature gives access to additional traffic distribution paths beyond the best path such as
  - through an intermediary location, referred to as site-3 (highlighted in blue), or
  - through a generic core infrastructure (highlighted in green).

  The alternative paths are available as part of unequal equal-cost multipath (uECMP) or weighted unequal equal-cost multipath (wuECMP).

Guidelines and Limitations for VXLAN EVPN TE - Multi-Site Egress Load-Balancing

- Beginning with Cisco NX-OS Release 10.4(3)F, the VXLAN EVPN TE - Multi-Site Egress Load-Balancing feature is supported on Cisco Nexus 9300 FX/FX2/FX3/GX/GX2 switches, and 9700-FX/GX line cards. However, only BGP-based underlay routing is currently supported.

- Beginning with Cisco NX-OS Release 10.5(2)F, the VXLAN EVPN TE - Multi-Site Egress Load-Balancing feature is supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.
- The VXLAN EVPN TE - Multi-Site Egress Load-Balancing feature is not supported on Cisco Nexus 9500 modular switches with 9500-FM-E Fabric Modules.
- VXLAN EVPN TE - Multi-Site Egress Load-Balancing supports the following features:
  - Load-balancing (LB) of egress traffic across underlay paths that may not be all best paths to remote sites.
  - Explicit and automatic LB policies for associating a specific “weight” (or load) to each multisite path. This comes with the two options of Weighted Equal-Cost Multi-Path (wECMP) and Weighted Unequal Equal-Cost Multipath (wuECMP).
  - BGP-based underlay routing.
  - AS-Path based uECMP path selection.
  - Layer 3 Unicast uECMP/wuECMP in Underlay.
  - Layer 3 (EVPN Type-5) and Layer 2 (EVPN Type-2) Overlay Unicast ECMP/wuECMP.
  - BUM forwarding:
    - BUM traffic will not be subject to egress load-balancing.
    - However, when ingress-replication is used for BUM forwarding across sites (DCI IR), the underlay next-hop path selection can use one among the egress load-balanced paths part of the multipath set and hence can benefit for uECMP in the underlay.
  - Software forwarding: In the baseline case for software forwarding, both Layer 2 and Layer 3 traffic support uECMP/wuECMP on an EVPN VXLAN Multisite setup.
  - CloudSec (PIP next-hop).
  - Firewall Clustering (PIP next-hop).
  - Downstream VNI.



Note The **dci-advertise-pip** command is required on the BGWs to start advertising Type-2 and Type-5 EVPN prefixes with PIP as next-hop (instead than the Multi-Site VIP). In Dynamic weight (wuECMP) in Overlay and Underlay, and with AIGP in Underlay, on page 370 section, more information is provided on why it is required to perform wuECMP load-balancing for overlay traffic destined to not equal cost next-hop PIP addresses.

  - VXLAN OAM.
  - VXLAN Policy-Based Routing (PBR).
- VXLAN EVPN TE - Multi-Site Egress Load-Balancing does not support the following features:
  - VXLAN with IPv6 Underlay.

- The **hardware profile ecmp resilient** configuration with weighted ECMP/uECMP.
- Multicast overlay traffic.
- The wuECMP is not supported on Cisco Nexus 9500 Series switches with 9700-FX/FX3 line cards.

- For a successful downgrade from Cisco NX-OS Release 10.4(3)F to a prior release, ensure that the VXLAN EVPN TE - Multi-Site Egress Load-Balancing configuration has been removed.

VXLAN EVPN TE - Multi-Site egress load balancing configurations

The configuration of the VXLAN EVPN TE - Multi-Site egress load balancing feature includes three main configurations.



Note Apply these configurations to the BGW only.

- **Creation of a Filter Policy**: Identifies the remote destination overlay next-hop IP address to which we want to distribute traffic across the underlay path part of the same multipath set. This address could be the common Multi-Site VIP of the remote BGWs or their unique PIPs (when **dci-advertise-pip** is configured).
- **Creation of a Multipath Policy**: Defines the criteria that classify underlay paths as part of a multipath set. This definition enables the provisioning of multiple use cases, such as uECMP and wuECMP with static or dynamic weights. Specifically, in cases where multiple overlay next-hops are present (such as BGWs' PIP addresses), wuECMP can be extended to include the selection of next-hop addresses, creating a “multi-level” load-sharing effect.
- **Enabling Resolution in the ELB VRF**: Enables the resolution of underlay paths using the underlay table in **egress-loadbalance-resolution- VRF** to reach the destination overlay next-hop IP address, rather than using the routing table in the default VRF.

Egress load balance resolution VRF

The **egress-loadbalance-resolution- VRF** is a new internal VRF that is created automatically by the system. This VRF is not configurable and cannot be deleted.

When the VXLAN EVPN TE - Multi-Site egress load balancing feature is configured, this new VRF serves to

- import and install the underlay protocol (currently BGP only) routes into this table,
- perform overlay next-hop resolution in this table instead of the default table, and
- allow more underlay paths for intersite communication, on top of the best paths installed in the default VRF.

Create an egress load balance filter policy for underlay

You can configure the filter policy to identify the underlay routes, such as remote VTEPs of interest, which require egress Load-balancing across multiple underlay paths.

To configure the egress load balance filter policy on the BGW, follow these steps:

Procedure

Step 1 Run the **ip prefix-list** *prefix-list-name* **seq** *value* **permit** *nexthop-ipaddress* command in configuration mode to configure the prefix list to match the remote next-hops.

Example:

```
switch# config terminal
switch(config)# ip prefix-list remote_nexthop seq 5 permit 10.10.112.1/32
```

Step 2 Run the **route-map** *route-map-name* command to create a route map.

Example:

```
switch(config)# route-map ROUTE_MAP_1
switch(config-route-map)#
```

The egress load balance logic applies only to VTEP routes or next-hops specified in the filter policy using a route map.

Step 3 Match the underlay routes using either the community attribute or a prefix-list.

Example:

```
switch(config-route-map)# match ip address prefix-list remote_nexthop
```

OR

```
switch(config-route-map)# match community BGPCommunity
```

- **match ip address prefix-list** *prefix-list-name*

  OR

- **match community** *community-list*

Step 4 Run the **router bgp** *as-number* command in global configuration mode to enter BGP router configuration mode.

Example:

```
switch(config)# router bgp 65001
switch(config-router)#
```

Step 5 Run the **address-family ipv4 unicast** command to configure the IPv4 unicast address family.

Example:

```
switch(config-router)# address-family ipv4 unicast
switch(config-router-af)#
```

Step 6 Run the **[no] load-balance egress filter-policy route-map** *route-map-name* command to restrict the egress load balance (ELB) to the VTEP routes of interest by enabling the filter policy.

Example:

```
switch(config-router-af)# load-balance egress filter-policy route-map ROUTE_MAP_1
```

You can tag the underlay routes with either a community attribute or a prefix-list. Use the command in the **no** form to remove the filter policy.

Note

- When filtered by community, the advertising egress BGW must tag all paths for a given route with the same community.
- Ensure that you configure the filter policy so that the system can compute ELB's paths for the prefixes.

Creating an Egress Load-Balancing Auto-Multipath Policy for Underlay

You can configure an auto-multipath policy using a route-map that specifies the maximum number of underlay paths to be computed and the criteria for assigning those underlay paths to the same multi-path sets. This policy can match one or more configured thresholds, such as AIGP metric or AS-Path difference, when compared to the best path. In the absence of an auto-multipath policy, only the best path is installed.

To configure an Egress Load Balancing Auto-Multipath policy for the underlay, follow these steps:

SUMMARY STEPS

1. **configure terminal**
2. **ip prefix-list** *prefix-list-name* **seq** *value* **permit** *nexthop-ipaddress*
3. **route-map** *route-map-name*
4. **exit**
5. **router bgp** *as-number*
6. **address-family ipv4 unicast**
7. **[no] load-balance egress multipath auto-policy route-map** *route-map-name*

DETAILED STEPS

Procedure

Step 1 **configure terminal**

Example:

```
switch# config terminal
switch(config)#
```

Enters global configuration mode.

Note
Proceed with step 2 only if the prefix-list has not been created.

Step 2 **ip prefix-list** *prefix-list-name* **seq** *value* **permit** *nexthop-ipaddress*

Example:

```
switch(config)# ip prefix-list remote_nexthop seq 5 permit 10.10.112.1/32
```

Configures the prefix list to match the remote next-hops.

Step 3 **route-map** *route-map-name*

This route-map is to group underlay paths as part of the same multipath set even if they are unequal (and inferior) to the best path. This can be done based on the configured difference of AIGP-metric or AS-Path length for those underlay paths when compared to these values for the best-path.

Note
Use the 'match' and 'set' commands to configure the system according to the specified requirements.

The egress load-balance automatic multipath policies can be enabled as below:

Example:

```
switch(config-route-map)# route-map ROUTE_MAP_2
```

a) **set maximum-paths** *max-path-value*

Example:

```
switch(config-route-map)# set maximum-paths 5
```

Configures the maximum number of multipath to be computed and installed for egress load-balancing. The range is 1–64.

AND

b) **set as-path-length difference** *as-path-diff-value*

Example:

```
switch(config-route-map)# set as-path-length difference 5
```

Configures the difference in AS-Path-length compared to the best path that underlay paths must have to be considered for unequal cost load balance. The range is 1–255.

c) **set aigp-metric difference** *value*

Example:

```
switch(config-route-map)# set aigp-metric difference 100
```

Configures the difference in AIGP metric value compared to best path that underlay paths must have to be considered for unequal cost load balance. The range is 1–4294967295.

Note
For more information on how to configure and use AIGP metric information, see Configuration examples for VXLAN EVPN TE - Multi-Site egress load balancing, on page 356.

Step 4 **exit**

Example:

```
switch(config-route-map)# exit
switch(config)#
```

Enter BGP router configuration mode.

Step 5 **router bgp** *as-number*

Example:

```
switch(config)# router bgp 65001
switch(config-router)#
```

Enter BGP router configuration mode.

Step 6 **address-family ipv4 unicast**

Example:

```
switch(config-router)# address-family ipv4 unicast
switch(config-router-af)#
```

Configures the IPv4 unicast address family.

Step 7 **[no] load-balance egress multipath auto-policy route-map** *route-map-name*

Example:

```
switch(config-router-af)# load-balance egress multipath auto-policy route-map ROUTE_MAP_2
```

Configures the parameters to control automatic multipath selection and load-sharing in BGP.

Use the **no** form of this command to remove the parameters configuration.

In the absence of the auto multipath policy only the best-path is installed.

Note
If a community match is used to select the prefix, all paths for this prefix should be tagged with the same community by the advertising egress BGW, which can be done by tagging the prefix during origination into BGP.

Weight Derivation in Underlay

The configuration described in the previous section allowed to add unequal underlay paths to the same multipath set together with the best-paths, to enable equal load balancing of traffic across all these paths (uECMP).

This section describes instead how to statically assign a weight to the underlay paths part of the same multipath set to better distribute overlay traffic flows among them. Two options are available for this: the first consists of assigning a different static weight to the best-path(s) and a different static weight to the set of underlay paths that are unequal (and inferior) to the best-path but still are added to the multipath set based on specific criteria (AS-Path length or AIGP metric). The second option allows instead to explicitly assign a static weight to a specific underlay path.

Load Share Weight Calculation

If the auto-multipath policy contains either or both of the following commands, then BGP will use Load Share Weight Calculation (LSWC) to derive the weight:

- **set load-share multipath-equal-group**
- **set load-share multipath-unequal-group**

Note

- A set of underlay paths that are equivalent to the best path in terms of AS-Path length, is categorized as a **multipath-equal-group**.
- A set of underlay paths that do not match the best path's quality but fall within a specified difference threshold for AS-Path length is categorized as a **multipath-unequal-group**.

In the following example-1, BGP will use LSWC to derive the weight of multiple underlay paths. The unequal underlay paths are added to the **multipath-unequal-group** if their AS-Path length falls withing the configure difference (4) when compared to the best-path:

```
route-map auto-multipath  permit 10
  match ip address prefix-list site_A_BGW2
  set as-path length difference 4
  set maximum-paths 64
  set load-share multipath-equal-group 40
  set load-share multipath-unequal-group 60
```

In the following example-2, BGP the unequal underlay paths are instead added to the **multipath-unequal-group** if their AIGP metric falls withing the configure difference (10) when compared to the best-path AIGP metric:

```
route-map auto-multipath  permit 10
  match ip address prefix-list site_A_BGW2
  set aigp-metric difference 10
  set maximum-paths 64
  set load-share multipath-equal-group 40
  set load-share multipath-unequal-group 60
```

In the preceding example, if only one of the multipath groups is configured (**load-share multipath-equal-group** or **load-share multipath-unequal-group**) the other one will be implicitly assumed as configured by BGP with a value of 1.

Explicit Load Share Weight Calculation

Explicit load share paths requires the definition of route-maps that match a specific next-hop or path associated with a destination IP address. This configuration has the characteristics mentioned below:

- These are specific route-map entries that precede the auto-multipath route-map entry in precedence.
- An auto-multipath rule may or may not be present with explicit load share rules.
- There are a maximum of 64 explicit path entries allowed per destination. This is independent of the maximum path attribute specified in auto-multipath.
- For any explicit load share rule that matches a path for a particular destination, that path is not included or processed by the auto-multipath rule if it exists. This may also be a path that doesn't qualify as ECMP or uECMP.
- With auto-multipath (alone) there is always a path that qualifies for the **multipath-equal-group**, which is the best path. If there is an explicit load share rule that matches the best path for a destination, it's possible for the **multipath-equal-group** to have 0 path. (This would happen if there were no other ECMP paths toward that destination).
- The explicit load share value for a path is proportional to ECMP or uECMP load share values specified in the auto-multipath (if it exists).

- With auto-multipath, a calculation is done to produce individual weights for each of the paths in the ECMP or uECMP groups. This value is a function of the number of paths in each group and the specified load share values.
- If only explicit load-share paths are defined (with no auto-multipath), the values specified in the command will be used as the weights for load sharing.
- If there is no load-share specified for the explicit load share weight policy, there is no default value, and the policy will have no effect.
- The path identified as "best path" is always assigned a weight and downloaded to URIB. This occurs even if there is no explicit load-share policy specified for the best path. If no explicit load-share is specified for the best path, and there is no auto-multipath policy exists, the best path will appear with a weight of 1 in the URIB.

To enable load-balance egress explicit policy, follow these steps:

SUMMARY STEPS

1. **configure terminal**
2. **router bgp** *as-number*
3. **route-map** *routemap-name* **permit** *value*
4. **match ip address prefix-list** *dest-IPaddress*
5. **match ip next-hop prefix-list path** *value*
6. **set load-share** *value*

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# config terminal`
	`switch(config)#`
		Enters global configuration mode.

Step 2	**router bgp** *as-number*
	Example:
	`switch(config)# router bgp 100`
	`switch(config-router)#`
		Enter BGP router configuration mode.

Step 3	**route-map** *routemap-name* **permit** *value* **Example:** `switch(config)# route-map load_distribute permit 6`	Creates a route map or enters route-map configuration mode.
Step 4	**match ip address prefix-list** *dest-IPaddress* **Example:**	The underlay routes are matched using a prefix-list.

	Command or Action	Purpose
	`switch(config-route-map)# match ip address prefix-list match_ip`	
Step 5	**match ip next-hop prefix-list path** *value* **Example:** `switch(config-route-map)# match ip next-hop prefix-list path1`	Explicit load share paths match a specific next-hop or path associated with a destination IP address.
Step 6	**set load-share** *value* **Example:** `switch(config-route-map)# set load-share 199`	Sets the explicit load share paths as specified in the configuration.

Example

If an explicit load-share command, such as **set load-share**, is configured within an auto-multipath policy based on a prefix or next-hop match, it will only take effect if the policy also contains either of one or more LSWC multipath-group commands.

In the following example-1, BGP will use LSWC to derive the weight although **set aigp-metric difference** is defined and, in this case, the explicit load-share will be computed for the matching prefix and next-hop:

```
route-map auto-multipath permit 6
  match ip address prefix-list match_ip
  match ip next-hop prefix-list path1
  set load-share 100

route-map auto-multipath permit 10
   match ip address prefix-list site_A_BGW2
   set as-path-length difference 4
   set aigp-metric difference 100
   set maximum-paths 64
   set load-share multipath-equal-group 40
   set load-share multipath-unequal-group 60
```

In the following example-2, BGP will use AMWC to compute the weight and the explicit load-share will be ignored:

```
route-map auto-multipath permit 6
  match ip address prefix-list match_ip
  match ip next-hop prefix-list path1
  set load-share 100

route-map auto-multipath permit 10
   match ip address prefix-list site_A_BGW2
   set as-path-length difference 4
   set aigp-metric difference 100
   set maximum-paths 64
```

Enabling Egress Load-Balancing for Overlay

For enabling overlay (EVPN) prefixes next-hop resolution through **egress-loadbalance-resolution- vrf** in underlay, follow these steps:

SUMMARY STEPS

1. **configure terminal**
2. **router bgp** *as-number*
3. **address-family l2vpn evpn**
4. **[no] nexthop load-balance egress multisite**

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# config terminal`
	`switch(config)#`
		Enters global configuration mode.

Step 2	**router bgp** *as-number*
	Example:
	`switch(config)# router bgp 100`
	`switch(config-router)#`
		Enter BGP router configuration mode.

Step 3	**address-family l2vpn evpn**
	Example:
	`switch(config-router)# address-family l2vpn evpn`
	`switch((config-router-af)#`
		Configures the L2VPN address family.

Step 4	**[no] nexthop load-balance egress multisite** **Example:** `switch(config-router-af)# nexthop load-balance egress multisite`	Enables overlay (EVPN) next-hop resolution using the corresponding IPv4 or IPv6 table in **egress-loadbalance-resolution- VRF**. For more information on egress-loadbalance-resolution- VRF's table, see VXLAN EVPN TE - Multi-Site egress load balancing configurations, on page 344. Use the **no** form of this command to remove the egress load balancing configuration for overlay. The **multisite** option restricts this functionality to only the EVPN next-hops that are learned from the multi-site network. This applies to both Type-2 and Type-5 routes imported into various VRFs. **Note** This configuration must be enabled after enabling egress-load-balance computation in the underlay table.

Enabling uECMP or wuECMP Load-Balancing for Overlay

You can enable unequal cost or weighted load-balance among multiple overlay next-hops.

Note This configuration does not support VIP next-hops.

Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# config terminal`
	`switch(config)#`
		Enters global configuration mode.

Step 2	**router bgp** *as-number*
	Example:
	`switch(config)# router bgp 100`
	`switch(config-router)#`
		Enter BGP router configuration mode.

Step 3	**address-family l2vpn evpn**
	Example:
	`switch(config-router)# address-family l2vpn evpn`
	`switch((config-router-af)#`
		Configures the L2VPN address family.

Step 4	**nexthop load-balance egress multisite** **Example:** `switch(config-router-af)# nexthop load-balance egress multisite`	Enables overlay (EVPN) next-hop resolution using the corresponding IPv4 or IPv6 table in **egress-loadbalance-resolution- VRF's**. For more information on **egress-loadbalance-resolution- VRF's** table, see VXLAN EVPN TE - Multi-Site egress load balancing configurations, on page 344
Step 5	**[no] maximum-path** *value* **Example:** `switch(config-router-af)# maximum-path 64`	Configures the maximum path as specified for egress load balancing. Use the **no** form of this command to remove the maximum path for egress load balancing. **Note** If the specified value is >1, along with the below command (maximum-paths unequal-cost) then the wuECMP for the overlay next-hops will be enabled automatically deriving the weight based on next-hop metric for those next-hops.
Step 6	**[no] maximum-paths unequal-cost** **Example:** `switch(config-router-af)# maximum-path unequal-cost`	Configures the Unequal Multipath in Overlay. Use the **no** form of this command to disable Unequal Multipath in Overlay **Note** • If the **nexthop load-balance egress multisite** command is configured along with **maximum-path**, and **maximum-path unequal** commands, the overlay

	Command or Action	Purpose
		next-hops will be programmed with weight only if there are multiple overlay next-hops and the **igp_metrics** of those next-hops are different. The overlay next-hops will be resolved using **egress-loadbalance-resolution- VRF** table. • If the **nexthop load-balance egress multisite** command is not configured but **maximum-path**, and **maximum-path unequal** commands are configured, the overlay next-hops will be programmed with weight only if there are multiple overlay next-hops and the **igp_metrics** of those next-hops are different. The overlay next-hops will be resolved using **default** table.

Step 7	**exit**
	Example:
	`switch(config-router-af)# exit`
	`switch(config-router)#`
		Exits command mode.

Step 8	(Optional) **[no] bestpath igp-metric ignore** **Example:** `switch(config-router)# bestpath igp-metric ignore`	This Configuration will cause the igp_metric of the overlay next-hops to be ignored by best path and hence will have ECMP (from wuECMP) in overlay even if **maximum-paths unequal-cost** is configured. Use the **no** form of this command to disable the ECMP/wuECMP.

Verifying VXLAN EVPN TE - Multi-Site Egress Load-Balancing Configuration

To display the VXLAN EVPN TE - Multi-Site Egress Load-Balancing configuration information, enter one of the following commands:

Command	Purpose
show ip|ipv6 route [detail] vrf egress-loadbalance-resolution-	Displays a special internal VRF that is automatically created. This VRF will be implicitly used internally when the egress load-balance configuration is enabled under BGP. When BGP is configured with an ELB filter-policy and an auto-multipath policy, it will inherit the best path for a route from the default table and will include additional ELB paths based on the ELB policy. When the **detail** option is enabled, it will display the weight that BGP sends to the RIB, if wuECMP is configured. **Note** Beginning with Cisco NX-OS Release 10.4(3)F, the table id for the **egress-loadbalance-resolution-** VRF will be statically allocated with a value of 4089/0x0ff9 and will be outside the "limit-resource vrf" pool of allocation. It will not impact any existing user configuration.
show ip|ipv6 route [detail] vrf tenant_vrf	Displays the overlay prefix in an EVPN-VXLAN tenant VRF where the nexthop is resolved via **egress-loadbalance-resolution-** table instead of default table. When the **detail** option is enabled, it will display the weight assigned to the next-hop that is sent from BGP to RIB, if wuECMP is configured.
show bgp ipv4 unicast *ipaddress* **vrf egress-loadbalance-resolution-**	Displays the underlay BGP routes and next-hops, including the derived AIGP metric if AIGP is configured in the underlay. For wuECMP cases, it shows the weight, which is derived dynamically (i.e., from the AIGP metric or from a statically configured load-share-weight). In the case of uECMP or ECMP, there will be no weight displayed.
show l2route evpn mac all detail	Displays the next-hops with weight for the mac routes, if wuECMP is configured.
show l2route evpn ead es detail	Displays the weight associated with the next-hops for the EAD/ES route if wuECMP is configured.

Configuration examples for VXLAN EVPN TE - Multi-Site egress load balancing

This section explains how to configure and verify the use cases of the VXLAN EVPN TE Multi-Site egress load balancing feature.

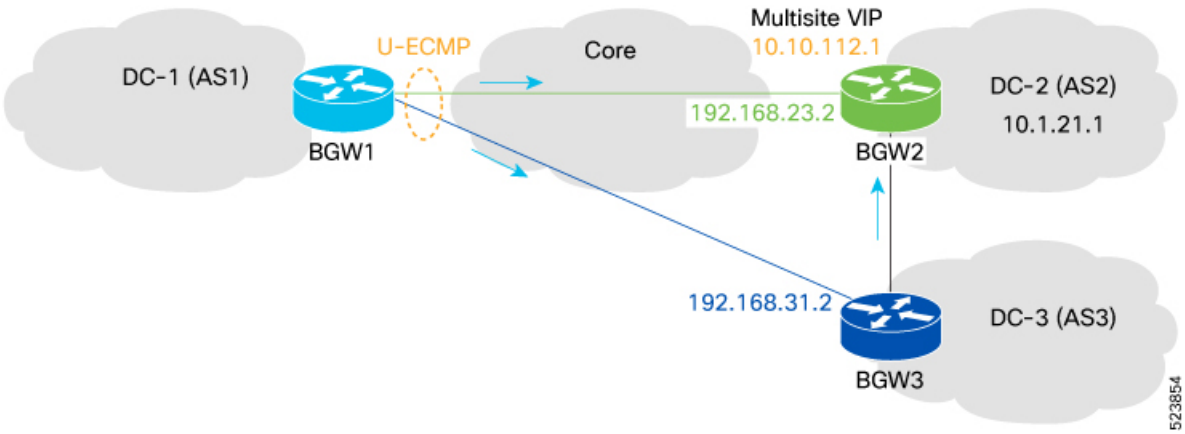
- uECMP in the underlay and single EVPN next-hop for overlay prefixes.
- Static wuECMP with load-share, explicit load-share in the underlay, and single EVPN next-hop for overlay prefixes.
- Dynamic weight (wuECMP) in overlay and underlay and with AIGP in underlay.

uECMP in Underlay, single EVPN next-hop for Overlay prefixes

As highlighted in the figure below, in this use case all the overlay prefixes advertised from DC-2 toward DC-1 have a single EVPN next-hop represented by the Multi-Site VIP address shared by all the BGW devices in DC-2 (only one device, BGW2, is shown for simplicity in the figure).

Underlay reachability from the BGW1 device in DC-1 to the Multi-Site VIP in DC-2 is possible using the following two underlay paths:

1. Green path: The best path connecting BGW1 to BGW2
2. Blue path: The less favorable path going through BGW3 in DC-3 (alternatively, the suboptimal blue path could go through one or more router devices part of the Core infrastructure).



The goal in this use case is to ensure that both the green and blue underlay paths can be equally used for any overlay communication between DC-1 and DC2. For this to be possible, the two paths must be considered as uECMP paths part of a same multipath set and the configuration steps below applied to the BGW1 device in DC-1 achieve this purpose.

Enable VXLAN EVPN TE - Multi-Site Egress Load-Balancing uECMP in Underlay

All the commands below must be configured on the BGW1 device in DC-1.

1. Create a Filter-policy.

- Specify the underlay routes to enable ELB uECMP. In this case, the underlay route is the Multi-Site VIP address in DC-2 representing the next-hop for the overlay prefixes advertised to DC-1.

```
ip prefix-list site2_ms_vip seq 5 permit 10.10.112.1/32
```

- Create a route-map and apply the previously configured prefix-list in the match condition.

```
route-map Filter-Policy permit 10
  match ip address prefix-list site2_ms_vip
```

2. Enable ELB Filter-policy (under the IPv4 or IPv4 BGP address-family).

```
router bgp 1
  address-family ipv4 unicast
    load-balance egress filter-policy route-map Filter-Policy
```

3. Verify the installation of DC-2 Multi-Site VIP route in the **Egress-loadbalance-resolution-** VRF table. At this point, only the green best-path is considered to reach the destination.

```
BGW1# show ip route 10.10.112.1 vrf egress-loadbalance-resolution-

10.10.112.1/32, ubest/mbest: 1/0
  *via 192.168.23.2%default, [20/0], 16:36:15, bgp-1, external, tag 2, uecmp ! Green
path
```

4. Create a Multipath Auto-policy.

- Provide the criteria to assign unequal underlay paths to the multipath set together with the best path. Differences in BGP attributes, such as AS-Path length, can be considered to group unequal underlay paths as part of the same multipath set.
- Specify the maximum number of underlay paths part of the multipath set to be computed and installed for 'Egress Load Balancing'.

```
route-map Auto-Policy permit 10
  set as-path-length difference 1 <1 to 255>
  set maximum-paths 2 <1 to 64>
```

5. Enable the ELB Multipath Auto-policy under the IPv4 or IPv6 BGP address-family.

```
router bgp 1
  address-family ipv4 unicast
    load-balance egress multipath auto-policy route-map Auto-Policy
```

6. Verify routes in Egress-loadbalance-resolution VRF table of URIB. The blue suboptimal path has been added to the green best-path as a viable option to reach the Multi-Site VIP address in DC-2.

```
BGW1# show ip route 10.10.112.1 vrf egress-loadbalance-resolution-

10.10.112.1/32, ubest/mbest: 2/0
  *via 192.168.23.2%default, [20/0], 00:44:17, bgp-1, external, tag 2, uecmp ! Green
path
  *via 192.168.31.2%default, [20/0], 00:44:17, bgp-1, external, tag 3, uecmp ! Blue path
```

Enabling Resolution in the ELB VRF for uECMP in Overlay

1. Enable Overlay Next-hop Resolution using the Egress Load Balancing uECMP Underlay Paths.

- This command enables resolution of EVPN next-hops using the **Egress-loadbalance-resolution-** VRF table. Therefore, both underlay paths (the green and the unequal blue one) would be used for equal cost load-balancing of intersite traffic.

- This command must be enabled under the BGP L2VPN EVPN address-family after enabling the Egress Load Balancing computation in the underlay table.

```
router bgp 1
  address-family l2vpn evpn
    nexthop load-balance egress multisite
```

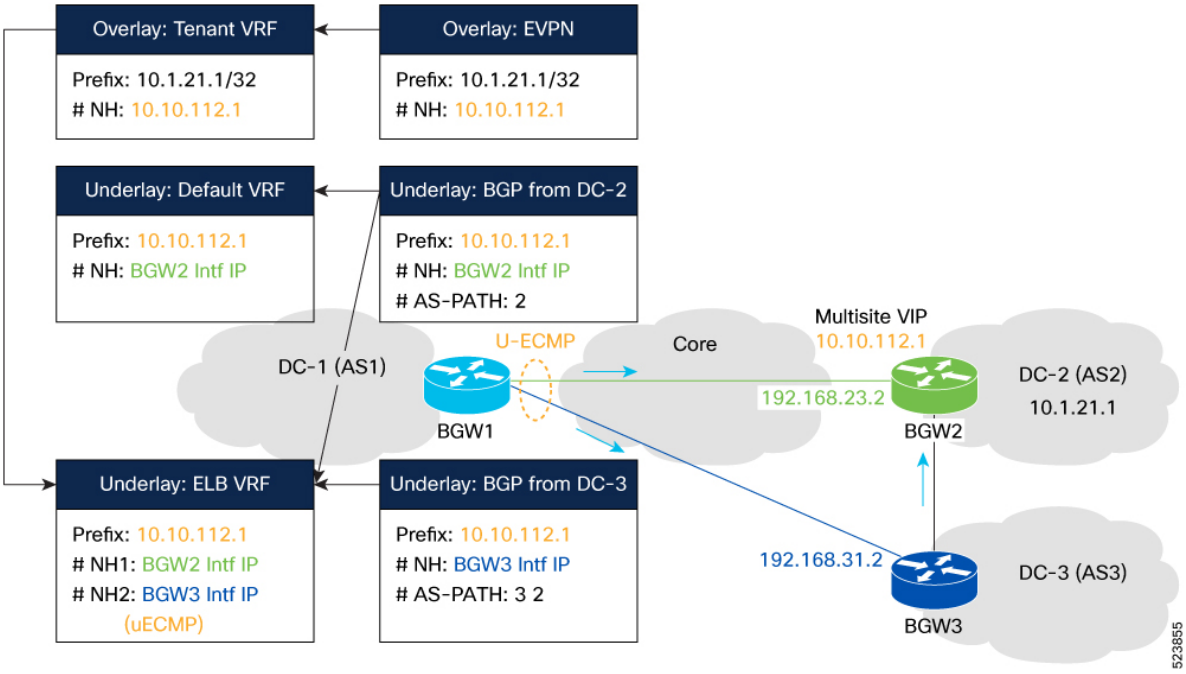
2. Verify overlay routes in Tenant VRF table. As shown below, reachability to an overlay prefix in DC-2 from the BGW1 device in DC-1 is through the DC-2 Multi-Site VIP address (10.10.112.1) and the lookup for that is to be performed in the **egress-loadbalance-resolution-** VRF. As a result, traffic will be distributed evenly across the two installed underlay paths. (as shown above).

```
BGW1# show ip route 10.1.21.1 vrf 3001

10.1.21.1/32, ubest/mbest: 1/0
  *via 10.10.112.1%egress-loadbalance-resolution-, [20/2000], 04:43:40, bgp-1, external,
 tag 2,
  eLB, segid: 3003001 tunnelid: 0x67027001 encap: VXLAN
```

uECMP Routing details

This section provides more detailed troubleshooting information for the uECMP in Underlay, single EVPN next-hop for Overlay prefixes use case.



1. **Underlay Route Programming**

   The DC-2 Multi-Site VIP routereceived from remote BGWs via uECMP underlay paths is programmed on the BGW1 device in DC-1 into the BGP routing table and unicast routing table for the egress-loadbalance-resolution- VRF.

   The following example shows the sample output for the following components:

- BGP - uECMP: Note how the second path is labeled as **multipath uecmp**, since the AS-Path difference is 1, matching the previously defined criteria to considered underlay paths part of the same multi-path set.

```
BGW1# show bgp ipv4 unicast 10.10.112.1 vrf egress-loadbalance-resolution-
<Truncated>
  Advertised path-id 1
  Path type: external, path is valid, is best path, no labeled nexthop, in rib
             Imported from 10.10.112.1/32 (VRF default)
  AS-Path: 2 , path sourced external to AS
    192.168.23.2 (metric 0) from 192.168.23.2 (101.2.33.1)
      Origin incomplete, MED 0, localpref 100, weight 0

  Path type: external, path is valid, not best reason: newer EBGP path, multipath,
uecmp, no labeled nexthop, in rib
             Imported from 10.10.112.1/32 (VRF default)
  AS-Path: 3 2 , path sourced external to AS
    192.168.31.2 (metric 0) from 192.168.31.2 (101.3.33.1)
      Origin incomplete, MED not set, localpref 100, weight 0
```

- URIB - uECMP: The following example shows how the two uECMP paths equally leveraged to reach DC-2 Multi-Site VIP destination.

```
BGW1# show ip route 10.10.112.1 vrf egress-loadbalance-resolution-
<Truncated>
10.10.112.1/32, ubest/mbest: 2/0
    *via 192.168.23.2%default, [20/0], 17:58:44, bgp-1, external, tag 2, uecmp !
Green path
    *via 192.168.31.2%default, [20/0], 17:58:44, bgp-1, external, tag 3, uecmp !
Green path
```

- FIB - uECMP:

```
BGW1# show forwarding route 10.10.112.1 vrf egress-loadbalance-resolution-
<Truncated>
-----------------+--------------------+---------------+----------------+---------------
Prefix           | Next-hop           | Interface     | Labels         | Partial
 Install
-----------------+--------------------+---------------+----------------+---------------
*10.10.112.1/32      192.168.23.2           Ethernet1/6
                     192.168.31.2           Ethernet1/7
```

2. **Overlay Route Programming**

   The **egress-loadbalance-resolution- VRF** will be used to resolve the remote VIP next-hops associated to the received overlay routes.

   The following example shows the sample output for following components:

   - BGP:

```
BGW1# show bgp l2vpn evpn 10.1.21.1
<Truncated>
  Advertised path-id 1
  Path type: external, path is valid, is best path, no labeled nexthop, has esi_gw
             Imported to 3 destination(s)
             Imported paths list: 3001 L3-3003001 L2-2001001
  AS-Path: 2 , path sourced external to AS
    10.10.112.1 (metric 0) from 101.2.33.1 (101.2.33.1)
      Origin IGP, MED 2000, localpref 100, weight 0
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 SOO:102.2.121.1:0
```

   - URIB:

```
BGW1# show ip route 10.1.21.1 vrf 3001
<Truncated>
10.1.21.1/32, ubest/mbest: 1/0
    *via 10.10.112.1%egress-loadbalance-resolution-, [20/2000], 00:28:03, bgp-1,
external,
    tag 2, eLB, segid: 3003001 tunnelid: 0x67027001 encap: VXLAN
```

- FIB:

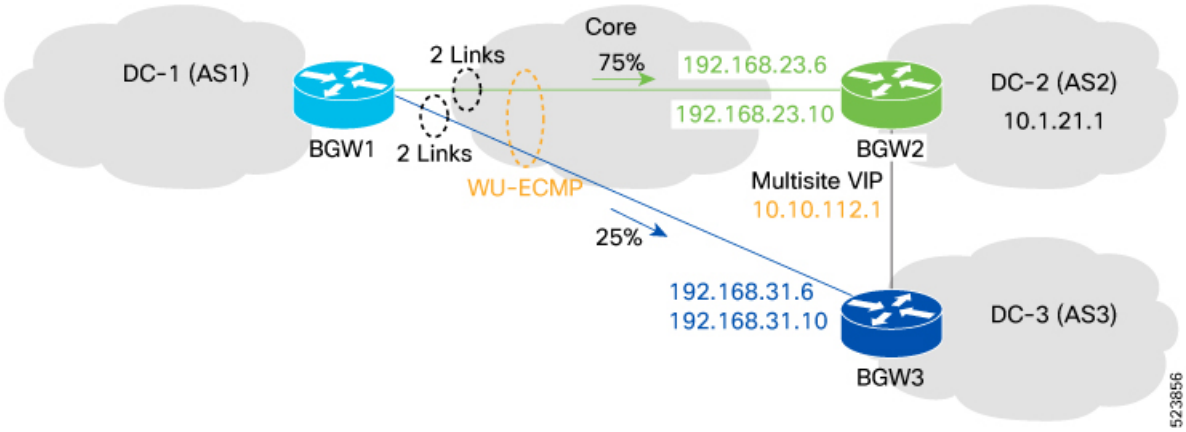
```
BGW1# show forwarding route 10.1.21.1 vrf 3001
<Truncated>
------------------+--------------------+---------------+----------------+----------------
Prefix            | Next-hop           | Interface     | Labels         | Partial
 Install
------------------+--------------------+---------------+----------------+----------------
10.1.21.1/32         10.10.112.1           nve1

BGW1# show forwarding route 10.10.112.1 vrf egress-loadbalance-resolution-
<Truncated>
------------------+--------------------+---------------+----------------+----------------
Prefix            | Next-hop           | Interface     | Labels         | Partial
 Install
------------------+--------------------+---------------+----------------+----------------
*10.10.112.1/32      192.168.23.2           Ethernet1/6
                     192.168.31.2           Ethernet1/7
```

Static wuECMP with Load-Share and Explicit load-share in Underlay, and single EVPN next-hop for Overlay Prefixes

Enable VXLAN EVPN TE - Multi-Site Egress Load-Balancing Static wuECMP with Load-Share in Underlay

As highlighted in the figure below, also in this use case all the overlay prefixes advertised from DC-2 toward DC-1 have a single EVPN next-hop represented by the Multi-Site VIP address shared by all the BGW devices in DC-2. Underlay reachability from the BGW1 device in DC-1 to the Multi-Site VIP in DC-2 is possible via four underlay paths: two green ECMP best-paths connecting BGW1 to BGW2 and two less favorable underlay paths going through BGW3 in DC-3 (alternatively, the suboptimal blue paths could go through one or more router devices part of the Core infrastructure).



The goal in this use case is to ensure that both green and blue underlay paths can be used for overlay communication between DC-1 and DC2, but with a different traffic load-share (75% of the traffic should use

the green paths and only 25% the blue paths). The configuration steps below applied to the BGW1 device in DC-1 achieve this purpose.

1. Create a Filter-policy.
   - Specify the underlay routes to enable ELB wuECMP. In this case, the underlay route is the Multi-Site VIP address in DC-2 representing the next-hop for the overlay prefixes advertised to DC-1.

     ```
     ip prefix-list site2_ms_vip seq 5 permit 10.10.112.1/32
     ```

   - Create a route-map and apply the previously configured prefix-list in the match condition.

     ```
     route-map Filter-Policy permit 10
       match ip address prefix-list site2_ms_vip
     ```

2. Enable ELB Filter-policy (under the IPv4 or IPv4 BGP address-family).

   ```
   router bgp 1
     address-family ipv4 unicast
       load-balance egress filter-policy route-map Filter-Policy
   ```

3. Verify routes in **Egress-loadbalance-resolution**- VRF table. By default, only the two green ECMP underlay paths are installed to reach the Multi-Site VIP address in DC-2.

   ```
   BGW1# show ip route 10.10.112.1 vrf egress-loadbalance-resolution-

   10.10.112.1/32, ubest/mbest: 1/0
     *via 192.168.23.6%default, [20/0], 16:36:15, bgp-1, external, tag 2, uecmp ! Green
   Path
     *via 192.168.23.10%default, [20/0], 16:37:41, bgp-1, external, tag 2, uecmp ! Green
   Path
   ```

4. Create a Multipath Auto-policy. For more information, see Load Share Weight Calculation, on page 348.
   - Provide differences in BGP attributes such as AS-Path length that can be considered to group underlay paths as part of the same multipath set.
   - Specify the maximum number of multipaths to be computed and installed for ‘Egress Load Balance’. In this case, the value is 4 (two green and two blue paths).
   - Specify the Load-Share Weight to be associated to the ECMP Path-set and uECMP Path-set of underlay paths. The goal, as shown in the figure above, is to send 75% of the traffic on the green best path underlay links and 25% on the blue unequal underlay links.

     ```
     route-map Auto-Policy permit 10
       set as-path-length difference 1 <1 to 255>
       set maximum-paths 4 <1 to 64>
       set load-share multipath-equal-group 3 <1 to 255>
       set load-share multipath-unequal-group 1 <1 to 255>
     ```

5. Enable ELB Multipath Auto-policy under the IPv4 or IPv6 BGP address-family.

   ```
   router bgp 1
     address-family ipv4 unicast
       load-balance egress multipath auto-policy route-map Auto-policy
   ```

6. Verify wuECMP with load-share routes in **Egress-loadbalance-resolution** VRF table of URIB. The blue suboptimal paths have been added to the green best-paths as viable options to reach the Multi-Site VIP address in DC-2, but with a different weight (1).

   ```
   BGW1# show ip route 10.10.112.1 detail vrf egress-loadbalance-resolution-

   10.10.112.1/32, ubest/mbest: 4/0
   ```

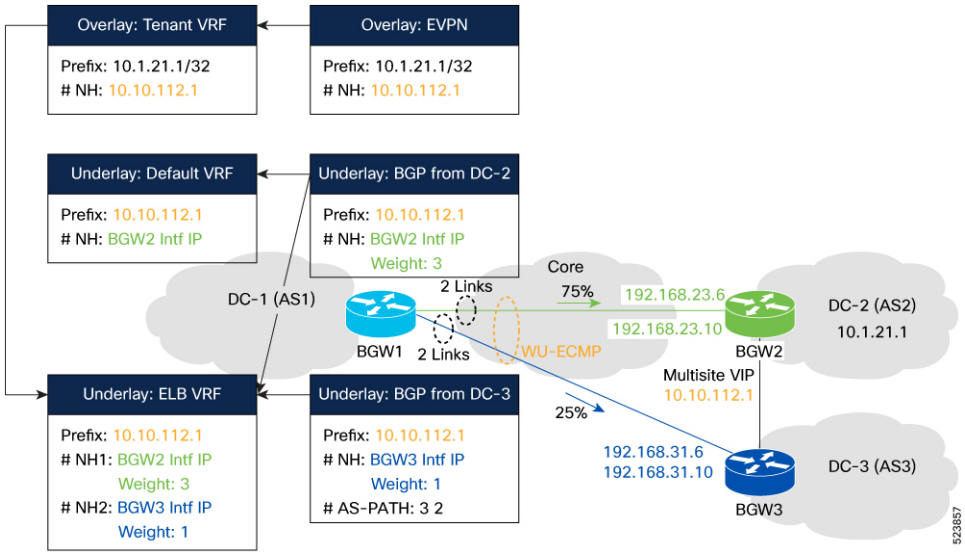
```
    *via 192.168.23.6%default, [20/0], 1w0d, bgp-1, weight:3, external, tag 2, uecmp !
Green Path
    *via 192.168.23.10%default, [20/0], 1w0d, bgp-1, weight:3, external, tag 2, uecmp !
 Green Path
    *via 192.168.31.6%default, [20/0], 1w0d, bgp-1, weight:1, external, tag 3, uecmp !
Blue Path
    *via 192.168.31.10%default, [20/0], 1w0d, bgp-1, weight:1, external, tag 3, uecmp !
 Blue Path
```

Because of the automatic load-share configuration applied to ECMP and uECMP paths, the two green path has been assigned a total weight of 3 (1.5 each), whereas the two blue paths has a total weight of 1 (0.5 each). Consequently, 75% of the overlay traffic from DC-1 to DC-2 is routed through the green paths, while the remaining 25% traverses the blue paths:

Green paths: (3 + 3) / (3+3+1+1) = 0.75

Blue paths: (1 + 1) / (3+3+0.5+0.5) = 0.25

Static wuECMP with Load-Share Routing details



1. **Underlay Route Programming**

   The DC-2 Multi-Site VIP route received from remote BGWs via uECMP underlay paths is programmed on the BGW1 device in DC-1 into the BGP routing table and unicast routing table for the **egress-loadbalance-resolution-** VRF.

   The following example shows the sample output for following components:

   - BGP: Note how the second path is labeled as "**multipath uecmp**" even if it is actually an ECMP path with the first best-path one. The other two paths are proper "**multipath uecmp**", since the AS-Path difference is 1, matching the previously defined criteria to considered underlay paths part of the same multi-path set. Additionally, a 3:1 weight ratio is applied to the two sets of paths because of the load-share configuration.

     ```
     BGW1# show bgp ipv4 unicast 10.10.112.1 vrf egress-loadbalance-resolution-
     <Truncated>
       Advertised path-id 1
       Path type: external, path is valid, is best path, no labeled nexthop, in rib
                  Imported from 10.10.112.1/32 (VRF default)
     ```

```
  AS-Path: 2 , path sourced external to AS
    192.168.23.6 (metric 0) from 192.168.23.6 (101.2.33.1)
      Origin incomplete, MED 0, localpref 100, weight 0, load share weight 3

  Path type: external, path is valid, not best reason: newer EBGP path, multipath,
no labeled nexthop, in rib
             Imported from 10.10.112.1/32 (VRF default)
  AS-Path: 2 , path sourced external to AS
    192.168.23.10 (metric 0) from 192.168.23.10 (101.2.33.1)
      Origin incomplete, MED 0, localpref 100, weight 0, load share weight 3
  Path type: external, path is valid, not best reason: newer EBGP path, multipath,
uecmp, no labeled nexthop, in rib
             Imported from 10.10.112.1/32 (VRF default)
  AS-Path: 3 2 , path sourced external to AS
    192.168.31.6 (metric 0) from 192.168.31.6 (101.3.33.1)
     Origin incomplete, MED not set, localpref 100, weight 0, load share weight 1

  Path type: external, path is valid, not best reason: newer EBGP path, multipath,
uecmp, no labeled nexthop, in rib
             Imported from 10.10.112.1/32 (VRF default)
  AS-Path: 3 2 , path sourced external to AS
    192.168.31.10 (metric 0) from 192.168.31.10 (101.3.33.1)
      Origin incomplete, MED not set, localpref 100, weight 0, load share weight 1
```

- URIB:

```
BGW1# show ip route 10.10.112.1 detail vrf egress-loadbalance-resolution-
<Truncated>
10.10.112.1/32, ubest/mbest: 4/0
    *via 192.168.23.6%default, [20/0], 1w0d, bgp-1, weight:3, external, tag 2, uecmp
 ! Green Path
<Truncated>
    *via 192.168.23.10%default, [20/0], 1w0d, bgp-1, weight:3, external, tag 2, uecmp
 ! Green Path
<Truncated>
    *via 192.168.31.6%default, [20/0], 1w0d, bgp-1, weight:1, external, tag 3, uecmp
 ! Blue Path
<Truncated>
    *via 192.168.31.10%default, [20/0], 1w0d, bgp-1, weight:1, external, tag 3, uecmp
 ! Blue Path
<Truncated>
```

- FIB:

```
BGW1# show forwarding route 10.10.112.1 vrf egress-loadbalance-resolution-
<Truncated>
-----------------+--------------------+---------------+----------------+---------------
Prefix           | Next-hop           | Interface     | Labels         | Partial
 Install
-----------------+--------------------+---------------+----------------+---------------
*10.10.112.1/32      192.168.23.6           Ethernet1/22 >> 24 Entries
                     192.168.23.6           Ethernet1/22
                     .....
                     192.168.23.10          Ethernet1/23 >> 24 Entries
                     192.168.23.10          Ethernet1/23
                     .....
                     192.168.31.6           Ethernet1/32 >> 8 Entries
                     192.168.31.6           Ethernet1/33
                     .....
                     192.168.31.10          Ethernet1/32 >> 8 Entries
                     192.168.31.10          Ethernet1/33
                     .....
```

The following summarizes the FIB forwarding entries created based on weight of ECMP (3) and uECMP (1) path sets:

- ECMP Path-Set: 48 Entries
- uECMP Path-Set: 16 Entries
- Traffic Ratio is 48 : 16 = 3 : 1

2. Overlay Route Programming

The **egress-loadbalance-resolution-VRF** table is used to resolve the VIP next-hop for the Overlay prefixes advertised from DC-2 to DC-1.

The following example shows the sample output for following components:

- BGP:

```
BGW1# show bgp l2vpn evpn 10.1.21.1
<Truncated>
  Advertised path-id 1
  Path type: external, path is valid, is best path, no labeled nexthop, has esi_gw
             Imported to 3 destination(s)
             Imported paths list: 3001 L3-3003001 L2-2001001
  AS-Path: 2 , path sourced external to AS
    10.10.112.1 (metric 0) from 101.2.33.1 (101.2.33.1)
      Origin IGP, MED 2000, localpref 100, weight 0
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 SOO:102.2.121.1:0
```

- URIB:

```
BGW1# show ip route 10.1.21.1 vrf 3001
<Truncated>
10.1.21.1/32, ubest/mbest: 1/0
    *via 10.10.112.1%egress-loadbalance-resolution-, [20/2000], 00:28:03, bgp-1,
external,
    tag 2, eLB, segid: 3003001 tunnelid: 0x67027001 encap: VXLAN
```

- FIB:

```
BGW1# show forwarding route 10.1.21.1 vrf 3001
<Truncated>
-----------------+--------------------+---------------+----------------+---------------
Prefix           | Next-hop           | Interface     | Labels         | Partial
 Install
-----------------+--------------------+---------------+----------------+---------------
10.1.21.1/32         10.10.112.1           nve1

BGW1# show forwarding route 10.10.112.1 vrf egress-loadbalance-resolution-
<Truncated>
-----------------+--------------------+---------------+----------------+---------------
Prefix           | Next-hop           | Interface     | Labels         | Partial
 Install
-----------------+--------------------+---------------+----------------+---------------
*10.10.112.1/32      192.168.23.6           Ethernet1/22 >> 24 Entries
                     192.168.23.6           Ethernet1/22
                     .....
                     192.168.23.10          Ethernet1/23 >> 24 Entries
                     192.168.23.10          Ethernet1/23
                     .....
                     192.168.31.6           Ethernet1/32 >> 8 Entries
                     192.168.31.6           Ethernet1/33
                     .....
```

```
192.168.31.10           Ethernet1/32 >> 8 Entries
192.168.31.10           Ethernet1/33
.....
```

The following summarizes the FIB forwarding entries created based on weight of ECMP (3) and uECMP (1) path sets:

- ECMP Path-Set: 48 Entries
- uECMP Path-Set: 16 Entries
- Traffic Ratio is 48 : 16 = 3 : 1

Enabling Resolution in the ELB VRF for Static wuECMP with Load-Share in Overlay

For both the use cases described above, it is then required to ensure that the resolution for the Multi-Site VIP next-hop address is done in the **Egress-loadbalance-resolution-** VRF table.

1. Enable Overlay Next-hop Resolution using Egress Load Balancing uECMP Underlay Paths.
   - This command enables resolution of EVPN next-hops using the **Egress-loadbalance-resolution-** VRF table.
   - This command must be enabled after enabling the Egress Load Balancing computation in the underlay table.

   ```
   router bgp 1
     address-family l2vpn evpn
       nexthop load-balance egress multisite
   ```

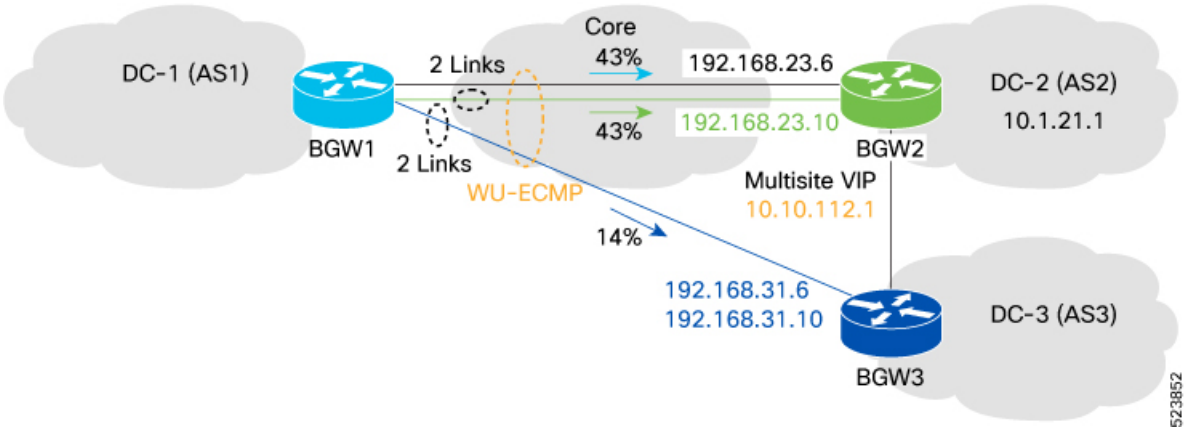
2. Verify overlay routes in Tenant VRF table.

   ```
   BGW1# show ip route 10.1.21.1 vrf 3001

   10.1.21.1/32, ubest/mbest: 1/0
     *via 10.10.112.1%egress-loadbalance-resolution-, [20/2000], 04:43:40, bgp-1, external,
    tag 2,
     eLB, segid: 3003001 tunnelid: 0x67027001 encap: VXLAN
   ```

Enable VXLAN EVPN TE - Multi-Site Egress Load-Balancing Static wuECMP with Explicit Load-Share in Underlay

The figure below shows a slight modification to the previously mentioned use case. In this use case, an explicit load-share configuration can be applied to one of the two green paths. By doing this, the particular green path is excluded from the **multipath-equal-group** set, resulting in a change to the overall distribution of traffic load among the green paths.



The configuration steps below are applied to BGW1 and are very similar to the ones described in the previous use case, with the only addition of the explicit load-share configuration.

1. Create a Filter-policy.
   - Specify the underlay routes to enable ELB wuECMP. Also in this case, the underlay route is the Multi-Site VIP address in DC-2 representing the next-hop for the overlay prefixes advertised to DC-1.

     ```
     ip prefix-list site2_ms_vip seq 5 permit 10.10.112.1/32
     ```

   - Create a route-map and apply the previously configured prefix-list in the match condition.

     ```
     route-map Filter-Policy permit 10
       match ip address prefix-list site2_ms_vip
     ```

2. Enable ELB Filter-policy (under the IPv4 or IPv4 BGP address-family).

   ```
   router bgp 1
     address-family ipv4 unicast
       load-balance egress filter-policy route-map Filter-Policy
   ```

3. Verify routes in **Egress-loadbalance-resolution-** VRF table.

   ```
   BGW1# show ip route 10.10.112.1 vrf egress-loadbalance-resolution-

   10.10.112.1/32, ubest/mbest: 2/0
     *via 192.168.23.6%default, [20/0], 16:36:15, bgp-1, external, tag 2, uecmp ! Green
   Path
     *via 192.168.23.10%default, [20/0], 16:37:41, bgp-1, external, tag 2, uecmp ! Green
   Path
   ```

4. Create a Multipath Auto-policy. For more information, see Load Share Weight Calculation, on page 348.
   - Provide differences in BGP attributes such as AS-Path length that can be considered when choosing multipath.
   - Specify the maximum number of multipaths to be computed and installed for ‘Egress Load Balance’.
   - For **Explicit Load share**, the load-share weight can be assigned to the explicit path that matches a specific next-hop associated with overlay next-hop (in this specific example is the green path with next-hop 192.168.23.6).

     ```
     ip prefix-list MS_NH_S2_1 seq 5 permit 192.168.23.6/32

     route-map Auto-Policy permit 5
     ```

```
  match ip address prefix-list site2_ms_vip
  match ip next-hop prefix-list MS_NH_S2_1
  set load-share 3 <1 to 255>

route-map Auto-Policy permit 10
   set as-path-length difference 1 <1 to 255>
   set maximum-paths 4 <1 to 64>
   set load-share multipath-equal-group 3 <1 to 255>
   set load-share multipath-unequal-group 1 <1 to 255>
```

5. Enable ELB Multipath Auto-policy.

```
router bgp 1
  address-family ipv4 unicast
    load-balance egress multipath auto-policy route-map Auto-policy
```

6. Verify wuECMP with explicit load-share routes in **Egress-loadbalance-resolution**VRF table of URIB. The blue suboptimal paths have been added to the green best-paths as a viable options to reach the Multi-Site VIP address in DC-2.

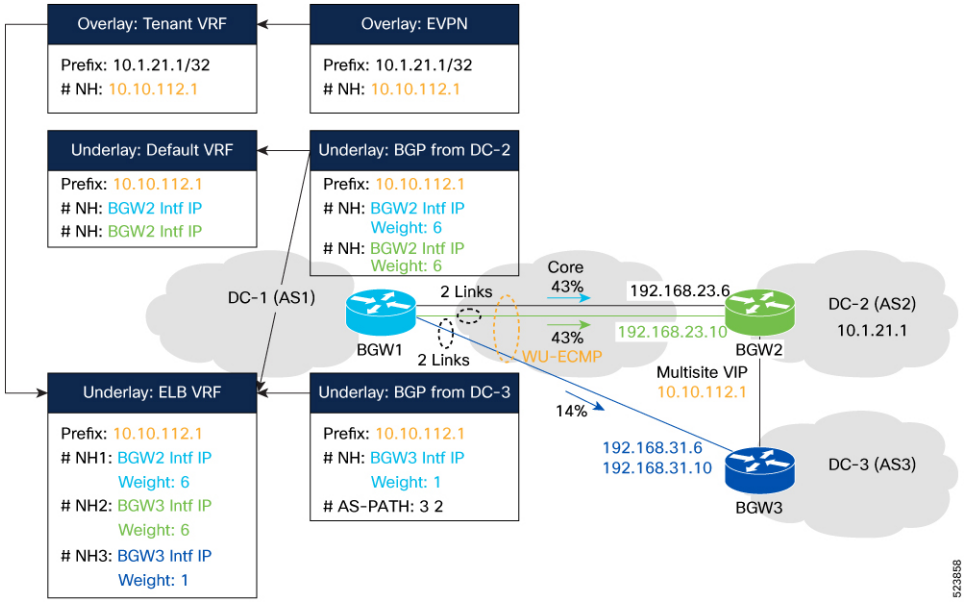
```
BGW1# show ip route 10.10.112.1 detail vrf egress-loadbalance-resolution-
<Truncated>
10.10.112.1/32, ubest/mbest: 4/0
    *via 192.168.23.6%default, [20/0], 1w0d, bgp-1, weight:6, external, tag 2, uecmp !
Green Path
<Truncated>
    *via 192.168.23.10%default, [20/0], 1w0d, bgp-1, weight:6, external, tag 2, uecmp !
 Green Path
<Truncated>
    *via 192.168.31.6%default, [20/0], 1w0d, bgp-1, weight:1, external, tag 3, uecmp !
Blue Path
<Truncated>
    *via 192.168.31.10%default, [20/0], 1w0d, bgp-1, weight:1, external, tag 3, uecmp !
 Blue Path
<Truncated>
```

When comparing the output above with the one shown in the previous automatic load-share use case, the main difference is that now, because of the explicit load-share configuration, one of the green link has been taken out of the multipath-equal-group and hence got individually assigned a weight of 3. The other green link, only one remaining in the multipath-equal-group, also got assigned an individual weight of 3, whereas the two blue links continue to get assigned a total weight of 1. As a result, 86% of traffic is now using the green links (43% for each green link) and 14% is using the blue links (7% each):

Green paths: (6 + 6) / (6+6+1+1) = 0.86

Blue paths: (1 + 1) / (6+6+1+1) = 0.14

Static wuECMP with Explicit Load-Share Routing details



1. **Underlay Route Programming**

   Underlay route with multipath received from remote BGWs is programmed to BGP routing table and Unicast routing table for egress-loadbalance-resolution- VRF.

   The following example shows the sample output for following components:

   - BGP:

```
BGW1# show bgp ipv4 unicast 10.10.112.1 vrf egress-loadbalance-resolution-
<Truncated>
  Advertised path-id 1
  Path type: external, path is valid, is best path, no labeled nexthop, in rib
             Imported from 10.10.112.1/32 (VRF default)
  AS-Path: 2 , path sourced external to AS
    192.168.23.6 (metric 0) from 192.168.23.6 (101.2.33.1)
      Origin incomplete, MED 0, localpref 100, weight 0, explicit  weight 6

  Path type: external, path is valid, not best reason: newer EBGP path, multipath,
no labeled nexthop, in rib
             Imported from 10.10.112.1/32 (VRF default)
  AS-Path: 2 , path sourced external to AS
    192.168.23.10 (metric 0) from 192.168.23.10 (101.2.33.1)
      Origin incomplete, MED 0, localpref 100, weight 0, load share weight 6

  Path type: external, path is valid, not best reason: newer EBGP path, multipath,
uecmp, no labeled nexthop, in rib
             Imported from 10.10.112.1/32 (VRF default)
  AS-Path: 3 2 , path sourced external to AS
    192.168.31.6 (metric 0) from 192.168.31.6 (101.3.33.1)
      Origin incomplete, MED not set, localpref 100, weight 0, load share weight 1

  Path type: external, path is valid, not best reason: newer EBGP path, multipath,
uecmp, no labeled nexthop, in rib
             Imported from 10.10.112.1/32 (VRF default)
  AS-Path: 3 2 , path sourced external to AS
    192.168.31.10 (metric 0) from 192.168.31.10 (101.3.33.1)
      Origin incomplete, MED not set, localpref 100, weight 0, load share weight 1
```

- URIB:

```
BGW1# show ip route 10.10.112.1 detail vrf egress-loadbalance-resolution-
<Truncated>
10.10.112.1/32, ubest/mbest: 4/0
    *via 192.168.23.6%default, [20/0], 1w0d, bgp-1, weight:6, external, tag 2, uecmp
 ! Blue Path
<Truncated>
    *via 192.168.23.10%default, [20/0], 1w0d, bgp-1, weight:6, external, tag 2, uecmp
 ! Green Path
<Truncated>
    *via 192.168.31.6%default, [20/0], 1w0d, bgp-1, weight:1, external, tag 3, uecmp
 ! Blue Path
<Truncated>
    *via 192.168.31.10%default, [20/0], 1w0d, bgp-1, weight:1, external, tag 3, uecmp
 ! Blue Path
<Truncated>
```

- FIB:

```
BGW1# show forwarding route 10.10.112.1 vrf egress-loadbalance-resolution-
<Truncated>
-----------------+--------------------+---------------+----------------+----------------
Prefix           | Next-hop           | Interface     | Labels         | Partial
 Install
-----------------+--------------------+---------------+----------------+----------------
*10.10.112.1/32      192.168.23.6           Ethernet1/22 >> 27 Entries
                     192.168.23.6           Ethernet1/22
                     .....
                     192.168.23.10          Ethernet1/23 >> 27 Entries
                     192.168.23.10          Ethernet1/23
                     .....
                     192.168.31.6           Ethernet1/32 >> 5 Entries
                     192.168.31.6           Ethernet1/33
                     .....
                     192.168.31.10          Ethernet1/32 >> 5 Entries
                     192.168.31.10          Ethernet1/33
                     .....
```

The following summarizes the FIB forwarding entries created based on weight of explicit (3), ECMP (3), and uECMP (1) path sets:

- Explicit Path: 27 Entries
- ECMP Path-Set: 27 Entries
- uECMP Path-Set: 10 Entries
- Traffic Ratio is 27:27:10 = 3:3:1

2. Overlay Route Programming

Overlay routes will use **egress-loadbalance-resolution-** VRF table to resolve the VIP/PIP next-hops for the fabric external path.

The following example shows the sample output for following components:

- BGP:

```
BGW1# show bgp l2vpn evpn 10.1.21.1
<Truncated>
  Advertised path-id 1
  Path type: external, path is valid, is best path, no labeled nexthop, has esi_gw
```

```
            Imported to 3 destination(s)
            Imported paths list: 3001 L3-3003001 L2-2001001
  AS-Path: 2 , path sourced external to AS
    10.10.112.1 (metric 0) from 101.2.33.1 (101.2.33.1)
      Origin IGP, MED 2000, localpref 100, weight 0
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 SOO:102.2.121.1:0
```

- URIB:

```
BGW1# show ip route 10.1.21.1 vrf 3001
<Truncated>
10.1.21.1/32, ubest/mbest: 1/0
    *via 10.10.112.1%egress-loadbalance-resolution-, [20/2000], 00:28:03, bgp-1,
external,
    tag 2, eLB, segid: 3003001 tunnelid: 0x67027001 encap: VXLAN
```

- FIB:

```
BGW1# show forwarding route 10.1.21.1 vrf 3001
<Truncated>
------------------+--------------------+---------------+----------------+---------------
Prefix            | Next-hop           | Interface     | Labels         | Partial
 Install
------------------+--------------------+---------------+----------------+---------------
10.1.21.1/32         10.10.112.1           nve1

BGW1# show forwarding route 10.10.112.1 vrf egress-loadbalance-resolution-
<Truncated>
------------------+--------------------+---------------+----------------+---------------
Prefix            | Next-hop           | Interface     | Labels         | Partial
 Install
------------------+--------------------+---------------+----------------+---------------
*10.10.112.1/32      192.168.23.6           Ethernet1/22 >> 27 Entries
                     192.168.23.6           Ethernet1/22
                     .....
                     192.168.23.10          Ethernet1/23 >> 27 Entries
                     192.168.23.10          Ethernet1/23
                     .....
                     192.168.31.6           Ethernet1/32 >> 5 Entries
                     192.168.31.6           Ethernet1/33
                     .....
                     192.168.31.10          Ethernet1/32 >> 5 Entries
                     192.168.31.10          Ethernet1/33
                     .....
```

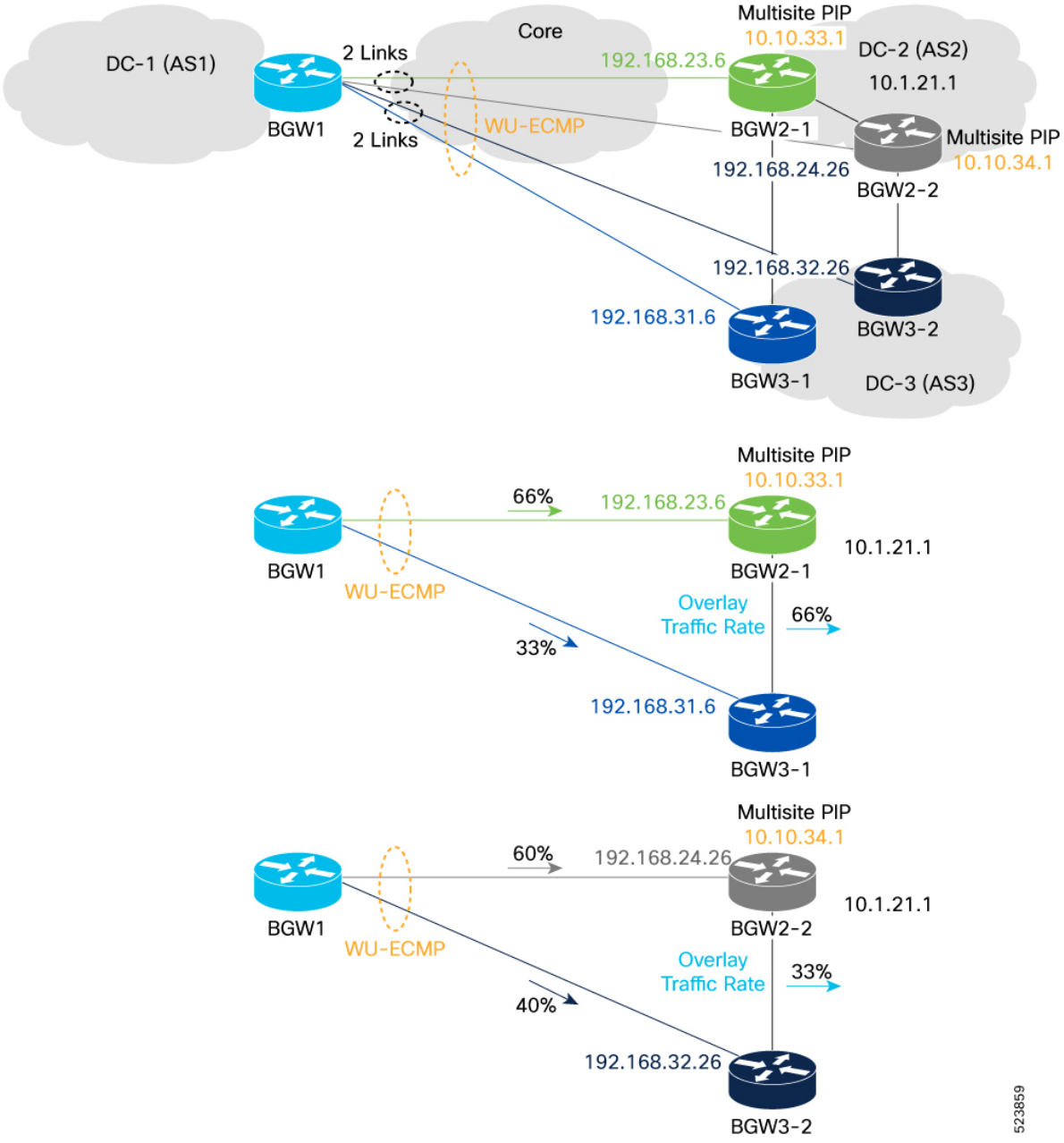
The following summarizes the FIB forwarding entries created based on weight of explicit (3), ECMP (3), and uECMP (1) path sets:

- Explicit Path: 27 Entries
- ECMP Path-Set: 27 Entries
- uECMP Path-Set: 10 Entries
- Traffic Ratio is 27:27:10 = 3:3:1

Dynamic weight (wuECMP) in Overlay and Underlay, and with AIGP in Underlay

The figure below shows the scenario where the BGW devices in DC-2 have been configured with the "dci-advertise-pip" command, so that all the overlay prefixes get advertised to DC-1 using their unique

Multi-Site PIP addresses as next-hop. Because of the specific topology shown below, BGW1 has multiple underlay paths to reach each of the Multi-Site PIP addresses of the BGWs in DC-2, some more direct and some indirect through the BGW nodes in DC-3 (as always, those could simply be routers in the Core instead). Differently to the static weight examples previously discussed, this section covers how to use a more dynamic approach, based on the use of AIGP metric, to assign different weights to the different underlay paths used to reach the two Multi-Site PIP addresses in DC-2 (10.10.33.1 and 10.10.34.1).



Dynamic wuECMP with AIGP in Underlay

1. Originate underlay routes with AIGP metric.

- Specify the underlay routes that need to be advertised with associated an AIGP metric. In this use case, those routes will be the Multi-Site PIP addresses of the BGW nodes in DC-2, representing the next-hops of the overlay prefixes advertised to BGW1 in DC-1.
- On the originator nodes (BGW2-1/BGW2-2), the following two options are available to originate prefixes with AIGP metric:
  - IGP cost
  - Static metric

In the specific use cases being discussed, where the prefix advertised includes the AIGP metric and represents the BGW PIP loopback IP address, utilizing the IGP cost allows for the advertisement of such a prefix with an AIGP metric value of 0 to neighboring devices. Alternatively, it is possible to assign a specific metric value.

The configuration samples below must be applied to the BGW nodes in DC-2:

BGW2-1:

```
interface loopback1
  description #NVE_Source#
  ip address 10.10.33.1/32 tag 54321

route-map RMAP-REDIST-DIRECT permit 10
  match tag 54321
  set aigp-metric igp-cost

router bgp 2
  address-family ipv4 unicast
    redistribute direct route-map RMAP-REDIST-DIRECT
```

BGW2-2:

```
interface loopback1
  description #NVE_Source#
  ip address 10.10.34.1/32 tag 54321

route-map RMAP-REDIST-DIRECT permit 10
  match tag 54321
  set aigp-metric 4 <0 to 4294967295>

router bgp 2
  address-family ipv4 unicast
    redistribute direct route-map RMAP-REDIST-DIRECT
```



Note You can either configure with **set aigp-metric** *value* or with **set aigp-metric igp-cost**. However, both variants cannot co-exist simultaneously.

2. Enable AIGP.
   - Enable AIGP under BGP IPv4 AF for each BGP neighbor on all nodes including the originator BGWs (BGW1/BGW3-1/BGW3-2, core routers, etc..).

```
router bgp 1
  template peer BGP
```

```
address-family ipv4 unicast
  aigp
```

3. Create a Filter-policy



Note Steps 3 to 5 should be applied to the BGWs in DC1.

- Specify the underlay routes to enable ELB wuECMP. The underlay routes are the Multi-Site PIP addresses of the BGWs in DC-2 representing the next-hop for the overlay prefixes advertised to DC-1.

```
ip prefix-list site2_ms_pip1 seq 5 permit 10.10.33.1/32
ip prefix-list site2_ms_pip2 seq 5 permit 10.10.34.1/32
```

- Create a route-map and apply the previously configured prefix-list in the match condition.

```
route-map Filter-Policy permit 10
  match ip address prefix-list site2_ms_pip1 site2_ms_pip2
```

4. Enable ELB Filter-policy (under the IPv4 or IPv4 BGP address-family).

```
router bgp 1
  address-family ipv4 unicast
    load-balance egress filter-policy route-map Filter-Policy
```

5. Verify routes in **Egress-loadbalance-resolution-** VRF table. As observed in the output below, BGW1 by default utilizes the direct underlay paths to reach each PIP address in DC-2.

```
BGW1# show ip route 10.10.33.1 vrf egress-loadbalance-resolution-

10.10.33.1, ubest/mbest: 1/0
  *via 192.168.23.6%default, [20/4], 1d04h, bgp-1, external, tag 2, uecmp ! Green path
BGW1# show ip route 10.10.34.1 vrf egress-loadbalance-resolution-
10.10.34.1/32, ubest/mbest: 1/0
  *via 192.168.24.26%default, [20/8], 1d04h, bgp-1, external, tag 2, uecmp ! Green Path
```

6. Create the Multipath Auto-policy.

- Provide differences in BGP attributes such as AS-Path length and AIGP-metric that can be considered when choosing multipath.
- Specify the maximum number of underlay paths to be computed and installed for ‘Egress Load Balancing’.

```
route-map Auto-Policy permit 10
  set as-path-length difference 1 <1 to 255>
  set aigp-metric difference 10 <1 to 4294967295>
  set maximum-paths 8 <1 to 64>
```



Note If the difference of derived AIGP metric between best path and non-best path is less than 10, non-best path is added to the multipath set of underlay links used to reach the destination PIP addresses.

7. Enable ELB Multipath Auto-policy

```
router bgp 1
  address-family ipv4 unicast
    load-balance egress multipath auto-policy route-map Auto-policy
```

8. Verify routes in **Egress-loadbalance-resolution** VRF table of URIB. As observed in the output below, two uECMP routes are now installed to reach each of the Multi-Site PIP addresses in DC-2. A different weight is associated to each path, based on the AIGP metric information calculated on BGW1 for the received prefixes.

```
BGW1# show ip route 10.10.33.1 vrf egress-loadbalance-resolution-

10.10.33.1/32, ubest/mbest: 4/0
    *via 192.168.23.6%default, [20/4], 23:19:35, bgp-1, weight:2, external, tag 2, uecmp
 ! Green path
    *via 192.168.31.6%default, [20/4], 23:23:59, bgp-1, weight:1, external, tag 3, uecmp
 ! Blue path

BGW1# show ip route 10.10.34.1 vrf egress-loadbalance-resolution-

10.10.34.1/32, ubest/mbest: 4/0
    *via 192.168.24.26%default, [20/8], 1d00h, bgp-1, weight:3, external, tag 2, uecmp
! Green path
    *via 192.168.32.26%default, [20/8], 1d00h, bgp-1, weight:2, external, tag 3, uecmp
! Blue path
```

Dynamic wuECMP with AIGP in Overlay

1. Enable Overlay Next-hop Resolution using the Egress Load Balancing uECMP Underlay Paths.
   - This command enables resolution of EVPN next-hops using the **Egress-loadbalance-resolution-** VRF table.
   - This command must be enabled after enabling the Egress Load Balancing computation in the underlay table.

```
router bgp 1 !BGW1
  address-family l2vpn evpn
    nexthop load-balance egress multisite
```

2. Enable Overlay wuECMP. This command is needed on the BGW nodes in DC2 to ensure that overlay prefixes are advertised with PIP as next-hop.

```
evpn multisite border-gateway 11 !BGW2-1/BGW2-2
  dci-advertise-pip

router bgp 1 !BGW1
  address-family l2vpn evpn
    maximum-paths 8
    maximum-paths unequal-cost
  vrf 3001
    address-family ipv4 unicast
      maximum-paths 8
      maximum-paths unequal-cost
```

3. Verify overlay routes in Tenant VRF table on BGW1. The overlay prefixes are learned with both Multi-Site PIP addresses as next-hops. Based on the output previously shown above, we know that traffic destined to each PIP address will be unequally load-balanced (with different weight) using the green and blue paths. Additionally, as shown in the output below, the underlay AIGP metric information is also leveraged to assign a different metric to the next-hop PIP address, so that unequal load-balancing (with different weight) can also be applied when deciding to which remote BGW node traffic should be encapsulated to.

```
BGW1# show ip route 10.1.21.1 vrf 3001

10.1.21.1/32, ubest/mbest: 2/0
    *via 10.10.33.1%egress-loadbalance-resolution-, [20/2000], 1d02h, bgp-1, weight:2,
```

```
external,
    tag 2, eLB, segid: 3003001 tunnelid: 0x66022101 encap: VXLAN
    *via 10.10.34.1%egress-loadbalance-resolution-, [20/2000], 1d02h, bgp-1, weight:1,
external,
    tag 2, eLB, segid: 3003001 tunnelid: 0x66022201 encap: VXLAN
```

Optional Configuration

1. Enable **bestpath aigp ignore**.

   - To configure a device that is running BGP to NOT evaluate AIGP metric during the best path selection process between two paths when one path does not have the AIGP metric:

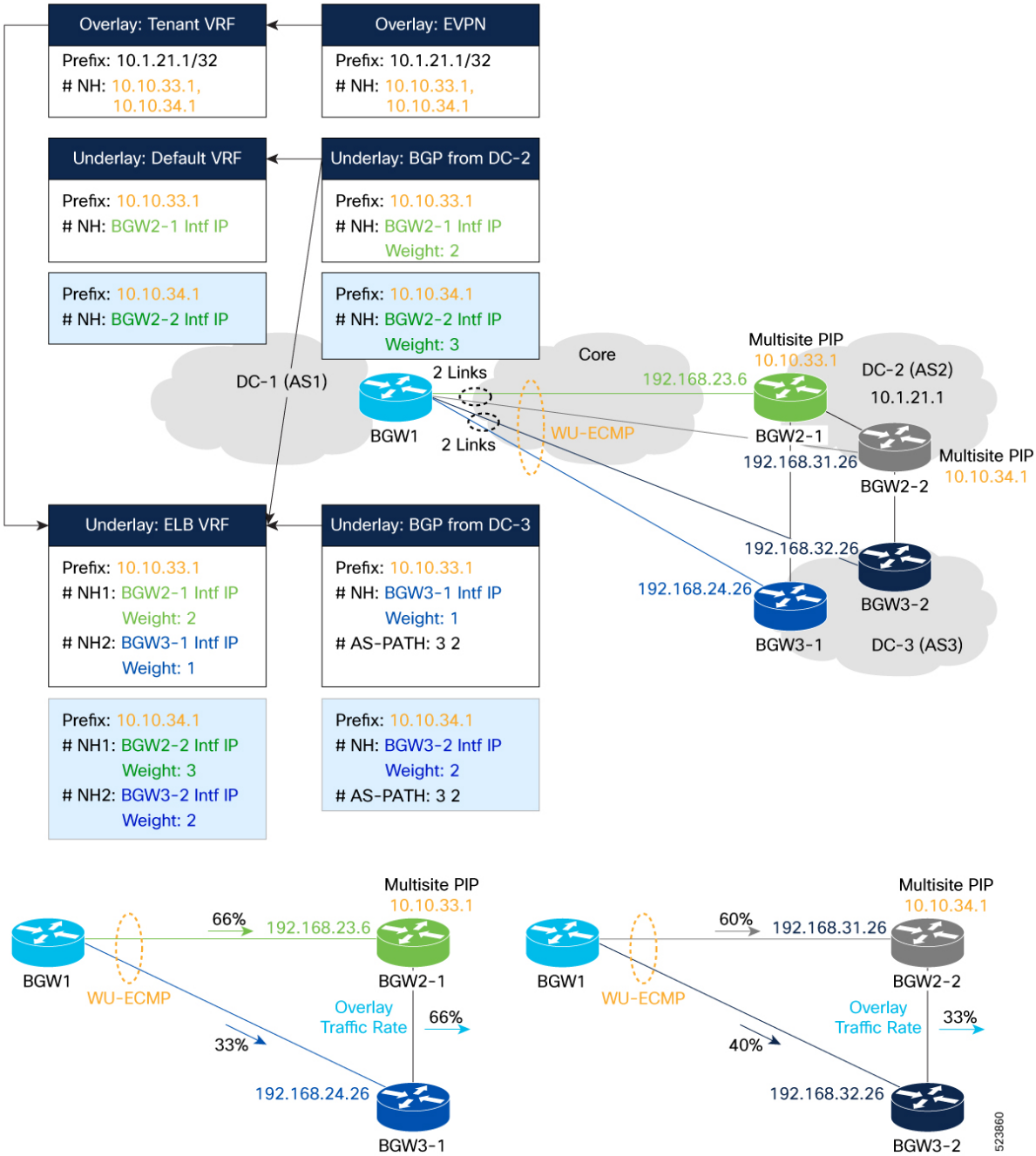
     ```
     router bgp 1
       bestpath aigp ignore
     ```

2. Enable **reference-bandwidth**.

   - It is possible that for 2 eBGP peer R1 and R2, there is a direct link between them on which no IGP is running. To derive the cost of such a link, the reference bandwidth needs to be configured using the following command:

     ```
     router bgp 1
       reference-bandwidth ?
       <1-4000000>  Rate in Mbps (bandwidth) (Default)
                    *Default value is 40000
       <1-4000>     Rate in Gbps (bandwidth)
     ```

Dynamic wuECMP with AIGP Routing details



1. **Dynamic wuECMP with AIGP Underlay Route Programming**

   Underlay route with multipath received from remote BGWs in DC-2is programmed to BGP routing table and Unicast routing table for egress-loadbalance-resolution- VRF on BGW1.

   The following example shows the sample output for the following components. For more information see, Load Share Weight Calculation, on page 348.

   - BGP:

```
BGW1# show bgp ipv4 unicast 10.10.33.1 vrf egress-loadbalance-resolution-
<Truncated>
  Advertised path-id 1
  Path type: external, path is valid, is best path, no labeled nexthop, in rib
             Imported from 10.10.33.1/32 (VRF default)
  AS-Path: 2 , path sourced external to AS
    192.168.23.6 (metric 0) from 192.168.23.6 (10.10.33.1)
      Origin incomplete, MED 0, localpref 100, weight 0, aigp metric weight 2, aigp
 0 derived aigp = 4

  Path type: external, path is valid, not best reason: newer EBGP path, multipath,
no labeled nexthop, in rib
             Imported from 10.10.33.1/32 (VRF default)
  AS-Path: 2 , path sourced external to AS

  Path type: external, path is valid, not best reason: newer EBGP path, multipath,
uecmp, no labeled nexthop, in rib
             Imported from 10.10.33.1/32 (VRF default)
  AS-Path: 3 2 , path sourced external to AS
    192.168.31.6 (metric 0) from 192.168.31.6 (10.10.33.1)
      Origin incomplete, MED not set, localpref 100, weight 0, aigp metric weight
1, aigp 4 derived aigp = 8
```

```
BGW1# show bgp ipv4 unicast 10.10.34.1 vrf egress-loadbalance-resolution-
<Truncated>
  Advertised path-id 1
  Path type: external, path is valid, is best path, no labeled nexthop, in rib
             Imported from 10.10.34.1/32 (VRF default)
  AS-Path: 2 , path sourced external to AS
    192.168.24.26 (metric 0) from 192.168.24.26 (10.10.34.1)
      Origin incomplete, MED 0, localpref 100, weight 0, aigp metric weight 3, aigp
 4 derived aigp = 8

  Path type: external, path is valid, not best reason: newer EBGP path, multipath,
uecmp, no labeled nexthop, in rib
             Imported from 10.10.34.1/32 (VRF default)
  AS-Path: 3 2 , path sourced external to AS
    192.168.32.26 (metric 0) from 192.168.32.26 (10.10.34.1)
      Origin incomplete, MED not set, localpref 100, weight 0, aigp metric weight
2, aigp 8 derived aigp = 12
```

- URIB:

```
BGW1# show ip route 10.10.33.1 detail vrf egress-loadbalance-resolution-
<Truncated>
10.10.33.1/32, ubest/mbest: 4/0
    *via 192.168.23.6%default, [20/4], 23:19:35, bgp-1, weight:2, external, tag 2,
uecmp ! Green path
<Truncated>
    *via 192.168.31.6%default, [20/4], 23:23:59, bgp-1, weight:1, external, tag 3,
uecmp ! Blue path
<Truncated>
```

```
BGW1# show ip route 10.10.34.1 detail vrf egress-loadbalance-resolution-
<Truncated>
10.10.34.1/32, ubest/mbest: 4/0
    *via 192.168.24.26%default, [20/8], 1d00h, bgp-1, weight:3, external, tag 2,
uecmp ! Green path
<Truncated>
    *via 192.168.32.26%default, [20/8], 1d00h, bgp-1, weight:2, external, tag 3,
uecmp ! Blue path
<Truncated>
```

- FIB:

```
BGW1# show forwarding route 10.10.33.1 vrf egress-loadbalance-resolution-
<Truncated>
------------------+--------------------+---------------+----------------+---------------
Prefix            | Next-hop           | Interface     | Labels         | Partial
 Install
------------------+--------------------+---------------+----------------+---------------
*10.10.33.1/32       192.168.23.6           Ethernet1/22 ! 21 Entries
                     192.168.23.6           Ethernet1/22
                     .....
                     192.168.31.6           Ethernet1/32 ! 11 Entries
                     192.168.31.6           Ethernet1/33
                     .....
```

The following summarizes the FIB forwarding entries created based on weight of ECMP (2) and uECMP (1) path sets:

- ECMP Path-Set: 21 Entries
- uECMP Path-Set: 11 Entries
- Traffic Ratio is 21 : 11 = 2 : 1

```
BGW1# show forwarding route 10.10.34.1 vrf egress-loadbalance-resolution-
<Truncated>
------------------+--------------------+---------------+----------------+---------------
Prefix            | Next-hop           | Interface     | Labels         | Partial
 Install
------------------+--------------------+---------------+----------------+---------------
*10.10.34.1/32       192.168.24.26          Ethernet1/27 ! 19 Entries
                     192.168.24.26          Ethernet1/27
                     .....
                     192.168.32.26          Ethernet1/37 ! 13 Entries
                     192.168.32.26          Ethernet1/37
                     .....
```

The following summarizes the FIB forwarding entries created based on weight of ECMP (3) and uECMP (2) path sets:

- ECMP Path-Set: 19 Entries
- uECMP Path-Set: 13 Entries
- Traffic Ratio is 19 : 13 = 3 : 2

2. **Dynamic wuECMP with AIGP Overlay Route Programming**

   Overlay routes will use **egress-loadbalance-resolution-** VRF table on BGW1 to resolve the PIP next-hops for the overlay prefixes received from DC-2.

   The following example shows the sample output for following components. For more information see, Weight Derivation in Underlay, on page 348.

   - BGP:

```
BGW1# show bgp l2vpn evpn 10.1.21.1
<Truncated>
  Advertised path-id 1
  Path type: external, path is valid, is best path, no labeled nexthop, has esi_gw
             Imported to 3 destination(s)
             Imported paths list: 3001 L3-3003001 L2-2001001
  AS-Path: 2 , path sourced external to AS
```

```
    10.10.33.1 (metric 4) from 10.10.33.1 (10.10.33.1)
<Truncated>
  Path type: external, path is valid, not best reason: NH metric, multipath, no
labeled nexthop, has esi_gw
             Imported to 3 destination(s)
             Imported paths list: 3001 L3-3003001 L2-2001001
  AS-Path: 2 , path sourced external to AS
    10.10.34.1 (metric 8) from 10.10.34.1 (10.10.34.1)

BGW1# show bgp ipv4 unicast 10.1.21.1 vrf 3001
<Truncated>
  Advertised path-id 1, VPN AF advertised path-id 1
  Path type: external, path is valid, is best path, no labeled nexthop, in rib, has
 esi_gw
             Imported from
21:2001001:[2]:[0]:[0]:[48]:[0010.0100.2101]:[32]:[10.1.21.1]/272
  AS-Path: 2 , path sourced external to AS
    10.10.33.1 (metric 4) from 10.10.33.1 (10.10.33.1)
      Origin IGP, MED 2000, localpref 100, weight 0, igp metric weight 2
<Truncated>
  Path type: external, path is valid, not best reason: NH metric, multipath, no
labeled nexthop, in rib, has esi_gw
             Imported from
21:2001001:[2]:[0]:[0]:[48]:[0010.0100.2101]:[32]:[10.1.21.1]/272
  AS-Path: 2 , path sourced external to AS
    10.10.34.1 (metric 8) from 10.10.34.1 (10.10.34.1)
      Origin IGP, MED 2000, localpref 100, weight 0, igp metric weight 1
```

- URIB:

```
BGW1# show ip route 10.1.21.1 detail vrf 3001
<Truncated>
10.1.21.1/32, ubest/mbest: 2/0
   *via 10.10.33.1%egress-loadbalance-resolution-, [20/2000], 1d02h, bgp-1, weight:2,
 external,
    tag 2, eLB, segid: 3003001 tunnelid: 0x66022101 encap: VXLAN
<Truncated>
   *via 10.10.34.1%egress-loadbalance-resolution-, [20/2000], 1d02h, bgp-1, weight:1,
 external,
    tag 2, eLB, segid: 3003001 tunnelid: 0x66022201 encap: VXLAN
```

- FIB:

```
BGW1# show forwarding route 10.1.21.1 vrf 3001
<Truncated>
-----------------+--------------------+---------------+-----------------+---------------
Prefix           | Next-hop           | Interface     | Labels          | Partial
 Install
-----------------+--------------------+---------------+-----------------+---------------
10.1.21.1/32         10.10.33.1           nve1 ! 21 Entries
                     10.10.33.1           nve1
                     .....
                     10.10.34.1           nve1 ! 11 Entries
                     10.10.34.1           nve1
                     .....
```

- L2RIB: Mac Route (EVPN Type-2) with Weight:

  Beginning with Cisco NX-OS Release 10.5(3)F, the labeled nexthop and an asymmetric VNI flag is added as shown. For symmetric VNI, labels and flags will not be shown as part of nexthops for EAD and PL.

```
switch# show l2route evpn mac evi 1001 detail

Flags -(Rmac):Router MAC (Stt):Static (L):Local (R):Remote (V):vPC link
```

```
(Dup):Duplicate (Spl):Split (Rcv):Recv (AD):Auto-Delete (D):Del Pending
(S):Stale (C):Clear, (Ps):Peer Sync (O):Re-Originated (Nho):NH-Override
(Pf):Permanently-Frozen, (Orp): Orphan

Topology    Mac Address    Prod   Flags         Seq No     Next-Hops

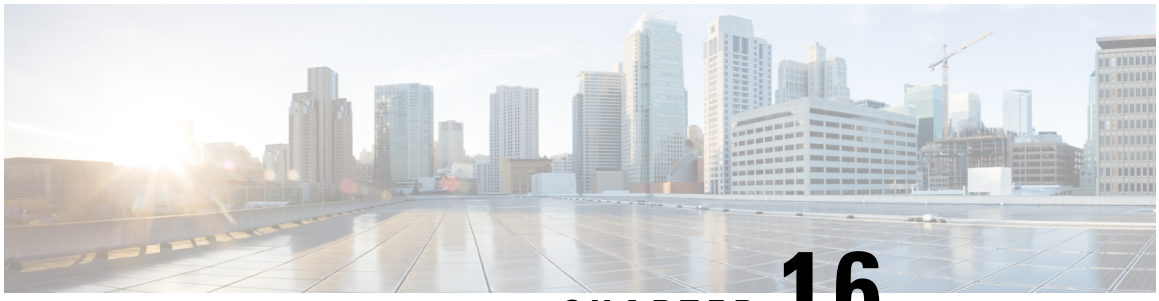
----------- -------------- ------ ------------- ----------
--------------------------------------
100         abcd.ef12.3456 BGP    Rcv               0          10.10.33.1 (Label:
20000)(Flags: Asy)
                                                               10.10.34.1 (Label:
10000)(Flags: Asy)
            Route Resolution Type: ESI
            Forwarding State: Resolved (PL)
            Resultant PL: 10.10.33.1(Wt: 2) (Label: 20000)(Flags: Asy)
                          10.10.34.1(Wt: 1) (Label: 10000)(Flags: Asy)
            Sent To: L2FM
            ESI : aaaa.aaaa.aaaa.aaaa.99aa


BGW1# show l2route evpn path-list all detail
(R) = Remote Global EAD NH Peerid resolved,
(UR) = Remote Global EAD NH Peerid unresolved
Flags - (A):All-Active (Si):Single-Active

NH Flags: Asy = Asymmetric VNI
Topology ID  Prod   ESI                       ECMP Label        Flags  Client Ctx
MACs       Sent To
------------ ------ ------------------------- ---------- ------ ----------- ----------
       ----------
 100          UFDM   aaaa.aaaa.aaaa.aaaa.99aa  0          A             0
 1          UFDM
                    CP Next-Hops: 5.5.5.5, 6.6.6.6
                    Gbl EAD Next-Hops: 10.10.33.1 (5,R)
                                       10.10.34.1 (6,R)
                    Res Next-Hops: 10.10.33.1(Wt: 2) (Label: 20000)(Flags: Asy)
                                   10.10.34.1(Wt: 1) (Label: 10000)(Flags: Asy)
                    Bkp Next-Hops:
                  Res Next-Hops from UFDM: 10.10.33.1(Wt: 2) (Label: 20000)(Flags:
 Asy)
                                           10.10.34.1(Wt: 1) (Label: 10000)(Flags:
 Asy)
```

- L2FM - Mac Route (EVPN Type-2) with Weight:

```
BGW1# show mac address-table vlan 1001 address 0010.0100.2101
<Truncated>
 VLAN     MAC Address      Type      age     Secure NTFY Ports
---------+-----------------+--------+---------+------+----+------------------
C 1001     0010.0100.2101   dynamic  NA         F      F    nve1(10.10.33.1[Wt: 2]
10.10.34.1[Wt: 1])
```



CHAPTER 16

Configuring Tenant Routed Multicast (TRM)

This chapter contains these sections:

- About Tenant Routed Multicast, on page 382
- About Tenant Routed Multicast Mixed Mode, on page 383
- About Tenant Routed Multicast with IPv6 Overlay, on page 383
- About Multicast Flow Path Visibility for TRM Flows, on page 384
- About Configuring VXLAN EVPN and TRM with IPv6 Underlay, on page 384
- Guidelines and Limitations for Tenant Routed Multicast, on page 385
- Guidelines and Limitations for Layer 3 Tenant Routed Multicast, on page 387
- Guidelines and Limitations for Layer 2/Layer 3 Tenant Routed Multicast (Mixed Mode), on page 390
- Guidelines and Limitations for VXLAN EVPN and TRM with IPv6 in the Multicast Underlay, on page 391
- Rendezvous Point for Tenant Routed Multicast, on page 392
- Configuring a Rendezvous Point for Tenant Routed Multicast, on page 393
- Configuring a Rendezvous Point Inside the VXLAN Fabric, on page 394
- Configuring an External Rendezvous Point, on page 395
- Configuring RP Everywhere with PIM Anycast, on page 397
- Configuring RP Everywhere with MSDP Peering, on page 403
- Configuring Layer 3 Tenant Routed Multicast, on page 410
- Configuring TRM on the VXLAN EVPN Spine, on page 415
- Configuring Tenant Routed Multicast in Layer 2/Layer 3 Mixed Mode, on page 417
- Configuring VXLAN EVPN and TRM with IPv6 Multicast Underlay, on page 422
- Configuring Layer 2 Tenant Routed Multicast, on page 425
- Configuring TRM with vPC Support, on page 426
- Configuring TRM with vPC Support (Cisco Nexus 9504-R and 9508-R), on page 429
- Flex Stats for TRM, on page 432
- Configuring Flex Stats for TRM, on page 433
- Configuring TRM Data MDT, on page 434
- Configuring IGMP Snooping, on page 437
- Verifying VXLAN EVPN and TRM with IPv6 Multicast Underlay, on page 438
- Example Configuration for VXLAN EVPN and TRM with IPv6 Multicast Underlay, on page 442

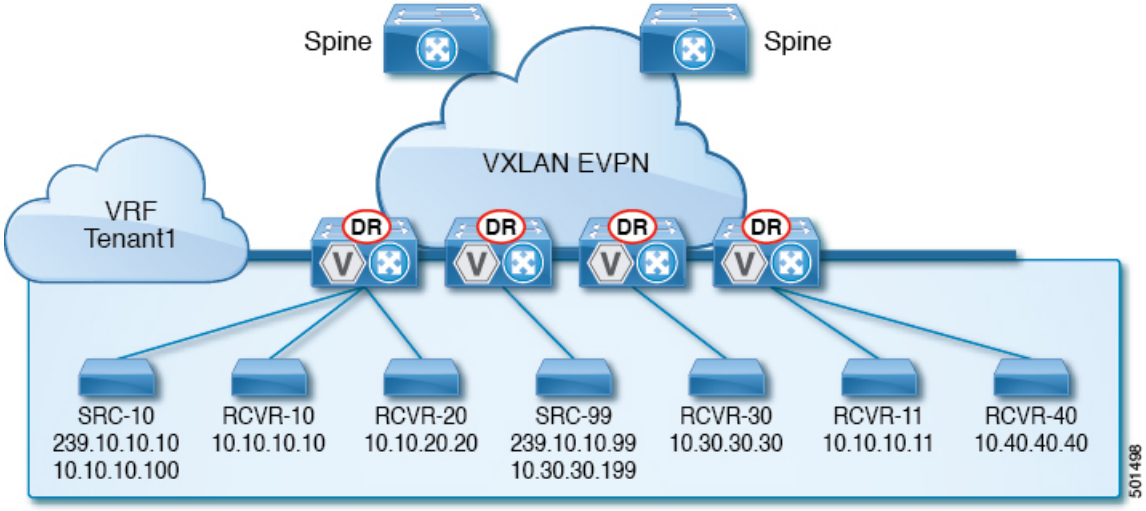
About Tenant Routed Multicast

Tenant Routed Multicast (TRM) enables multicast forwarding on the VXLAN fabric that uses a BGP-based EVPN control plane. TRM provides multi-tenancy aware multicast forwarding between senders and receivers within the same or different subnet local or across VTEPs.

This feature brings the efficiency of multicast delivery to VXLAN overlays. It is based on the standards-based next generation control plane (ngMVPN) described in IETF RFC 6513, 6514. TRM enables the delivery of customer IP multicast traffic in a multitenant fabric, and thus in an efficient and resilient manner. The delivery of TRM improves Layer-3 overlay multicast functionality in our networks.

While BGP EVPN provides the control plane for unicast routing, ngMVPN provides scalable multicast routing functionality. It follows an “always route” approach where every edge device (VTEP) with distributed IP Anycast Gateway for unicast becomes a Designated Router (DR) for Multicast. Bridged multicast forwarding is only present on the edge-devices (VTEP) where IGMP snooping optimizes the multicast forwarding to interested receivers. Every other multicast traffic beyond local delivery is efficiently routed.

Figure 34: VXLAN EVPN TRM

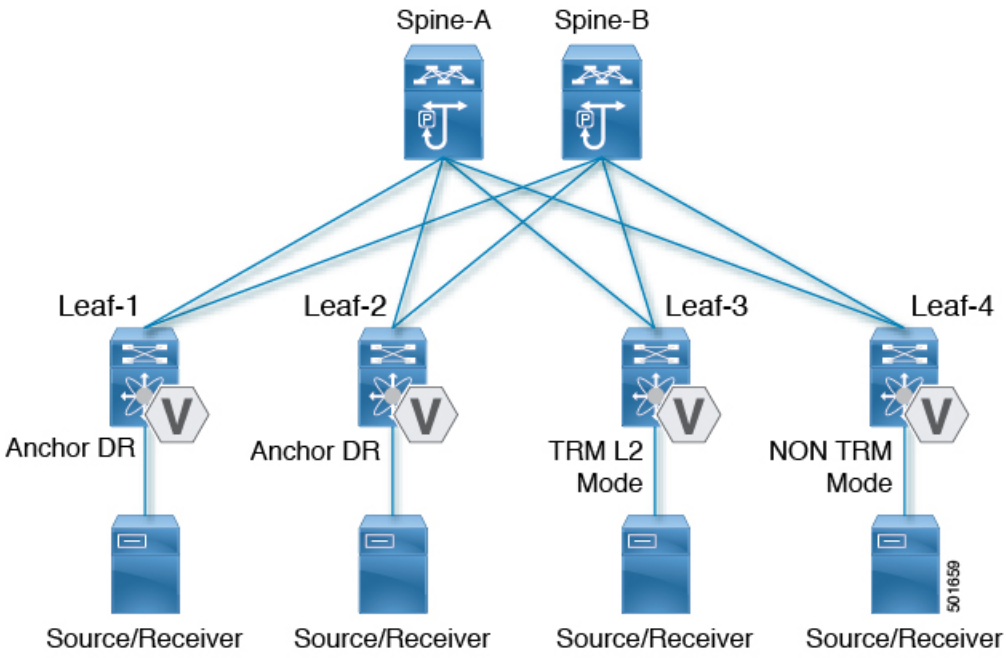


With TRM enabled, multicast forwarding in the underlay is leveraged to replicate VXLAN encapsulated routed multicast traffic. A Default Multicast Distribution Tree (Default-MDT) is built per-VRF. This is an addition to the existing multicast groups for Layer-2 VNI Broadcast, Unknown Unicast, and Layer-2 multicast replication group. The individual multicast group addresses in the overlay are mapped to the respective underlay multicast address for replication and transport. The advantage of using a BGP-based approach allows the VXLAN BGP EVPN fabric with TRM to operate as fully distributed Overlay Rendezvous-Point (RP), with the RP presence on every edge-device (VTEP).

A multicast-enabled data center fabric is typically part of an overall multicast network. Multicast sources, receivers, and multicast rendezvous points, might reside inside the data center but might also be inside the campus or externally reachable via the WAN. TRM allows a seamless integration with existing multicast networks. It can leverage multicast rendezvous points external to the fabric. Furthermore, TRM allows for tenant-aware external connectivity using Layer-3 physical interfaces or subinterfaces.

About Tenant Routed Multicast Mixed Mode

Figure 35: TRM Layer 2/Layer 3 Mixed Mode



About Tenant Routed Multicast with IPv6 Overlay

Beginning with Cisco NX-OS Release 10.2(1), Tenant Routed Multicast (TRM) supports IPv6 in the overlay.

Guidelines and Limitations for TRM with IPv6 Overlay

The following are supported by TRM with IPv6 Overlay:

- Multicast IPv4 underlay within fabric. Bidir and SSM are not supported.
- IPv4 Underlay in the data center core for multisite.
- IPv4 overlay only, IPv6 overlay Only, combination of IPv4 and IPv6 overlays
- IPv6 in the underlay
- Anycast Border Gateway with Border Leaf Role
- vPC support on Border Gateway and Leaf
- Virtual MCT on Leaf
- Anycast RP (internal, external, and RP-everywhere)
- Multisite Border Gateway is supported on Cisco Nexus 9300 -FX3, -GX, GX2, -H2R, and -H1 TORs.
- RP-everywhere with Anycast RP is supported.

- TRMv6 is supported only on default system routing mode.
- MLD snooping with VxLAN VLANs with TRM
- PIM6 SVI and MLD snooping configuration on the VLAN are not supported.
- TRM with IPv6 Overlay is supported on Cisco Nexus 9300 -FX, - FX2, -FX3, -GX, - GX2, -H2R, -H1 TORs.

The following are not supported by TRM with IPv6 Overlay:

- L2 TRM
- VXLAN flood mode on L2 VLANs with L3TRM is not supported
- L2-L3 TRM Mixed Mode
- VXLAN Ingress Replication within a single site
- MLD snooping with VXLAN VLANs without TRM
- PIM6 SVI configuration without MLD snooping
- MSDP

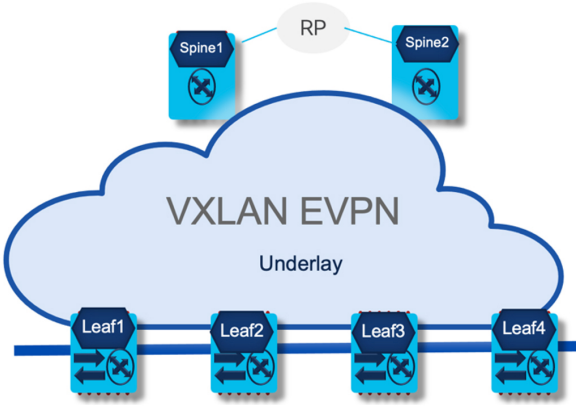
About Multicast Flow Path Visibility for TRM Flows

Beginning with Cisco NX-OS Release 10.3(2)F, the Multicast Flow Path Visualization (FPV) for TRM Flows feature is supported for TRM L3 mode and underlay multicast along with the already supported multicast flows. This feature enables you to export all multicast states in a Cisco Nexus 9000 Series switch. This helps to have a complete and reliable traceability of the flow path from the source to a receiver. To enable Multicast Flow Path Data Export on Cisco Nexus 9000 Series switches, use the **multicast flow-path export** command.

About Configuring VXLAN EVPN and TRM with IPv6 Underlay

Beginning with Cisco NX-OS Release 10.4(2)F, the support is provided for VXLAN with IPv6 Multicast in the Underlay. Hosts in the overlay can be IPv4 or IPv6. This requires IPv6 versions of the unicast routing protocols and using IPv6 multicast in the underlay (PIMv6). Any multi-destination overlay traffic (such as TRM, BUM) can use the IPv6 multicast underlay.

Figure 36: Topology - VXLAN EVPN with IPv6 Multicast Underlay



The above topology shows four leafs and two spines in a VXLAN EVPN fabric. The underlay is an IPv6 Multicast running PIMv6. RP is positioned in the spine with anycast RP.

Beginning with Cisco NX-OS Release 10.4(3)F, the combination of PIMv6 underlay on the fabric side and Ingress Replication (IPv6) on Data Center Interconnect (DCI) side is supported on Cisco Nexus 9300-FX/FX2/FX3/GX/GX2/H2R/H1 ToR switches and 9500 switches with X9716D-GX and X9736C-FX line cards.

Beginning with Cisco NX-OS Release 10.5(1)F, the underlay network supports the following combinations for VXLAN EVPN:

- In the data center fabric, both Multicast Underlay (PIMv6) and Ingress Replication (IPv6) are supported.
- In the Data Center Interconnect (DCI), only Ingress Replication (IPv6) is supported.

Guidelines and Limitations for Tenant Routed Multicast

Tenant Routed Multicast (TRM) has the following guidelines and limitations:

- Beginning with Cisco NX-OS Release 10.1(2), TRM Multisite with vPC BGW is supported.
- Beginning with Cisco NX-OS Release 10.2(1q)F, VXLAN TRM is supported on Cisco Nexus N9K-C9332D-GX2B platform switches.
- Beginning with Cisco NX-OS Release 10.2(3)F, VXLAN TRM is supported on Cisco Nexus 9364D-GX2A, and 9348D-GX2A platform switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, VXLAN TRM is supported on Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, VXLAN TRM is supported on Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, VXLAN TRM is supported on Cisco Nexus 9364C-H1 switches.
- Support straight-through FEX connected to a standalone VXLAN VTEP and multicast source/receiver behind the FEX port.

- With TRM enabled, we do not support a multicast receiver behind active-active FEX and vPC behind straight-through FEX.
- Cisco Nexus 9300-FX, and 9300-FX2 switches support FEX.
- FEX is not supported on Cisco Nexus 9500 platform switches.
- If VXLAN TRM feature is enabled on a VTEP, it would stop to send IGMP messages to the VXLAN fabric.
- The Guidelines and Limitations for VXLAN also apply to TRM.
- With TRM enabled, SVI as a core link is not supported.
- If TRM is configured, ISSU is disruptive.
- TRM supports IPv4 and IPv6 multicast underlay.
- TRM supports overlay PIM ASM and PIM SSM only. PIM BiDir is not supported in the overlay.
- RP has to be configured either internal or external to the fabric.
- The internal RP must be configured on all TRM-enabled VTEPs including the border nodes.
- The external RP must be external to the border nodes.
- The RP must be configured within the VRF pointing to the external RP IP address (static RP). This ensures that unicast and multicast routing is enabled to reach the external RP in the given VRF.
- In a Transit Routing Multicast (TRM) deployment, the RP-on-stick model can sometimes lead to traffic drops if there is flapping on the Protocol Independent Multicast (PIM) enabled interface. Use the **ip pim spt-switch-graceful** command on the turnaround router that leads to the RP. This command allows for a graceful switch to the Shortest Path Tree (SPT) during flapping, which can minimize traffic drops.

  Beginning with Cisco NX-OS Release 10.6(1)F, the **ip pim spt-switch-graceful** command is enabled by default. To disable the command, use **no ip pim spt-switch-graceful** command.
- Replication of first packet is supported only on Cisco Nexus 9300 – FX, FX2 family switches.
- Beginning with Cisco NX-OS Release 10.2(3)F, Replication of first packet is supported on the Cisco Nexus 9300-FX3 platform switches.
- TRM with Multi-Site is not supported on Cisco Nexus 9504-R platforms.
- TRM supports multiple border nodes. Reachability to an external RP/source via multiple border leaf switches is supported with ECMP and requires symmetric unicast routing.
- Both PIM and **ip igmp snooping vxlan** must be enabled on the L3 VNI's VLAN in a VXLAN vPC setup.
- For traffic streams with an internal source and external L3 receiver using an external RP, the external L3 receiver might send PIM S,G join requests to the internal source. Doing so triggers the recreation of S,G on the fabric FHR, and it can take up to 10 minutes for this S,G to be cleared.
- Beginning with Cisco NX-OS Release 10.3(1)F, the Real-time/flex statistics for TRM is supported on Cisco Nexus 9300-X Cloud Scale Switches.
- TRM supports vPC fabric peering leaf's as well as vPC/Anycast BGW.
- Beginning with Cisco NX-OS Release 10.6(1)F, Cisco Nexus 9336C-SE1 switches supports TRMv4 on vPC leaf, vPC fabric peering leaf, Anycast BGW, and standalone leaf.

VXLAN with TRM Upgrade limitations

After upgrading a Cisco NX-OS 9000 Series switches configured with VXLAN (specifically VRF-related configurations) from Cisco NX-OS Release 7.x through 9.3 to 10.3(6) or earlier, two issues arise:

- The startup-config displays both legacy and new Layer 3 VNID configuration modes
- TRM traffic's RPF changes to the new mode for S,Gs, causing multicast traffic forwarding problems.

To avoid these issues, follow these steps:

- Enable the REST configuration input using the following commands:

```
feature nxapi
            nxapi http port 80
```

- Open a browser and enter the management IP address of the switch. This will open the Sandbox page. Use the same credentials as the switch admin login to sign in.
- In the top input textbox, enter the following command for each VRF that has an issue with the VNI ID:

```
vrf context tenant-1
        no vni 50000 l3
```

- On the right side of the page, set the Method to **NXAPI-REST(DME)** and keep the Input Type as **cli**.
- Click the **Convert (with DN)** button in the middle of the page. This will generate the XML equivalent of the configuration change.
- When the XML appears in the second textbox, click **Send** to apply the changes and remove the VNI ID configuration from the switch.
- To ensure the changes are applied, run the command:

```
copy running-config startup-config
```

Guidelines and Limitations for Layer 3 Tenant Routed Multicast

Layer 3 Tenant Routed Multicast (TRM) has the following configuration guidelines and limitations:

- When upgrading from Cisco NX-OS Release 9.3(3) to Cisco NX-OS Release 9.3(6), if you do not retain configurations of the TRM enabled VRFs from Cisco NX-OS Release 9.3(3), or if you create new VRFs after the upgrade, the auto-generation of **ip multicast multipath s-g-hash next-hop-based** CLI, when **feature ngmvpn** is enabled, will not happen. You must enable the CLI manually for each TRM enabled VRF.
- Layer 3 TRM is supported for Cisco Nexus 9300-FX/FX2/FX3/FXP and 9300-GX platform switches.
- Beginning with Cisco NX-OS Release 10.2(3)F, Layer 3 TRM is supported on the Cisco Nexus 9300-GX2 platform switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, Layer 3 TRM is supported on the Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, Layer 3 TRM is supported on the Cisco Nexus 93400LD-H1 switches.

- Beginning with Cisco NX-OS Release 10.4(3)F, Layer 3 TRM is supported on the Cisco Nexus 9364C-H1 switches.
- Beginning with Cisco NX-OS Release 9.3(7), Cisco Nexus N9K-C9316D-GX, N9K-C9364C-GX, and N9K-X9716D-GX platform switches support the combination of Layer 3 TRM and EVPN Multi-Site.
- Cisco Nexus 9300-GX platform switches do not support the combination of Layer 3 TRM and EVPN Multi-Site in Cisco NX-OS Release 9.3(5).
- Beginning with Cisco NX-OS Release 10.2(3)F, the combination of Layer 3 TRM and EVPN Multi-Site is supported on the Cisco Nexus 9300-GX2 platform switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, the combination of Layer 3 TRM and EVPN Multi-Site is supported on the Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, the combination of Layer 3 TRM and EVPN Multi-Site is supported on the Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, the combination of Layer 3 TRM and EVPN Multi-Site is supported on the Cisco Nexus 9364C-H1 switches.
- Beginning with Cisco NX-OS Release 9.3(3), the Cisco Nexus 9504 and 9508 platform switches with -R/RX line cards support TRM in Layer 3 mode. This feature is supported on IPv4 overlays only. Layer 2 mode and L2/L3 mixed mode are not supported.

  The Cisco Nexus 9504 and 9508 platform switches with -R/RX line cards can function as a border leaf for Layer 3 unicast traffic. For Anycast functionality, the RP can be internal, external, or RP everywhere.
- When configuring TRM VXLAN BGP EVPN, the following platforms are supported:
  - Cisco Nexus 9300-FX/FX2/FX3/FXP platform switches.
  - Cisco Nexus 9300-GX/GX2 platform switches.
  - Cisco Nexus 9300-H2R/H1 platform switches.
  - Cisco Nexus 9500 platform switches with X97160YC-EX, 9700-FX line cards, 9700-FX3 line cards.
- Layer 3 TRM and VXLAN EVPN Multi-Site are supported on the same physical switch. For more information, see Configuring VXLAN EVPN Multi-Site, on page 293.
- TRM Multi-Site functionality is not supported on Cisco Nexus 9504 platform switches with -R/RX line cards.
- If one or both VTEPs is a Cisco Nexus 9504 or 9508 platform switch with -R/RX line cards, the packet TTL is decremented twice, once for routing to the L3 VNI on the source leaf and once for forwarding from the destination L3 VNI to the destination VLAN on the destination leaf.
- TRM with vPC border leafs is supported only for Cisco Nexus 9300-FX/FX2/FX3/GX/GX2/H2R/H1 platform switches and Cisco Nexus 9500 platform switches with X97160YC-EX, FX/FX3 or -R/RX line cards. The **advertise-pip** and **advertise virtual-rmac** commands must be enabled on the border leafs to support this functionality. For configuration information, see the "Configuring VIP/PIP" section.
- To support any Layer 3 source behind one of the vPC peers, whether physical or virtual MCT, a physical link configured as VRF-lite is required between the vPC peers. This setup is necessary to accommodate a receiver located behind the vPC peer, especially if it is the sole receiver in the fabric. This requirement applies to all scenarios where the vPC functions as a BGW, border Leaf, or an internal Leaf.

On the receiving vPC peer, the VRF-lite link must have a superior reachability metric to the L3 source compared to any other paths (iBGP or eBGP) to be selected as the RPF towards the L3 source. In this configuration, traffic will flow directly to the receiver without traversing the EVPN fabric.

- Well-known local scope multicast (224.0.0.0/24) is excluded from TRM and is bridged.
- When an interface NVE is brought down on the border leaf, the internal overlay RP per VRF must be brought down.
- Beginning with Cisco NX-OS Release 10.3(1)F, TRM support for the new L3VNI mode CLIs are provided on Cisco Nexus 9300-X Cloud Scale switches.
- Beginning Cisco NXOS release 10.2(1)F, TRM Flow Path Visualization is supported for flows within a single VXLAN EVPN site.
- Beginning Cisco NXOS Release 10.3(2)F, TRM Flow Path Visualization support has been extended to below traffic patterns on Cisco Nexus 9000 Series platform switches:
  - TRM Multisite DCI Multicast
  - TRM Multisite DCI IR
  - TRM Data MDT
  - TRM on Virtual MCT vPC
  - TRM using new L3VNI
  - BUM Traffic visibility is not supported.
- Beginning with Cisco NX-OS Release 10.4(3)F, the TRM Multi-Site Anycast BGW on Cisco Nexus 9808/9804 switches with Cisco Nexus X9836DM-A and X98900CD-A line cards support the following features:
  - TRMv4
  - Ingress Replication between DCI peers across the core
  - Multicast underlay for fabric peers.
  - Only new L3VNI mode is supported. However, the traditional L3VNI mode is not supported

  TRM Multi-Site Anycast BGW on Cisco Nexus 9808/9804 switches with Cisco Nexus X9836DM-A and X98900CD-A line cards do not support the following features:
  - TRMv6
  - Data MDT
  - Multicast underlay between DCI peers across the core is not supported.
- Beginning with Cisco NX-OS Release 10.6(1)F, Cisco Nexus 9336C-SE1 switches supports TRMv4 L3 on vPC leaf, vPC fabric peering leaf, Anycast BGW, and standalone leaf.

  Supported features are:
  - TRMv4,
  - Ingress Replication between DCI peers across the core,

- Multicast underlay for fabric peers, and
- VLAN based L3VNI configuration is deprecated. Only **vni vnid l3** command is supported.

Unsupported features are:

- TRMv6 and
- Data MDT.

Guidelines and Limitations for Layer 2/Layer 3 Tenant Routed Multicast (Mixed Mode)

Layer 2/Layer 3 Tenant Routed Multicast (TRM) has the following configuration guidelines and limitations:

- All TRM Layer 2/Layer 3 configured switches must be Anchor DR. This is because in TRM Layer 2/Layer 3, you can have switches configured with TRM Layer 2 mode that co-exist in the same topology. This mode is necessary if non-TRM and Layer 2 TRM mode edge devices (VTEPs) are present in the same topology.
- Anchor DR is required to be an RP in the overlay.
- An extra loopback is required for anchor DRs.
- Non-TRM and Layer 2 TRM mode edge devices (VTEPs) require an IGMP snooping querier configured per multicast-enabled VLAN. Every non-TRM and Layer 2 TRM mode edge device (VTEP) requires this IGMP snooping querier configuration because in TRM multicast control-packets are not forwarded over VXLAN.
- The IP address for the IGMP snooping querier can be re-used on non-TRM and Layer 2 TRM mode edge devices (VTEPs).
- The IP address of the IGMP snooping querier in a VPC domain must be different on each VPC member device.
- When interface NVE is brought down on the border leaf, the internal overlay RP per VRF should be brought down.
- The NVE interface must be shut and unshut while configuring the **ip multicast overlay-distributed-dr** command.
- Beginning with Cisco NX-OS Release 9.2(1), TRM with vPC border leafs is supported. Advertise-PIP and Advertise Virtual-Rmac need to be enabled on border leafs to support with functionality. For configuring advertise-pip and advertise virtual-rmac, see the "Configuring VIP/PIP" section.
- Anchor DR is supported only on the following hardware platforms:
  - Cisco Nexus 9300-FX/FX2 platform switches
  - Cisco Nexus 9500 platform switches with X97160YC-EX, 9700-FX/FX3 line cards.
- Beginning with Cisco NX-OS Release 10.2(3)F, Anchor DR is supported on the Cisco Nexus 9300-FX3 platform switches.

- Layer 2/Layer 3 Tenant Routed Multicast (TRM) is not supported on Cisco Nexus 9300-FX3/GX/GX2/H2R/H1 platform switches.

Guidelines and Limitations for VXLAN EVPN and TRM with IPv6 in the Multicast Underlay

VXLAN EVPN and TRM with IPv6 Multicast Underlay has the following guidelines and limitations:

- Spine-based static RP is supported in underlay.
- Cisco Nexus 9300-FX, FX2, FX3, GX, GX2, H2R, and H1 ToR switches are supported as the leaf VTEP.
- Cisco Nexus X9716D-GX and X9736C-FX line cards are supported only on the spine (EoR).
- When an EoR is deployed as a spine node with Multicast Underlay (PIMv6) Any-Source Multicast (ASM), it in mandatory to configure non-default template using one of the following commands in global configuration mode:
  - **system routing template-multicast-heavy**
  - **system routing template-multicast-ext-heavy**
- OSPFv3, ISIS, eBGP underlay is supported.
- PIMv6 ASM (sparse mode) is supported in underlay.
- PIMv6 Anycast RP is supported in underlay as RP redundancy.
- Underlay IPv6 Multicast is supported.
- Underlay IPv6 Multicast is not supported on EOR platforms as a leaf.
- For overlay traffic, each Cisco Nexus 9000 leaf switch is an RP. External RP is also supported.
- EVPN TRMv4 and TRMv6 with IPv6 Multicast Underlay are supported on the Fabric.
- Fabric Peering and Multisite are not supported with IPv6 multicast underlay.
- The global mcast-group under NVE should not be configured as SSM range, and vice versa. If there is no explicit SSM configuration, then 232/8 is the default in data plane. hence 232.0.0.0/8 should not be configured as SSM and vice versea.
- GPO is not supported with IPv6 multicast underlay.
- For EVPN TRMv4 and TRMv6 with IPv6 Multicast Underlay, the TCAM region for ingress sup region must be carved to 768.
  - Check the ingress sup region using **show hardware access-list tcam region** command.
  - If the ingress sup region is not 768 or above, you must configure using the **hardware access-list tcam region ing-sup 768** command.

Note If you get an error, "Aggregate ingress TCAM allocation failure" while configuring ing-sup as 768, you must borrow the amount from other TCAM regions.

- Reload the device after this configuration.

- Beginning with Cisco NX-OS Release 10.5(1)F, VXLAN EVPN in the data center fabric supports both Multicast Underlay (PIMv6) Any-Source Multicast (ASM) and Ingress Replication (IPv6) in the underlay. This support is available on the following switches and line cards:
  - Cisco Nexus 9300-FX, FX2, FX3, GX, GX2, H2R, and H1 ToR switches as the leaf VTEPs.
  - Cisco Nexus N9K-X9716D-GX and N9K-X9736C-FX line cards as spines if the underlay is configured for Multicast Underlay (PIMv6) Any-Source Multicast (ASM).
  - Cisco Nexus N9K-X9716D-GX and N9K-X9736C-FX line cards as VTEPs if the underlay uses Ingress Replication (IPv6).

Rendezvous Point for Tenant Routed Multicast

With TRM enabled Internal and External RP is supported. The following table displays the first release in which RP positioning is or is not supported.

	RP Internal	RP External	PIM-Based RP Everywhere
TRM L2 Mode	N/A	N/A	N/A

	RP Internal	RP External	PIM-Based RP Everywhere
TRM L3 Mode	7.0(3)I7(1), 9.2(x)	7.0(3)I7(4), 9.2(3)	Supported in 7.0(3)I7(x) releases starting from 7.0(3)I7(5) Not supported in 9.2(x) Supported in NX-OS releases beginning with 9.3(1) for the following Nexus 9000 switches: • Cisco Nexus 9300-FX/FX2 platform switches (excluding the Cisco Nexus 9300-FXP platform switch) Supported for Cisco Nexus 9300-FX3 platform switches beginning with Cisco NX-OS Release 9.3(5)
TRM L2L3 Mode	7.0(3)I7(1), 9.2(x)	N/A	N/A

	RP Internal	RP External
TRM L2 Mode	N/A	N/A
TRM L3 Mode	7.0(3)I7(1)	7.0(3)I7(4)
TRM L2L3 Mode	7.0(3)I7(1)	N/A

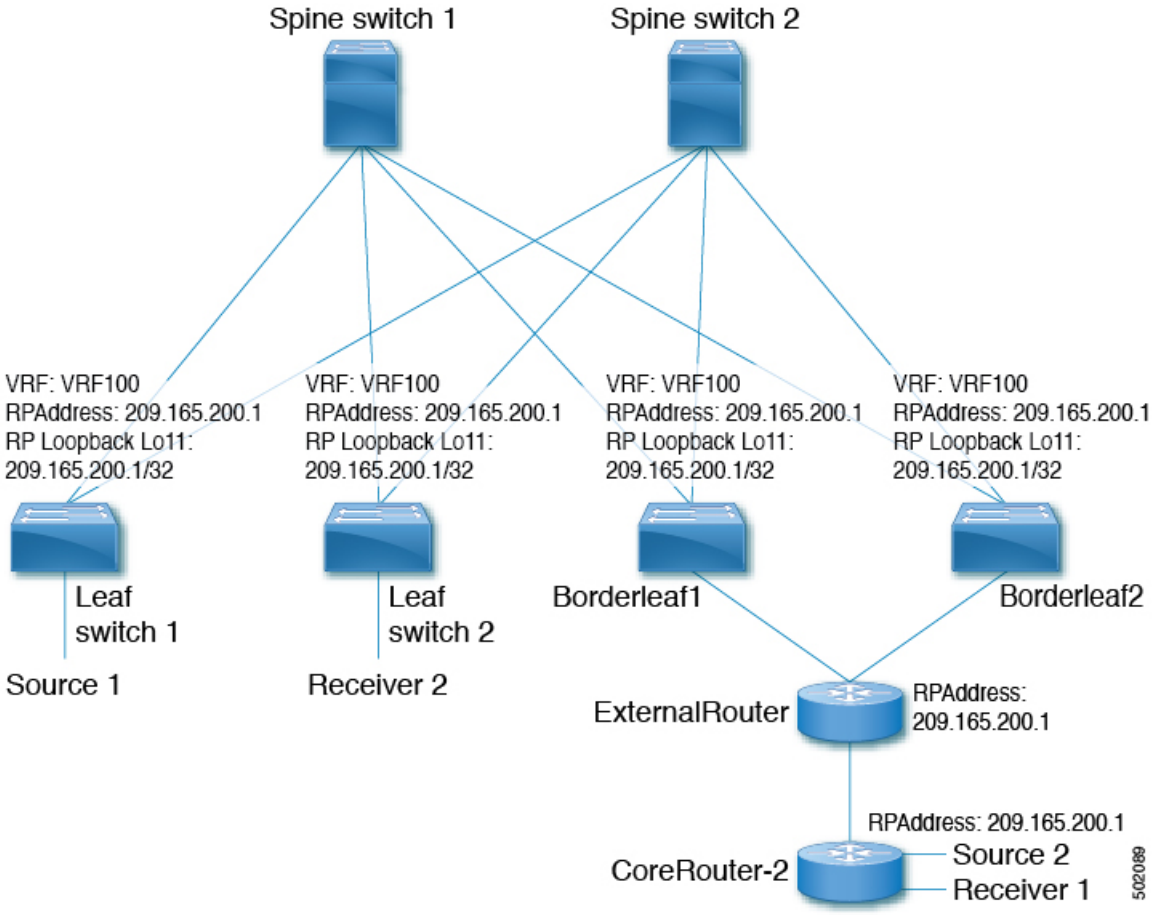
Configuring a Rendezvous Point for Tenant Routed Multicast

For Tenant Routed Multicast, the following rendezvous point options are supported:

- Configuring a Rendezvous Point Inside the VXLAN Fabric, on page 394
- Configuring an External Rendezvous Point, on page 395
- Configuring RP Everywhere with PIM Anycast, on page 397
- Configuring RP Everywhere with MSDP Peering, on page 403

Configuring a Rendezvous Point Inside the VXLAN Fabric

Configure the loopback for the TRM VRFs with the following commands on all devices (VTEP). Ensure it is reachable within EVPN (advertise/redistribute).



SUMMARY STEPS

1. **configure terminal**
2. **interface loopback** *loopback_number*
3. **vrf member** *vxlan-number*
4. **ip address** *ip-address*
5. **ip pim sparse-mode**
6. **vrf context** *vrf-name*
7. **ip pim rp-address** *ip-address-of-router* **group-list** *group-range-prefix*

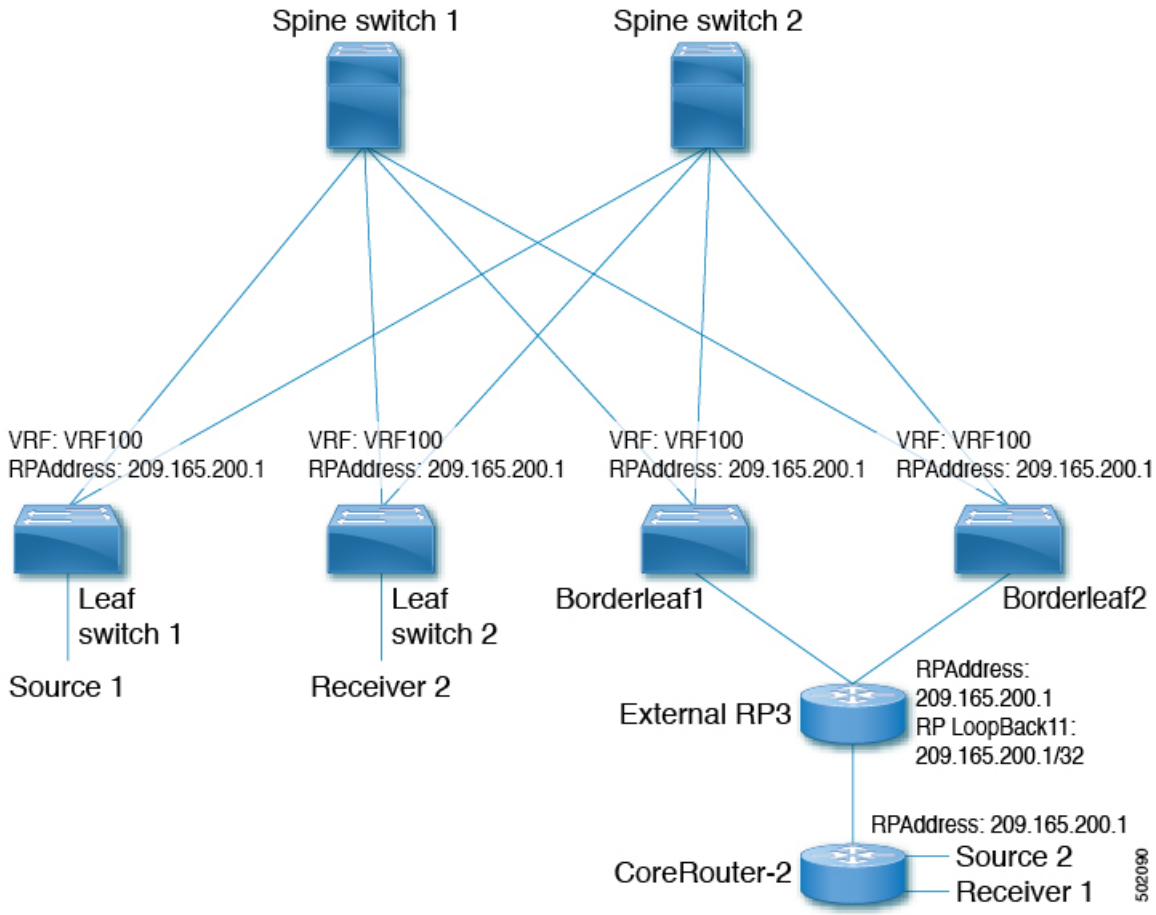
DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**interface loopback** *loopback_number* **Example:** `switch(config)# interface loopback 11`	Configure the loopback interface on all TRM-enabled nodes. This enables the rendezvous point inside the fabric.
Step 3	**vrf member** *vxlan-number* **Example:** `switch(config-if)# vrf member vrf100`	Configure VRF name.
Step 4	**ip address** *ip-address* **Example:** `switch(config-if)# ip address 209.165.200.1/32`	Specify IP address.
Step 5	**ip pim sparse-mode** **Example:** `switch(config-if)# ip pim sparse-mode`	Configure sparse-mode PIM on an interface.
Step 6	**vrf context** *vrf-name* **Example:** `switch(config-if)# vrf context vrf100`	Create a VXLAN tenant VRF.
Step 7	**ip pim rp-address** *ip-address-of-router* **group-list** *group-range-prefix* **Example:** `switch(config-vrf# ip pim rp-address 209.165.200.1 group-list 224.0.0.0/4`	The value of the *ip-address-of-router* parameter is that of the RP. The same IP address must be on all the edge devices (VTEPs) for a fully distributed RP.

Configuring an External Rendezvous Point

Configure the external rendezvous point (RP) IP address within the TRM VRFs on all devices (VTEP). In addition, ensure reachability of the external RP within the VRF via the border node. With TRM enabled and an external RP in use, ensure that only one routing path is active. Routing between the TRM fabric and the external RP must be via a single border leaf (non ECMP).



SUMMARY STEPS

1. **configure terminal**
2. **vrf context vrf100**
3. **ip pim rp-address** *ip-address-of-router* **group-list** *group-range-prefix*

DETAILED STEPS

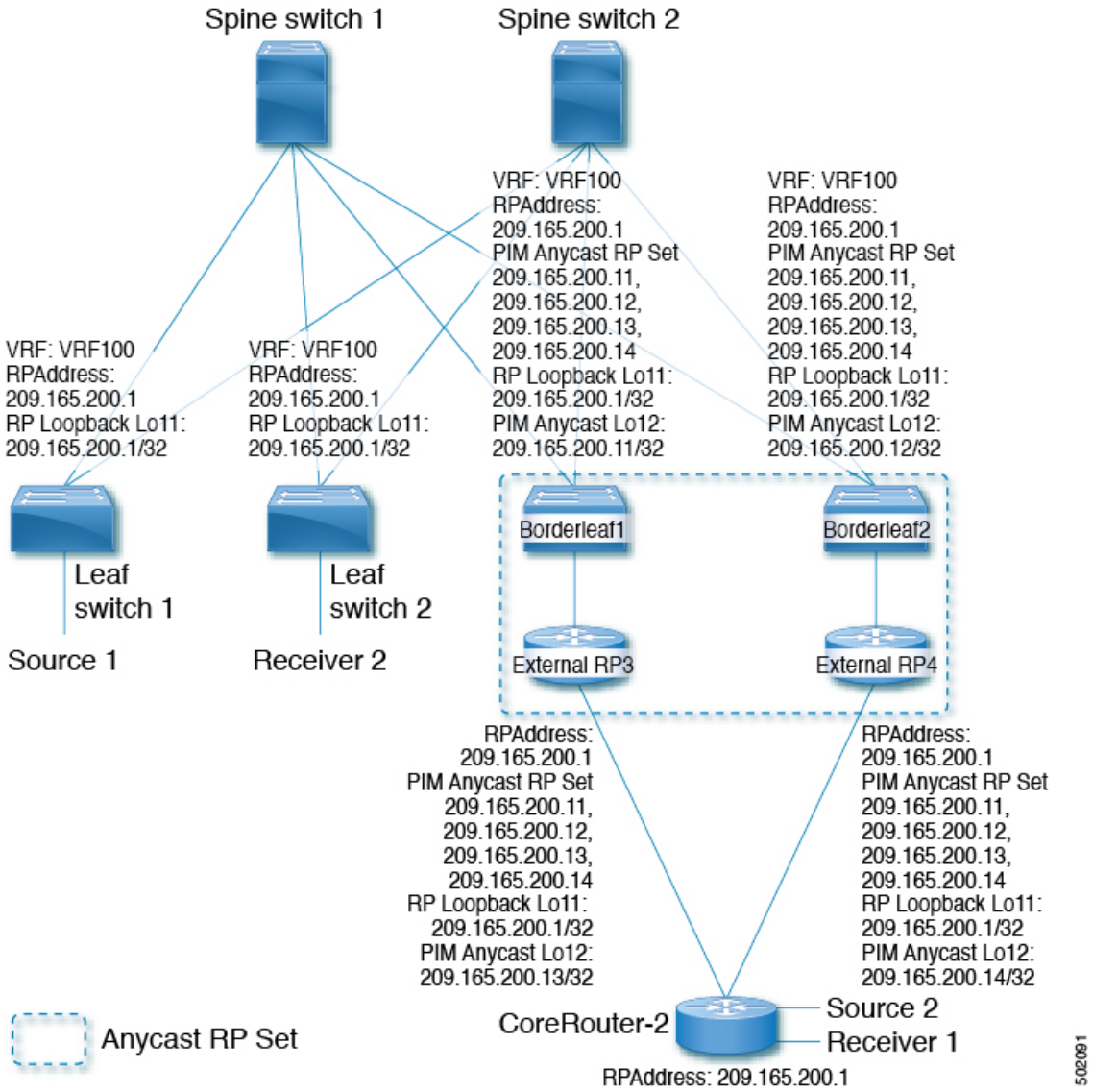
Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter configuration mode.
Step 2	**vrf context vrf100** **Example:** `switch(config)# vrf context vrf100`	Enter configuration mode.

	Command or Action	Purpose
Step 3	**ip pim rp-address** *ip-address-of-router* **group-list** *group-range-prefix* **Example:** `switch(config-vrf)# ip pim rp-address 209.165.200.1 group-list 224.0.0.0/4`	The value of the *ip-address-of-router* parameter is that of the RP. The same IP address must be on all of the edge devices (VTEPs) for a fully distributed RP.

Configuring RP Everywhere with PIM Anycast

RP Everywhere configuration with PIM Anycast solution.



For information about configuring RP Everywhere with PIM Anycast, see:

• Configuring a TRM Leaf Node for RP Everywhere with PIM Anycast, on page 398
• Configuring a TRM Border Leaf Node for RP Everywhere with PIM Anycast, on page 399
• Configuring an External Router for RP Everywhere with PIM Anycast, on page 401

Configuring a TRM Leaf Node for RP Everywhere with PIM Anycast

Configuration of Tenant Routed Multicast (TRM) leaf node for RP Everywhere.

SUMMARY STEPS

1. **configure terminal**
2. **interface loopback** *loopback_number*
3. **vrf member** *vrf-name*
4. **ip address** *ip-address*
5. **ip pim sparse-mode**
6. **vrf context** *vxlan*
7. **ip pim rp-address** *ip-address-of-router* **group-list** *group-range-prefix*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter configuration mode.
Step 2	**interface loopback** *loopback_number* **Example:** `switch(config)# interface loopback 11`	Configure the loopback interface on all VXLAN VTEP devices.
Step 3	**vrf member** *vrf-name* **Example:** `switch(config-if)# vrf member vrf100`	Configure VRF name.
Step 4	**ip address** *ip-address* **Example:** `switch(config-if)# ip address 209.165.200.1/32`	Specify IP address.
Step 5	**ip pim sparse-mode** **Example:** `switch(config-if)# ip pim sparse-mode`	Configure sparse-mode PIM on an interface.

	Command or Action	Purpose
Step 6	**vrf context** *vxlan* **Example:** `switch(config-if)# vrf context vrf100`	Create a VXLAN tenant VRF.
Step 7	**ip pim rp-address** *ip-address-of-router* **group-list** *group-range-prefix* **Example:** `switch(config-vrf# ip pim rp-address 209.165.200.1 group-list 224.0.0.0/4`	The value of the *ip-address-of-router* parameters is that of the RP. The same IP address must be on all the edge devices (VTEPs) for a fully distributed RP.

Configuring a TRM Border Leaf Node for RP Everywhere with PIM Anycast

Configuring the TRM Border Leaf Node for RP Anywhere with PIM Anycast.

SUMMARY STEPS

1. **configure terminal**
2. **ip pim evpn-border-leaf**
3. **interface loopback** *loopback_number*
4. **vrf member** *vrf-name*
5. **ip address** *ip-address*
6. **ip pim sparse-mode**
7. **interface loopback** *loopback_number*
8. **vrf member** *vxlan-number*
9. **ip address** *ip-address*
10. **ip pim sparse-mode**
11. **vrf context** *vrf-name*
12. **ip pim rp-address** *ip-address-of-router* **group-list** *group-range-prefix*
13. **ip pim anycast-rp** *anycast-rp-address address-of-rp*
14. **ip pim anycast-rp** *anycast-rp-address address-of-rp*
15. **ip pim anycast-rp** *anycast-rp-address address-of-rp*
16. **ip pim anycast-rp** *anycast-rp-address address-of-rp*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter configuration mode.

	Command or Action	Purpose
Step 2	**ip pim evpn-border-leaf** **Example:** `switch(config)# ip pim evpn-border-leaf`	Configure VXLAN VTEP as TRM border leaf node,
Step 3	**interface loopback** *loopback_number* **Example:** `switch(config)# interface loopback 11`	Configure the loopback interface on all VXLAN VTEP devices.
Step 4	**vrf member** *vrf-name* **Example:** `switch(config-if)# vrf member vrf100`	Configure VRF name.
Step 5	**ip address** *ip-address* **Example:** `switch(config-if)# ip address 209.165.200.1/32`	Specify IP address.
Step 6	**ip pim sparse-mode** **Example:** `switch(config-if)# ip pim sparse-mode`	Configure sparse-mode PIM on an interface.
Step 7	**interface loopback** *loopback_number* **Example:** `switch(config)# interface loopback 12`	Configure the PIM Anycast set RP loopback interface.
Step 8	**vrf member** *vxlan-number* **Example:** `switch(config-if)# vrf member vxlan-number`	Configure VRF name.
Step 9	**ip address** *ip-address* **Example:** `switch(config-if)# ip address 209.165.200.11/32`	Specify IP address.
Step 10	**ip pim sparse-mode** **Example:** `switch(config-if)# ip pim sparse-mode`	Configure sparse-mode PIM on an interface.
Step 11	**vrf context** *vrf-name* **Example:** `switch(config-if)# vrf context vrf100`	Create a VXLAN tenant VRF.
Step 12	**ip pim rp-address** *ip-address-of-router* **group-list** *group-range-prefix* **Example:**	The value of the *ip-address-of-router* parameters is that of the RP. The same IP address must be on all the edge devices (VTEPs) for a fully distributed RP.

	Command or Action	Purpose
	`switch(config-vrf)# ip pim rp-address 209.165.200.1 group-list 224.0.0.0/4`	
Step 13	**ip pim anycast-rp** *anycast-rp-address address-of-rp* **Example:** `switch(config-vrf)# ip pim anycast-rp 209.165.200.1 209.165.200.11`	Configure PIM Anycast RP set.
Step 14	**ip pim anycast-rp** *anycast-rp-address address-of-rp* **Example:** `switch(config-vrf)# ip pim anycast-rp 209.165.200.1 209.165.200.12`	Configure PIM Anycast RP set.
Step 15	**ip pim anycast-rp** *anycast-rp-address address-of-rp* **Example:** `switch(config-vrf)# ip pim anycast-rp 209.165.200.1 209.165.200.13`	Configure PIM Anycast RP set.
Step 16	**ip pim anycast-rp** *anycast-rp-address address-of-rp* **Example:** `switch(config-vrf)# ip pim anycast-rp 209.165.200.1 209.165.200.14`	Configure PIM Anycast RP set.

Configuring an External Router for RP Everywhere with PIM Anycast

Use this procedure to configure an external router for RP Everywhere.

SUMMARY STEPS

1. **configure terminal**
2. **interface loopback** *loopback_number*
3. **vrf member** *vrf-name*
4. **ip address** *ip-address*
5. **ip pim sparse-mode**
6. **interface loopback** *loopback_number*
7. **vrf member** *vxlan-number*
8. **ip address** *ip-address*
9. **ip pim sparse-mode**
10. **vrf context** *vxlan*
11. **ip pim rp-address** *ip-address-of-router* **group-list** *group-range-prefix*
12. **ip pim anycast-rp** *anycast-rp-address address-of-rp*
13. **ip pim anycast-rp** *anycast-rp-address address-of-rp*
14. **ip pim anycast-rp** *anycast-rp-address address-of-rp*
15. **ip pim anycast-rp** *anycast-rp-address address-of-rp*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter configuration mode.
Step 2	**interface loopback** *loopback_number* **Example:** `switch(config)# interface loopback 11`	Configure the loopback interface on all VXLAN VTEP devices.
Step 3	**vrf member** *vrf-name* **Example:** `switch(config-if)# vrf member vfr100`	Configure VRF name.
Step 4	**ip address** *ip-address* **Example:** `switch(config-if)# ip address 209.165.200.1/32`	Specify IP address.
Step 5	**ip pim sparse-mode** **Example:** `switch(config-if)# ip pim sparse-mode`	Configure sparse-mode PIM on an interface.
Step 6	**interface loopback** *loopback_number* **Example:** `switch(config)# interface loopback 12`	Configure the PIM Anycast set RP loopback interface.
Step 7	**vrf member** *vxlan-number* **Example:** `switch(config-if)# vrf member vrf100`	Configure VRF name.
Step 8	**ip address** *ip-address* **Example:** `switch(config-if)# ip address 209.165.200.13/32`	Specify IP address.
Step 9	**ip pim sparse-mode** **Example:** `switch(config-if)# ip pim sparse-mode`	Configure sparse-mode PIM on an interface.
Step 10	**vrf context** *vxlan* **Example:** `switch(config-if)# vrf context vrf100`	Create a VXLAN tenant VRF.

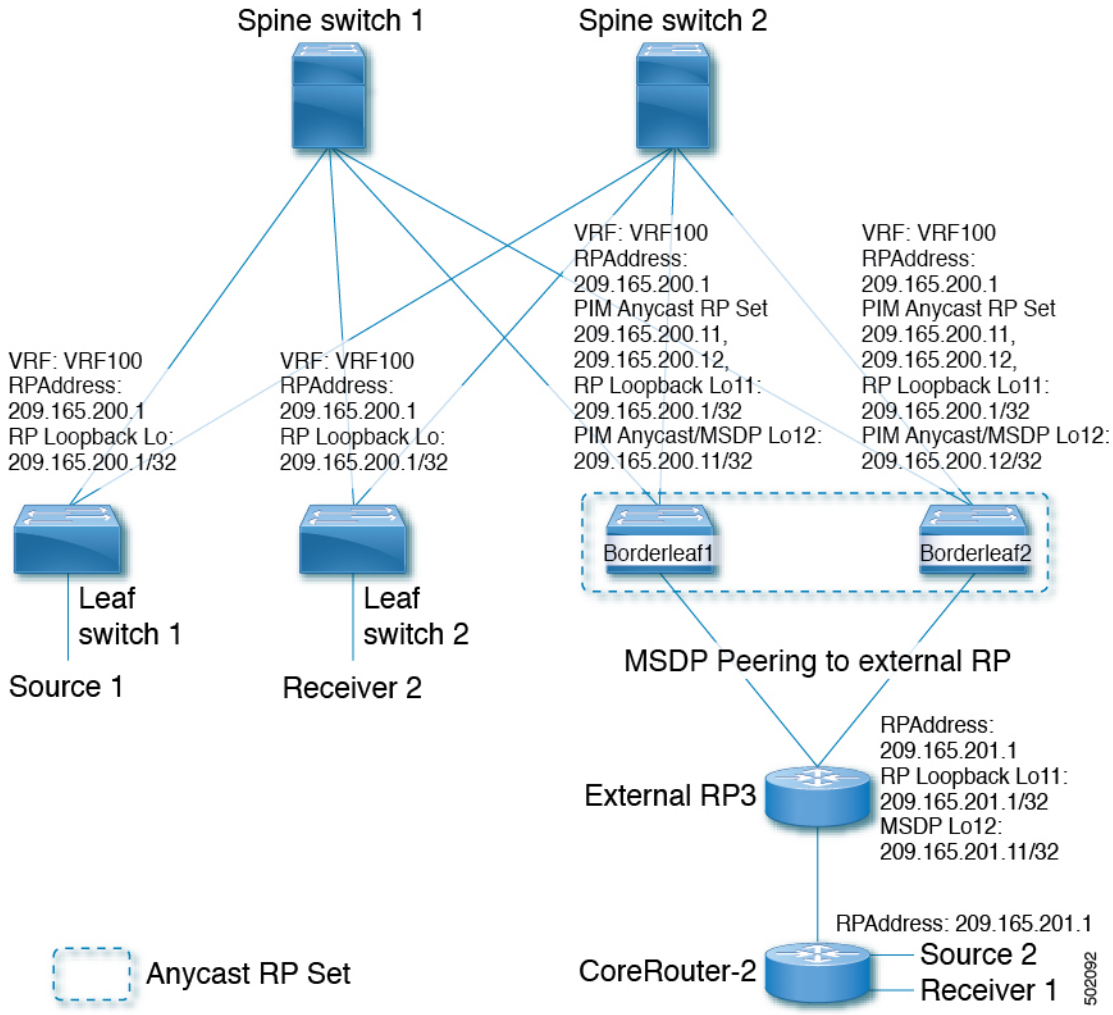
	Command or Action	Purpose
Step 11	**ip pim rp-address** *ip-address-of-router* **group-list** *group-range-prefix* **Example:** `switch(config-vrf)# ip pim rp-address 209.165.200.1 group-list 224.0.0.0/4`	The value of the *ip-address-of-router* parameters is that of the RP. The same IP address must be on all the edge devices (VTEPs) for a fully distributed RP.
Step 12	**ip pim anycast-rp** *anycast-rp-address address-of-rp* **Example:** `switch(config-vrf)# ip pim anycast-rp 209.165.200.1 209.165.200.11`	Configure PIM Anycast RP set.
Step 13	**ip pim anycast-rp** *anycast-rp-address address-of-rp* **Example:** `switch(config-vrf)# ip pim anycast-rp 209.165.200.1 209.165.200.12`	Configure PIM Anycast RP set.
Step 14	**ip pim anycast-rp** *anycast-rp-address address-of-rp* **Example:** `switch(config-vrf)# ip pim anycast-rp 209.165.200.1 209.165.200.13`	Configure PIM Anycast RP set.
Step 15	**ip pim anycast-rp** *anycast-rp-address address-of-rp* **Example:** `switch(config-vrf)# ip pim anycast-rp 209.165.200.1 209.165.200.14`	Configure PIM Anycast RP set.

Configuring RP Everywhere with MSDP Peering

The following figure represents the RP Everywhere configuration with MSDP RP solution.

For information about configuring RP Everywhere with MSDP Peering, see:

- Configuring a TRM Leaf Node for RP Everywhere with MSDP Peering, on page 404
- Configuring a TRM Border Leaf Node for RP Everywhere with MSDP Peering, on page 405
- Configuring an External Router for RP Everywhere with MSDP Peering, on page 408



Configuring a TRM Leaf Node for RP Everywhere with MSDP Peering

Configuring a TRM leaf node for RP Everywhere with MSDP peering.

SUMMARY STEPS

1. **configure terminal**
2. **interface loopback** *loopback_number*
3. **vrf member** *vrf-name*
4. **ip address** *ip-address*
5. **ip pim sparse-mode**
6. **vrf context** *vrf-name*
7. **ip pim rp-address** *ip-address-of-router* **group-list** *group-range-prefix*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter configuration mode.
Step 2	**interface loopback** *loopback_number* **Example:** `switch(config)# interface loopback 11`	Configure the loopback interface on all VXLAN VTEP devices.
Step 3	**vrf member** *vrf-name* **Example:** `switch(config-if)# vrf member vrf100`	Configure VRF name.
Step 4	**ip address** *ip-address* **Example:** `switch(config-if)# ip address 209.165.200.1/32`	Specify IP address.
Step 5	**ip pim sparse-mode** **Example:** `switch(config-if)# ip pim sparse-mode`	Configure sparse-mode PIM on an interface.
Step 6	**vrf context** *vrf-name* **Example:** `switch(config-if)# vrf context vrf100`	Create a VXLAN tenant VRF.
Step 7	**ip pim rp-address** *ip-address-of-router* **group-list** *group-range-prefix* **Example:** `switch(config-vrf# ip pim rp-address 209.165.200.1 group-list 224.0.0.0/4`	The value of the *ip-address-of-router* parameters is that of the RP. The same IP address must be on all the edge devices (VTEPs) for a fully distributed RP.

Configuring a TRM Border Leaf Node for RP Everywhere with MSDP Peering

Use this procedure to configure a TRM border leaf for RP Everywhere with PIM Anycast.

SUMMARY STEPS

1. **configure terminal**
2. **feature msdp**
3. **ip pim evpn-border-leaf**
4. **interface loopback** *loopback_number*

5. **vrf member** *vrf-name*
6. **ip address** *ip-address*
7. **ip pim sparse-mode**
8. **interface loopback** *loopback_number*
9. **vrf member** *vrf-name*
10. **ip address** *ip-address*
11. **ip pim sparse-mode**
12. **vrf context** *vrf-name*
13. **ip pim rp-address** *ip-address-of-router* **group-list** *group-range-prefix*
14. **ip pim anycast-rp** *anycast-rp-address address-of-rp*
15. **ip pim anycast-rp** *anycast-rp-address address-of-rp*
16. **ip msdp originator-id** *loopback*
17. **ip msdp peer** *ip-address* **connect-source** *loopback*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter configuration mode.
Step 2	**feature msdp** **Example:** `switch(config)# feature msdp`	Enable feature MSDP.
Step 3	**ip pim evpn-border-leaf** **Example:** `switch(config)# ip pim evpn-border-leaf`	Configure VXLAN VTEP as TRM border leaf node,
Step 4	**interface loopback** *loopback_number* **Example:** `switch(config)# interface loopback 11`	Configure the loopback interface on all VXLAN VTEP devices.
Step 5	**vrf member** *vrf-name* **Example:** `switch(config-if)# vrf member vrf100`	Configure VRF name.
Step 6	**ip address** *ip-address* **Example:** `switch(config-if)# ip address 209.165.200.1/32`	Specify IP address.
Step 7	**ip pim sparse-mode** **Example:**	Configure sparse-mode PIM on an interface.

	Command or Action	Purpose
	`switch(config-if)# ip pim sparse-mode`	
Step 8	**interface loopback** *loopback_number* **Example:** `switch(config)# interface loopback 12`	Configure the PIM Anycast set RP loopback interface.
Step 9	**vrf member** *vrf-name* **Example:** `switch(config-if)# vrf member vrf100`	Configure VRF name.
Step 10	**ip address** *ip-address* **Example:** `switch(config-if)# ip address 209.165.200.11/32`	Specify IP address.
Step 11	**ip pim sparse-mode** **Example:** `switch(config-if)# ip pim sparse-mode`	Configure sparse-mode PIM on an interface.
Step 12	**vrf context** *vrf-name* **Example:** `switch(config-if)# vrf context vrf100`	Create a VXLAN tenant VRF.
Step 13	**ip pim rp-address** *ip-address-of-router* **group-list** *group-range-prefix* **Example:** `switch(config-vrf)# ip pim rp-address 209.165.200.1 group-list 224.0.0.0/4`	The value of the *ip-address-of-router* parameter is that of the RP. The same IP address must be on all the edge devices (VTEPs) for a fully distributed RP.
Step 14	**ip pim anycast-rp** *anycast-rp-address address-of-rp* **Example:** `switch(config-vrf)# ip pim anycast-rp 209.165.200.1 209.165.200.11`	Configure PIM Anycast RP set.
Step 15	**ip pim anycast-rp** *anycast-rp-address address-of-rp* **Example:** `switch(config-vrf)# ip pim anycast-rp 209.165.200.1 209.165.200.12`	Configure PIM Anycast RP set.
Step 16	**ip msdp originator-id** *loopback* **Example:** `switch(config-vrf)# ip msdp originator-id loopback12`	Configure MSDP originator ID.
Step 17	**ip msdp peer** *ip-address* **connect-source** *loopback* **Example:**	Configure MSDP peering between border node and external RP router.

	Command or Action	Purpose
	`switch(config-vrf)# ip msdp peer 209.165.201.11 connect-source loopback12`	

Configuring an External Router for RP Everywhere with MSDP Peering

SUMMARY STEPS

1. **configure terminal**
2. **feature msdp**
3. **interface loopback** *loopback_number*
4. **vrf member** *vrf-name*
5. **ip address** *ip-address*
6. **ip pim sparse-mode**
7. **interface loopback** *loopback_number*
8. **vrf member** *vrf-name*
9. **ip address** *ip-address*
10. **ip pim sparse-mode**
11. **vrf context** *vrf-name*
12. **ip pim rp-address** *ip-address-of-router* **group-list** *group-range-prefix*
13. **ip msdp originator-id loopback12**
14. **ip msdp peer** *ip-address* **connect-source loopback12**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter configuration mode.
Step 2	**feature msdp** **Example:** `switch(config)# feature msdp`	Enable feature MSDP.
Step 3	**interface loopback** *loopback_number* **Example:** `switch(config)# interface loopback 11`	Configure the loopback interface on all VXLAN VTEP devices.
Step 4	**vrf member** *vrf-name* **Example:** `switch(config-if)# vrf member vrf100`	Configure VRF name.

	Command or Action	Purpose
Step 5	**ip address** *ip-address* **Example:** `switch(config-if)# ip address 209.165.201.1/32`	Specify IP address.
Step 6	**ip pim sparse-mode** **Example:** `switch(config-if)# ip pim sparse-mode`	Configure sparse-mode PIM on an interface.
Step 7	**interface loopback** *loopback_number* **Example:** `switch(config)# interface loopback 12`	Configure the PIM Anycast set RP loopback interface.
Step 8	**vrf member** *vrf-name* **Example:** `switch(config-if)# vrf member vrf100`	Configure VRF name.
Step 9	**ip address** *ip-address* **Example:** `switch(config-if)# ip address 209.165.201.11/32`	Specify IP address.
Step 10	**ip pim sparse-mode** **Example:** `switch(config-if)# ip pim sparse-mode`	Configure sparse-mode PIM on an interface.
Step 11	**vrf context** *vrf-name* **Example:** `switch(config-if)# vrf context vrf100`	Create a VXLAN tenant VRF.
Step 12	**ip pim rp-address** *ip-address-of-router* **group-list** *group-range-prefix* **Example:** `switch(config-vrf)# ip pim rp-address 209.165.201.1 group-list 224.0.0.0/4`	The value of the *ip-address-of-router* parameters is that of the RP. The same IP address must be on all the edge devices (VTEPs) for a fully distributed RP.
Step 13	**ip msdp originator-id loopback12** **Example:** `switch(config-vrf)# ip msdp originator-id loopback12`	Configure MSDP originator ID.
Step 14	**ip msdp peer** *ip-address* **connect-source loopback12** **Example:** `switch(config-vrf)# ip msdp peer 209.165.200.11 connect-source loopback12`	Configure MSDP peering between external RP router and all TRM border nodes.

Configuring Layer 3 Tenant Routed Multicast

This procedure enables the Tenant Routed Multicast (TRM) feature. TRM operates primarily in the Layer 3 forwarding mode for IP multicast by using BGP MVPN signaling. TRM in Layer 3 mode is the main feature and the only requirement for TRM enabled VXLAN BGP EVPN fabrics. If non-TRM capable edge devices (VTEPs) are present, the Layer 2/Layer 3 mode and Layer 2 mode have to be considered for interop.

To forward multicast between senders and receivers on the Layer 3 cloud and the VXLAN fabric on TRM vPC border leafs, the VIP/PIP configuration must be enabled. For more information, see Configuring VIP/PIP.



Note TRM follows an always-route approach and hence decrements the Time to Live (TTL) of the transported IP multicast traffic.

Before you begin

VXLAN EVPN **feature nv overlay** and **nv overlay evpn** must be configured.

The rendezvous point (RP) must be configured.

To enable/disable TRM v4/v6, PIM v4/v6 must be enabled.

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter configuration mode.
Step 2	**feature ngmvpn** **Example:** `switch(config)# feature ngmvpn`	Enables the Next-Generation Multicast VPN (ngMVPN) control plane. New address family commands become available in BGP. **Note** The **no feature ngmvpn** command will not remove MVPN configuration under BGP. You will get a syslog message when you enable this command. The message informs you that **ip multicast multipath s-g-hash next-hop-based** is the recommended multipath hashing algorithm and you need enable it for the TRM enabled VRFs. The auto-generation of **ip multicast multipath s-g-hash next-hop-based** command does not happen after you enable the **feature ngmvpn** command. You need to configure **ip multicast multipath s-g-hash next-hop-based** as part of the VRF configuration.

	Command or Action	Purpose
Step 3	**ip igmp snooping vxlan** **Example:** `switch(config)# ip igmp snooping vxlan`	Configure IGMP snooping for VXLAN VLANs.
Step 4	**interface nve1** **Example:** `switch(config)# interface nve 1`	Configure the NVE interface.
Step 5	**member vni** *vni-range* **associate-vrf** **Example:** `switch(config-if-nve)# member vni 200100 associate-vrf`	Configure the Layer 3 virtual network identifier. The range of *vni-range* is from 1 to 16,777,214.
Step 6	**mcast-group** *ip-prefix* **Example:** `switch(config-if-nve-vni)# mcast-group 225.3.3.3`	Builds the default multicast distribution tree for the VRF VNI (Layer 3 VNI). The multicast group is used in the underlay (core) for all multicast routing within the associated Layer 3 VNI (VRF). **Note** We recommend that underlay multicast groups for Layer 2 VNI, default MDT, and data MDT not be shared. Use separate, non-overlapping groups.
Step 7	**exit** **Example:** `switch(config-if-nve-vni)# exit`	Exits command mode.
Step 8	**exit** **Example:** `switch(config-if)# exit`	Exits command mode.
Step 9	**router bgp <as-number>** **Example:** `switch(config)# router bgp 100`	Set autonomous system number.
Step 10	**neighbor** *ip-addr* **Example:** `switch(config-router)# neighbor 1.1.1.1`	Configure IP address of the neighbor.
Step 11	**address-family ipv4 mvpn** **Example:** `switch(config-router-neighbor)# address-family ipv4 mvpn`	Configure multicast VPN.

	Command or Action	Purpose
Step 12	**send-community extended** **Example:** `switch(config-router-neighbor-af)# send-community extended`	Enables ngMVPN for address family signalization. The **send community extended** command ensures that extended communities are exchanged for this address family.
Step 13	**exit** **Example:** `switch(config-router-neighbor-af)# exit`	Exits command mode.
Step 14	**exit** **Example:** `switch(config-router-neighbor)# exit`	Exits command mode.
Step 15	**vrf context** *vrf_name* **Example:** `switch(config-router)# vrf context vrf100`	Configures VRF name.
Step 16	**vni** *number* **Example:** `switch(config-router)# vni 500001 l3`	Specifies the VNI for the tenant VRF. Beginning with Cisco NX-OS Release 10.3(1)F, the **L3** keyword is provided to indicate that the new L3VNI configuration is enabled. Beginning with Cisco NX-OS Release 10.4(3)F, this command with **L3** option is supported on Cisco Nexus 9808/9804 switches with Cisco Nexus X9836DMA and X98900CD-A line cards.
Step 17	**mvpn vri id** <*id*> **Example:** `switch(config-router)#mvpn vri 100`	Generates the VRI for TRM. Run this command under router bgp <as-number> submode. The vri id range is from 1 to 65535. **Note** This command is mandatory on vPC leaf nodes, and value has to be same across vPC pair and unique in TRM domain. Also the value must not collide with any site-id value. **Note** This command is required on BGWs if site-id value is greater than 2 bytes, and value has to be same across all same site BGWs and unique in TRM domain. Also the value must not collide with any site-id value.
Step 18	**[no] mdt** [*v4*|*v6*] vxlan **Example:**	Enables TRM v4/v6 on the specified VRF. The TRM v4/v6 is enabled by default.

	Command or Action	Purpose
	`switch(config-router)#mdt v4 vxlan`	The **no** option disables the TRM v4/v6 on the specified VRF. Run this command under the sub-mode of new L3VNI config. **Note** This command is applicable only to VRFs configured with new-L3VNI.
Step 19	**ip multicast multipath s-g-hash next-hop-based** **Example:** `switch(config-vrf)# ip multicast multipath s-g-hash next-hop-based`	Configures multicast multipath and initiates S, G, nexthop hashing (rather than the default of S/RP, G-based hashing) to select the RPF interface.
Step 20	**ip pim rp-address** *ip-address-of-router* **group-list** *group-range-prefix* **Example:** `switch(config-vrf)# ip pim rp-address 209.165.201.1 group-list 226.0.0.0/8`	The value of the *ip-address-of-router* parameter is that of the RP. The same IP address must be on all of the edge devices (VTEPs) for a fully distributed RP. For overlay RP placement options, see the Configuring a Rendezvous Point for Tenant Routed Multicast, on page 393 section.
Step 21	**address-family ipv4 unicast** **Example:** `switch(config-vrf)# address-family ipv4 unicast`	Configures unicast address family.
Step 22	**route-target both auto mvpn** **Example:** `switch(config-vrf-af-ipv4)# route-target both auto mvpn`	Defines the BGP route target that is added as an extended community attribute to the customer multicast (C_Multicast) routes (ngMVPN route type 6 and 7). Auto route targets are constructed by the 2-byte Autonomous System Number (ASN) and Layer 3 VNI.
Step 23	**ip multicast overlay-spt-only** **Example:** `switch(config)# ip multicast overlay-spt-only`	Gratuitously originate (S,A) route when the source is locally connected. The **ip multicast overlay-spt-only** command is enabled by default on all MVPN-enabled Cisco Nexus 9000 Series switches (typically leaf node).
Step 24	**interface***vlan_id* **Example:** `switch(config)# interface vlan11`	Configures the first-hop gateway (distributed anycast gateway for the Layer 2 VNI. No router PIM peering must ever happen with this interface.
Step 25	**no shutdown** **Example:** `switch(config-if)# no shutdown`	Disables an interface.
Step 26	**vrf member** *vrf-num* **Example:** `switch(config-if)# vrf member vrf100`	Configures VRF name.

	Command or Action	Purpose
Step 27	**ipv6 address** *ipv6_address* **Example:** `switch(config-if)# ip address 11.1.1.1/24`	Configures IP address.
Step 28	**ipv6 pim sparse-mode** **Example:** `switch(config-if)# ip pim sparse-mode`	Enables IGMP and PIM on the SVI. This is required is multicast sources and/or receivers exist in this VLAN.
Step 29	**fabric forwarding mode anycast-gateway** **Example:** `switch(config-if)# fabric forwarding mode anycast-gateway`	Configures Anycast Gateway Forwarding Mode.
Step 30	**ip pim neighbor-policy** *route-map-name* **Example:** `switch(config-if)# ip pim neighbor-policy route-map1`	Creates an IP PIM neighbor policy with a suitable route-map to deny any IPv4 addresses, preventing PIM from establishing PIM neighborship on the L2VNI SVI. **Note** Do not use Distributed Anycast Gateway for PIM Peerings.
Step 31	**exit** **Example:** `switch(config-if)# exit`	Exits command mode.
Step 32	**interface** *vlan_id* **Example:** `switch(config)# interface vlan100`	Configures Layer 3 VNI.
Step 33	**no shutdown** **Example:** `switch(config-if)# no shutdown`	Disable an interface.
Step 34	**vrf member vrf100** **Example:** `switch(config-if)# vrf member vrf100`	Configures VRF name.
Step 35	**ip forward** **Example:** `switch(config-if)# ip forward`	Enable IP forwarding on interface.
Step 36	**ip pim sparse-mode** **Example:** `switch(config-if)# ip pim sparse-mode`	Configures sparse-mode PIM on interface. There is no PIM peering happening in the Layer-3 VNI, but this command must be present for forwarding.

Configuring TRM on the VXLAN EVPN Spine

This procedure enables Tenant Routed Multicast (TRM) on a VXLAN EVPN spine switch.

Before you begin

The VXLAN BGP EVPN spine must be configured. See Configuring iBGP for EVPN on the Spine, on page 139.

SUMMARY STEPS

1. **configure terminal**
2. **route-map permitall permit 10**
3. **set ip next-hop unchanged**
4. **exit**
5. **router bgp [autonomous system]** *number*
6. **address-family ipv4 mvpn**
7. **retain route-target all**
8. **neighbor** *ip-address* [**remote-as** *number*]
9. **address-family ipv4 mvpn**
10. **disable-peer-as-check**
11. **rewrite-rt-asn**
12. **send-community extended**
13. **route-reflector-client**
14. **route-map permitall out**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter configuration mode.
Step 2	**route-map permitall permit 10** **Example:** `switch(config)# route-map permitall permit 10`	Configure the route-map. **Note** The route-map keeps the next-hop unchanged for EVPN routes • Required for eBGP • Options for iBGP
Step 3	**set ip next-hop unchanged**	Set next hop address.

	Command or Action	Purpose
	Example: `switch(config-route-map)# set ip next-hop unchanged`	**Note** The route-map keeps the next-hop unchanged for EVPN routes • Required for eBGP • Options for iBGP
Step 4	**exit** **Example:** `switch(config-route-map)# exit`	Return to exec mode.
Step 5	**router bgp [autonomous system]** *number* **Example:** `switch(config)# router bgp 65002`	Specify BGP.
Step 6	**address-family ipv4 mvpn** **Example:** `switch(config-router)# address-family ipv4 mvpn`	Configure the address family IPv4 MVPN under the BGP.
Step 7	**retain route-target all** **Example:** `switch(config-router-af)# retain route-target all`	Configure retain route-target all under address-family IPv4 MVPN [global]. **Note** Required for eBGP. Allows the spine to retain and advertise all MVPN routes when there are no local VNIs configured with matching import route targets.
Step 8	**neighbor** *ip-address* [**remote-as** *number*] **Example:** `switch(config-router-af)# neighbor 100.100.100.1`	Define neighbor.
Step 9	**address-family ipv4 mvpn** **Example:** `switch(config-router-neighbor)# address-family ipv4 mvpn`	Configure address family IPv4 MVPN under the BGP neighbor.
Step 10	**disable-peer-as-check** **Example:** `switch(config-router-neighbor-af)# disable-peer-as-check`	Disables checking the peer AS number during route advertisement. Configure this parameter on the spine for eBGP when all leafs are using the same AS but the spines have a different AS than leafs. **Note** Required for eBGP.
Step 11	**rewrite-rt-asn** **Example:**	Normalizes the outgoing route target's AS number to match the remote AS number. Uses the BGP configured neighbors

	Command or Action	Purpose
	`switch(config-router-neighbor-af)# rewrite-rt-asn`	remote AS. The **rewrite-rt-asn** command is required if the route target auto feature is being used to configure EVPN route targets.
Step 12	**send-community extended** **Example:** `switch(config-router-neighbor-af)# send-community extended`	Configures community for BGP neighbors.
Step 13	**route-reflector-client** **Example:** `switch(config-router-neighbor-af)# route-reflector-client`	Configure route reflector. **Note** Required for iBGP with route-reflector.
Step 14	**route-map permitall out** **Example:** `switch(config-router-neighbor-af)# route-map permitall out`	Applies route-map to keep the next-hop unchanged. **Note** Required for eBGP.

Configuring Tenant Routed Multicast in Layer 2/Layer 3 Mixed Mode

This procedure enables the Tenant Routed Multicast (TRM) feature. This enables both Layer 2 and Layer 3 multicast BGP signaling. This mode is only necessary if non-TRM edge devices (VTEPs) are present in the Cisco Nexus 9000 Series switches (1st generation). Only the Cisco Nexus 9000-FX switches can do Layer 2/Layer 3 mode (Anchor-DR).

To forward multicast between senders and receivers on the Layer 3 cloud and the VXLAN fabric on TRM vPC border leafs, the VIP/PIP configuration must be enabled. For more information, see Configuring VIP/PIP.

All Cisco Nexus 9300-FX platform switches must be in Layer 2/Layer 3 mode.

Before you begin

VXLAN EVPN must be configured.

The rendezvous point (RP) must be configured.

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter configuration mode.

	Command or Action	Purpose
Step 2	**feature ngmvpn** **Example:** `switch(config)# feature ngmvpn`	Enables the Next-Generation Multicast VPN (ngMVPN) control plane. New address family commands become available in BGP. **Note** The **no feature ngmvpn** command will not remove MVPN configuration under BGP.
Step 3	**advertise evpn multicast** **Example:** `switch(config)# advertise evpn multicast`	Advertises IMET and SMET routes into BGP EVPN towards non-TRM capable switches.
Step 4	**ip igmp snooping vxlan** **Example:** `switch(config)# ip igmp snooping vxlan`	Configure IGMP snooping for VXLAN VLANs.
Step 5	**ip multicast overlay-spt-only** **Example:** `switch(config)# ip multicast overlay-spt-only`	Gratuitously originate (S,A) route when source is locally connected. The **ip multicast overlay-spt-only** command is enabled by default on all MVPN-enabled Cisco Nexus 9000 Series switches (typically leaf nodes).
Step 6	**ip multicast overlay-distributed-dr** **Example:** `switch(config)# ip multicast overlay-distributed-dr`	Enables distributed anchor DR function on this VTEP. **Note** The NVE interface must be shut and unshut while configuring this command.
Step 7	**interface nve1** **Example:** `switch(config)# interface nve 1`	Configure the NVE interface.
Step 8	**[no] shutdown** **Example:** `switch(config-if-nve)# shutdown`	Shuts down the NVE interface. The **no shutdown** command brings up the interface.
Step 9	**member vni** *vni-range* **associate-vrf** **Example:** `switch(config-if-nve)# member vni 200100 associate-vrf`	Configure the Layer 3 virtual network identifier. The range of *vni-range* is from 1 to 16,777,214.
Step 10	**mcast-group** *ip-prefix* **Example:** `switch(config-if-nve-vni)# mcast-group 225.3.3.3`	Configures the multicast group on distributed anchor DR.
Step 11	**exit** **Example:** `switch(config-if-nve-vni)# exit`	Exits command mode.

	Command or Action	Purpose
Step 12	**interface loopback** *loopback_number* **Example:** `switch(config-if-nve)# interface loopback 10`	Configure the loopback interface on all distributed anchor DR devices.
Step 13	**ip address** *ip_address* **Example:** `switch(config-if)# ip address 100.100.1.1/32`	Configure IP address. This IP address is the same on all distributed anchor DR.
Step 14	**ip router ospf** *process-tag* **area** *ospf-id* **Example:** `switch(config-if)# ip router ospf 100 area 0.0.0.0`	OSPF area ID in IP address format.
Step 15	**ip pim sparse-mode** **Example:** `switch(config-if)# ip pim sparse-mode`	Configure sparse-mode PIM on interface.
Step 16	**interface nve1** **Example:** `switch(config-if)# interface nve1`	Configure NVE interface.
Step 17	**shutdown** **Example:** `switch(config-if-nve)# shutdown`	Disable the interface.
Step 18	**mcast-routing override source-interface loopback** *int-num* **Example:** `switch(config-if-nve)# mcast-routing override source-interface loopback 10`	Enables that TRM is using a different loopback interface than the VTEPs default source-interface. The *loopback10* variable must be configured on every TRM-enabled VTEP (Anchor DR) in the underlay with the same IP address. This loopback and the respective **override** command are needed to serve TRM VTEPs in co-existence with non-TRM VTEPs.
Step 19	**exit** **Example:** `switch(config-if-nve)# exit`	Exits command mode.
Step 20	**router bgp 100** **Example:** `switch(config)# router bgp 100`	Set autonomous system number.
Step 21	**neighbor** *ip-addr* **Example:** `switch(config-router)# neighbor 1.1.1.1`	Configure IP address of the neighbor.

	Command or Action	Purpose
Step 22	**address-family ipv4 mvpn** **Example:** `switch(config-router-neighbor)# address-family ipv4 mvpn`	Configure multicast VPN.
Step 23	**send-community extended** **Example:** `switch(config-router-neighbor-af)# send-community extended`	Send community attribute.
Step 24	**exit** **Example:** `switch(config-router-neighbor-af)# exit`	Exits command mode.
Step 25	**exit** **Example:** `switch(config-router)# exit`	Exits command mode.
Step 26	**vrf** *vrf_name* **vrf100** **Example:** `switch(config)# vrf context vrf100`	Configure VRF name.
Step 27	**ip pim rp-address** *ip-address-of-router* **group-list** *group-range-prefix* **Example:** `switch(config-vrf)# ip pim rp-address 209.165.201.1 group-list 226.0.0.0/8`	The value of the *ip-address-of-router* parameter is that of the RP. The same IP address must be on all of the edge devices (VTEPs) for a fully distributed RP. For overlay RP placement options, see the Configuring a Rendezvous Point for Tenant Routed Multicast, on page 393 - Internal RP section.
Step 28	**address-family ipv4 unicast** **Example:** `switch(config-vrf)# address-family ipv4 unicast`	Configure unicast address family.
Step 29	**route-target both auto mvpn** **Example:** `switch(config-vrf-af-ipv4)# route-target both auto mvpn`	Specify target for mvpn routes.
Step 30	**exit** **Example:** `switch(config-vrf-af-ipv4)# exit`	Exits command mode.
Step 31	**exit** **Example:** `switch(config-vrf)# exit`	Exits command mode.

	Command or Action	Purpose
Step 32	**interface** *vlan_id* **Example:** `switch(config)# interface vlan11`	Configure Layer 2 VNI.
Step 33	**no shutdown** **Example:** `switch(config-if)# no shutdown`	Disable an interface.
Step 34	**vrf member vrf100** **Example:** `switch(config-if)# vrf member vrf100`	Configure VRF name.
Step 35	**ip address** *ip_address* **Example:** `switch(config-if)# ip address 11.1.1.1/24`	Configure IP address.
Step 36	**ip pim sparse-mode** **Example:** e `switch(config-if)# ip pim sparse-mode`	Configure sparse-mode PIM on the interface.
Step 37	**fabric forwarding mode anycast-gateway** **Example:** `switch(config-if)# fabric forwarding mode anycast-gateway`	Configure Anycast Gateway Forwarding Mode.
Step 38	**ip pim neighbor-policy** *route-map-name* **Example:** `switch(config-if)# ip pim neighbor-policy route-map1`	Creates an IP PIM neighbor policy with a suitable route-map to deny any IPv4 addresses, preventing PIM from establishing PIM neighborship on the L2VNI SVI.
Step 39	**exit** **Example:** `switch(config-if)# exit`	Exits command mode.
Step 40	**interface** *vlan_id* **Example:** `switch(config)# interface vlan100`	Configure Layer 3 VNI.
Step 41	**no shutdown** **Example:** `switch(config-if)# no shutdown`	Disable an interface.

	Command or Action	Purpose
Step 42	**vrf member vrf100** **Example:** `switch(config-if)# vrf member vrf100`	Configure VRF name.
Step 43	**ip forward** **Example:** `switch(config-if)# ip forward`	Enable IP forwarding on interface.
Step 44	**ip pim sparse-mode** **Example:** `switch(config-if)# ip pim sparse-mode`	Configure sparse-mode PIM on the interface.

Configuring VXLAN EVPN and TRM with IPv6 Multicast Underlay

Configuring IPv6 multicast underlay in the VXLAN fabric involves the following configurations:

Configuring L2-VNI Based Multicast Group in Underlay

Under NVE configuration on a leaf, IPv6 multicast group (IPv6) is configured for each L2-VNI (VLAN).

SUMMARY STEPS

1. **configure terminal**
2. **interface nve1**
3. **member vni** *vni*
4. **mcast-group** *ipv6-prefix*
5. **global mcast-group** *ipv6-multicast-group* **l2**
6. **exit**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**interface nve1** **Example:** `switch(config)# interface nve1`	Configures the NVE interface.

	Command or Action	Purpose
Step 3	**member vni** *vni* **Example:** `switch(config-if-nve)# member vni 10501`	Configures the Layer 2 virtual network identifier.
Step 4	**mcast-group** *ipv6-prefix* **Example:** `switch(config-if-nve-vni)# mcast-group ff04::40`	Builds the default multicast distribution tree for the Layer 2 VNI.
Step 5	**global mcast-group** *ipv6-multicast-group* **l2** **Example:** `switch(config-if-nve)# global mcast-group ff04::40 l2`	Configures the global multicast group for the Layer 2 VNI.
Step 6	**exit** **Example:** `switch(config-if-nve)# exit`	Exits configuration mode.

Configuring L3-VNI Based Multicast Group in Underlay

IPv6 multicast group (IPv6) is configured for each L3-VNI (VRF).

SUMMARY STEPS

1. **configure terminal**
2. **interface nve1**
3. **member vni** *vni* **associate-vrf**
4. **mcast-group** *ipv6-prefix*
5. **global mcast-group** *ipv6-multicast-group* **l3**
6. **exit**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**interface nve1** **Example:** `switch(config)# interface nve1`	Configures the NVE interface.

	Command or Action	Purpose
Step 3	**member vni** *vni* **associate-vrf** **Example:** `switch(config-if-nve)# member vni 50001 associate-vrf`	Associates L3VNI to VRF.
Step 4	**mcast-group** *ipv6-prefix* **Example:** `switch(config-if-nve-vni)# mcast-group ff10:0:0:1::1`	Builds the default multicast distribution tree for the Layer 3 VNI.
Step 5	**global mcast-group** *ipv6-multicast-group* **l3** **Example:** `switch(config-if-nve)# global mcast-group ff04::40 l3`	Configures the global multicast group for the Layer 3 VNI.
Step 6	**exit** **Example:** `switch(config-if-nve)# exit`	Exits configuration mode.

Enabling PIMv6 for Underlay

PIMv6 in and underlay is configured as follows:

SUMMARY STEPS

1. **configure terminal**
2. **interface loopback** *number*
3. **ipv6 address** *ipv6-prefix*
4. **ipv6 pim sparse-mode**
5. **interface nve1**
6. **source-interface loopback** *number*
7. **exit**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**interface loopback** *number* **Example:**	Configures an interface loopback. This example configures interface loopback 1.

	Command or Action	Purpose
	`switch(config)# interface loopback 1`	
Step 3	**ipv6 address** *ipv6-prefix* **Example:** `switch(config-if)# ipv6 address 11:0:0:1::1/128`	Configures an IP address for this interface. It should be a unique IP address that helps to identify this router.
Step 4	**ipv6 pim sparse-mode** **Example:** `switch(config-if)# ipv6 pim sparse-mode`	Enables PIM6 sparse mode.
Step 5	**interface nve1** **Example:** `switch(config-if)# interface nve1`	Configures the NVE interface.
Step 6	**source-interface loopback** *number* **Example:** `switch(config-if-nve)# source-interface loopback 1`	Configures an source interface loopback.
Step 7	**exit** **Example:** `switch(config-if-nve)# exit`	Exits configuration mode. **Note** For the PIMv6 configuration see the *Cisco Nexus 9000 Series NX-OS Multicast Routing Configuration Guide*. For the TRM configuration see the *Cisco Nexus 9000 Series NX-OS VXLAN Configuration Guide*.

Configuring Layer 2 Tenant Routed Multicast

This procedure enables the Tenant Routed Multicast (TRM) feature. This enables Layer 2 multicast BGP signaling.

IGMP Snooping Querier must be configured per multicast-enabled VXLAN VLAN on all Layer-2 TRM leaf switches.

Before you begin

VXLAN EVPN must be configured.

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter configuration mode.

	Command or Action	Purpose
Step 2	**feature ngmvpn** **Example:** `switch(config)# feature ngmvpn`	Enables EVPN/MVPN feature. **Note** The **no feature ngmvpn** command will not remove MVPN configuration under BGP.
Step 3	**advertise evpn multicast** **Example:** `switch(config)# advertise evpn multicast`	Advertise L2 multicast capability.
Step 4	**ip igmp snooping vxlan** **Example:** `switch(config)# ip igmp snooping vxlan`	Configure IGMP snooping for VXLANs.
Step 5	**vlan configuration** *vlan-id* **Example:** `switch(config)# vlan configuration 101`	Enter configuration mode for VLAN 101.
Step 6	**ip igmp snooping querier** *querier-ip-address* **Example:** `switch(config-vlan-config)# ip igmp snooping querier 2.2.2.2`	Configure IGMP snooping querier for each multicast-enabled VXLAN VLAN.

Configuring TRM with vPC Support

This section provides steps to configure TRM with vPC support. Beginning with Cisco NX-OS Release 10.1(2), TRM Multisite with vPC BGW is supported.

SUMMARY STEPS

1. **configure terminal**
2. **feature vpc**
3. **feature interface-vlan**
4. **feature lacp**
5. **feature pim**
6. **feature ospf**
7. **ip pim rp-address** *address* **group-list** *range*
8. **vpc domain** *domain-id*
9. **peer switch**
10. **peer gateway**
11. **peer-keepalive destination** *ipaddress*
12. **ip arp synchronize**
13. **ipv6 nd synchronize**
14. Create vPC peer-link.

15. **system nve infra-vlans** *range*
16. **vlan** *number*
17. Create the SVI.
18. (Optional) **delay restore interface-vlan** *seconds*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.
Step 2	**feature vpc** **Example:** `switch(config)# feature vpc`	Enables vPCs on the device.
Step 3	**feature interface-vlan** **Example:** `switch(config)# feature interface-vlan`	Enables the interface VLAN feature on the device.
Step 4	**feature lacp** **Example:** `switch(config)# feature lacp`	Enables the LACP feature on the device.
Step 5	**feature pim** **Example:** `switch(config)# feature pim`	Enables the PIM feature on the device.
Step 6	**feature ospf** **Example:** `switch(config)# feature ospf`	Enables the OSPF feature on the device.
Step 7	**ip pim rp-address** *address* **group-list** *range* **Example:** `switch(config)# ip pim rp-address 100.100.100.1 group-list 224.0.0/4`	Defines a PIM RP address for the underlay multicast group range.
Step 8	**vpc domain** *domain-id* **Example:** `switch(config)# vpc domain 1`	Creates a vPC domain on the device and enters vpn-domain configuration mode for configuration purposes. There is no default. The range is from 1 to 1000.
Step 9	**peer switch** **Example:**	Defines the peer switch.

	Command or Action	Purpose
	`switch(config-vpc-domain)# peer switch`	
Step 10	**peer gateway** **Example:** `switch(config-vpc-domain)# peer gateway`	To enable Layer 3 forwarding for packets destined to the gateway MAC address of the virtual port channel (vPC), use the **peer-gateway** command.
Step 11	**peer-keepalive destination** *ipaddress* **Example:** `switch(config-vpc-domain)# peer-keepalive destination 172.28.230.85`	Configures the IPv4 address for the remote end of the vPC peer-keepalive link. **Note** The system does not form the vPC peer link until you configure a vPC peer-keepalive link. The management ports and VRF are the defaults. **Note** We recommend that you configure a separate VRF and use a Layer 3 port from each vPC peer device in that VRF for the vPC peer-keepalive link. For more information about creating and configuring VRFs, see the Cisco Nexus 9000 NX-OS Series Unicast Routing Config Guide, 9.3(x).
Step 12	**ip arp synchronize** **Example:** `switch(config-vpc-domain)# ip arp synchronize`	Enables IP ARP synchronize under the vPC Domain to facilitate faster ARP table population following device reload.
Step 13	**ipv6 nd synchronize** **Example:** `switch(config-vpc-domain)# ipv6 nd synchronize`	Enables IPv6 nd synchronization under the vPC domain to facilitate faster nd table population following device reload.

Step 14	Create vPC peer-link.
	Example:
	`switch(config)# interface port-channel 1`
	`switch(config)# switchport`
	`switch(config)# switchport mode trunk`
	`switch(config)# switchport trunk allowed vlan 1,10,100-200`
	`switch(config)# mtu 9216`
	`switch(config)# vpc peer-link`
	`switch(config)# no shut`
	
	`switch(config)# interface Ethernet 1/1, 1/21`
	`switch(config)# switchport`
	`switch(config)# mtu 9216`
	`switch(config)# channel-group 1 mode active`
	`switch(config)# no shutdown`
		Creates the vPC peer-link port-channel interface and adds two member interfaces to it.

Step 15	**system nve infra-vlans** *range* **Example:**	Defines a non-VXLAN enabled VLAN as a backup routed path.

	Command or Action	Purpose
	`switch(config)# system nve infra-vlans 10`	
Step 16	**vlan** *number* **Example:** `switch(config)# vlan 10`	Creates the VLAN to be used as an infra-VLAN.

Step 17	Create the SVI.
	Example:
	`switch(config)# interface vlan 10`
	`switch(config)# ip address 10.10.10.1/30`
	`switch(config)# ip router ospf process UNDERLAY area 0`
	`switch(config)# ip pim sparse-mode`
	`switch(config)# no ip redirects`
	`switch(config)# mtu 9216`
	`switch(config)# no shutdown`
		Creates the SVI used for the backup routed path over the vPC peer-link.

Step 18	(Optional) **delay restore interface-vlan** *seconds* **Example:** `switch(config-vpc-domain)# delay restore interface-vlan 45`	Enables the delay restore timer for SVIs. We recommend tuning this value when the SVI/VNI scale is high. For example, when the SCI count is 1000, we recommend that you set the delay restore for **interface-vlan** to 45 seconds.

Configuring TRM with vPC Support (Cisco Nexus 9504-R and 9508-R)

SUMMARY STEPS

1. **configure terminal**
2. **feature vpc**
3. **feature interface-vlan**
4. **feature lacp**
5. **feature pim**
6. **feature ospf**
7. **ip pim rp-address** *address* **group-list** *range*
8. **vpc domain** *domain-id*
9. **hardware access-list tcam region mac-ifacl**
10. **hardware access-list tcam region vxlan 10**
11. **reload**
12. **peer switch**
13. **peer gateway**
14. **peer-keepalive destination** *ipaddress*
15. **ip arp synchronize**
16. **ipv6 nd synchronize**
17. Create vPC peer-link.

18. **system nve infra-vlans** *range*
19. **vlan** *number*
20. Create the SVI.
21. (Optional) **delay restore interface-vlan** *seconds*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.
Step 2	**feature vpc** **Example:** `switch(config)# feature vpc`	Enables vPCs on the device.
Step 3	**feature interface-vlan** **Example:** `switch(config)# feature interface-vlan`	Enables the interface VLAN feature on the device.
Step 4	**feature lacp** **Example:** `switch(config)# feature lacp`	Enables the LACP feature on the device.
Step 5	**feature pim** **Example:** `switch(config)# feature pim`	Enables the PIM feature on the device.
Step 6	**feature ospf** **Example:** `switch(config)# feature ospf`	Enables the OSPF feature on the device.
Step 7	**ip pim rp-address** *address* **group-list** *range* **Example:** `switch(config)# ip pim rp-address 100.100.100.1 group-list 224.0.0/4`	Defines a PIM RP address for the underlay multicast group range.
Step 8	**vpc domain** *domain-id* **Example:** `switch(config)# vpc domain 1`	Creates a vPC domain on the device and enters vpn-domain configuration mode for configuration purposes. There is no default. The range is 1–1000.
Step 9	**hardware access-list tcam region mac-ifacl** **Example:**	Carves the TCAM region for the ACL database. **Note**

	Command or Action	Purpose
	`switch(config)# hardware access-list tcam region mac-ifacl 0`	This TCAM carving command is required to enable TRM forwarding for N9K-X9636C-RX line cards only. With no TCAM region carved for **mac-ifacl**, the TCAM resources are used for TRM instead.
Step 10	**hardware access-list tcam region vxlan 10** **Example:** `switch(config)# hardware access-list tcam region vxlan 10`	Assigns the the TCAM region for use by a VXLAN. **Note** This TCAM carving command is required to enable TRM forwarding for N9K-X9636C-RX line cards only.
Step 11	**reload** **Example:** `switch(config)# reload`	Reloads the switch config for the TCAM assignments to become active.
Step 12	**peer switch** **Example:** `switch(config-vpc-domain)# peer switch`	Defines the peer switch.
Step 13	**peer gateway** **Example:** `switch(config-vpc-domain)# peer gateway`	To enable Layer 3 forwarding for packets that are destined to the gateway MAC address of the virtual port channel (vPC), use the **peer-gateway** command.
Step 14	**peer-keepalive destination** *ipaddress* **Example:** `switch(config-vpc-domain)# peer-keepalive destination 172.28.230.85`	Configures the IPv4 address for the remote end of the vPC peer-keepalive link. **Note** The system does not form the vPC peer link until you configure a vPC peer-keepalive link. The management ports and VRF are the defaults. **Note** We recommend that you configure a separate VRF and use a Layer 3 port from each vPC peer device in that VRF for the vPC peer-keepalive link. For more information about creating and configuring VRFs, see the Cisco Nexus 9000 NX-OS Series Unicast Routing Config Guide, 9.3(x).
Step 15	**ip arp synchronize** **Example:** `switch(config-vpc-domain)# ip arp synchronize`	Enables IP ARP synchronize under the vPC Domain to facilitate faster ARP table population following device reload.
Step 16	**ipv6 nd synchronize** **Example:** `switch(config-vpc-domain)# ipv6 nd synchronize`	Enables IPv6 and synchronization under the vPC domain to facilitate faster and table population following device reload.

	Command or Action	Purpose

Step 17	Create vPC peer-link.
	Example:
	`switch(config)# interface port-channel 1`
	`switch(config)# switchport`
	`switch(config)# switchport mode trunk`
	`switch(config)# switchport trunk allowed vlan 1,10,100-200`
	`switch(config)# mtu 9216`
	`switch(config)# vpc peer-link`
	`switch(config)# no shut`
	
	`switch(config)# interface Ethernet 1/1, 1/21`
	`switch(config)# switchport`
	`switch(config)# mtu 9216`
	`switch(config)# channel-group 1 mode active`
	`switch(config)# no shutdown`
		Creates the vPC peer-link port-channel interface and adds two member interfaces to it.

Step 18	**system nve infra-vlans** *range* **Example:** `switch(config)# system nve infra-vlans 10`	Defines a non-VXLAN enabled VLAN as a backup routed path.
Step 19	**vlan** *number* **Example:** `switch(config)# vlan 10`	Creates the VLAN to be used as an infra-VLAN.

Step 20	Create the SVI.
	Example:
	`switch(config)# interface vlan 10`
	`switch(config)# ip address 10.10.10.1/30`
	`switch(config)# ip router ospf process UNDERLAY area 0`
	`switch(config)# ip pim sparse-mode`
	`switch(config)# no ip redirects`
	`switch(config)# mtu 9216`
	`switch(config)# no shutdown`
		Creates the SVI used for the backup routed path over the vPC peer-link.

Step 21	(Optional) **delay restore interface-vlan** *seconds* **Example:** `switch(config-vpc-domain)# delay restore interface-vlan 45`	Enables the delay restore timer for SVIs. We recommend tuning this value when the SVI/VNI scale is high. For example, when the SCI count is 1000, we recommend that you set the delay restore for **interface-vlan** to 45 seconds.

Flex Stats for TRM

Beginning with Cisco NX-OS Release 10.3(1)F, the Real-time/flex statistics for TRM is supported for Overlay routes on Cisco Nexus 9300-X Cloud Scale Switches. Flex Stats is not supported for Underlay Routes

Note VXLAN NVE VNI ingress and egress, NVE per-peer ingress and tunnel tx stats won't be supported.

In a VXLAN TRM setup, if you want mroute statistics for overlay mroutes you must configure the **hardware profile multicast flex-stats-enable** command in the default template. For more information on configuration, see Configuring Flex Stats for TRM, on page 433.

The following CLIs will not be supported after the flex stats CLI is enabled:

- `sh nve vni <vni_id>/<all> counters`
- `sh nve peers <peer-ip> interface nve 1 counters`
- `sh int tunnel <Tunnel interface number> counters`

Configuring Flex Stats for TRM

This procedure enables/disables the flex stats counters in a VXLAN TRM setup.

SUMMARY STEPS

1. **configure terminal**
2. **[no] hardware profile multicast flex-stats-enable**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter configuration mode.
Step 2	**[no] hardware profile multicast flex-stats-enable** **Example:** `switch(config)# hardware profile multicast flex-stats-enable`	Enables the flex stats on TRM. The **no** option disables the flex stats on TRM. **Note** To reflect the changes done during configuration, ensure that the switch is reloaded.

Configuring TRM Data MDT

About TRM Data MDT

Tenant Routed Multicast (TRM) enables multicast forwarding on the VXLAN fabric that uses a BGP-based EVPN control plane. TRM provides multi-tenancy aware multicast forwarding between senders and receivers within the same or different subnet local to the VTEP or across VTEPs.

Existing TRM solution enables multicast forwarding using default Multicast Distribution Tree (default MDT). With default MDT, nodes (PEs) will always receive traffic in the underlay irrespective of whether they have interested receiver on the overlay.

The solution described in this document enables optimized multicast forwarding using S-PMSI (data MDT). With S-PMSI, source traffic will be encapsulated in a selective multicast tunnel. Only the leafs that have interested receivers will join the selective multicast distribution tree.

Switchover to Data MDT can be immediate or based on the traffic bandwidth (threshold based configuration).

Guidelines and Limitations for TRM Data MDT

TRM Data MDT has the following guidelines and limitations:

- Beginning with Cisco NX-OS Release 10.3(2)F, TRM Data MDT is supported on Cisco Nexus 9300 FX/FX2/FX3/GX/GX2 switches, and 9500 switches with X97160YC-EX, 9700-FX/GX line cards.
- Beginning with Cisco NX-OS Release 10.5(2)F, TRM Data MDT is supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.
- Beginning with Cisco NX-OS Release 10.4(1)F, TRM Data MDT is supported on Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, TRM Data MDT is supported on Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, TRM Data MDT is supported on Cisco Nexus 9364C-H1 switches.
- Data MDT in fabric is supported only with DCI IR for a given VRF. Data MDT in fabric is not supported with DCI Multicast for a given VRF on the site BGW.
- Data MDT configuration is VRF specific and configured under L3 VRF.
- The following TRM Data MDT features are supported:
  - ASM and SSM group ranges are supported for Data MDT. PIM-Bider Underlay is not supported for Data MDT.
  - Data MDT supports IPv4 and IPv6 overlay multicast traffic.
  - Data MDT will be supported by vPC, VMCT leaf's as well as vPC/Anycast BGW. Also, L2, L3 orphan/external network can be connected to vPC nodes.
  - Data MDT config per L3 VRF.
  - Data MDT origination (immediate and threshold based).

- Data MDT encap route programming delay of 3 seconds. User-defined delays are currently not supported.

- L2, L2 -L3 mixed mode will not be supported.
- New L3VNI mode is supported.
- Ensure that the total number of underlay groups (L2 BUM, default MDT, and data MDT groups) is 512.
- Within the same VRF, a route map can only exist with the same sequence number and threshold across multiple data MDT lines. Configuring different thresholds for the same route map by omitting the sequence number on the command is invalid and results in inconsistent configurations.

Configuring TRM Data MDT

Follow this procedure to configure TRM Data MDT:

Before you begin

To enable switching to data MDT group based on real-time flow rate, the following command is needed:

hardware profile multicast flex-stats-enable

Note This command requires switch reloading.

SUMMARY STEPS

1. **configure terminal**
2. **vrf context** *vrf-name*
3. **address-family {ipv4 | ipv6} unicast**
4. **[no] mdt data vxlan** <*group-range-1*> **[threshold]** [**route-map** <*value*> <*policy-name_1*>] [**seq** <*sequence-number*>]

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**vrf context** *vrf-name* **Example:** `switch(config)# vrf context vrf1`	Configures the VRF.

	Command or Action	Purpose

Step 3	**address-family {ipv4 | ipv6} unicast**
	Example:
	For IPv4
	`switch(config-vrf)# address-family ipv4 unicast`
	For IPv6
	`switch(config-vrf)# address-family ipv6 unicast`
		Configures the IPv4 or IPv6 unicast address family.

Step 4	**[no] mdt data vxlan** <*group-range-1*> **[threshold]** [**route-map** <*value*> <*policy-name_1*>] [**seq** <*sequence-number*>] **Example:** `switch(config-vrf-af)# mdt data vxlan 224.7.8.0/24 route-map map1 10`	Data MDT can be enabled/disabled per address family. Cisco Nexus supports overlapping group ranges between VRF as well as within the VRF between the address families. • Threshold & route-maps are optional. The traffic threshold is the traffic of the source and is measured in kbps. When the threshold is exceeded, the traffic takes 3 seconds to switch over to data MDT. • Group-range is part of the command key. More than one group range can be configured per address family. • BUM & default MDT group should not overlap with data MDT group. • Data MDT can have overlapping config range.

Verifying TRM Data MDT Configuration

To display the TRM Data MDT configuration information, enter one of the following commands:

Command	Purpose
show nve vni { <*vni-id*> **| all } mdt [{ local | remote | peer-sync }] [{** <*cs*> <*cg*>**} | {** <*cs6*> <*cg6*>**}]**	Displays customer source (CS), customer group (DS), data group (DG) mapping information.
show nve vrf [x] mdt [local | remote | peer-sync] [y] [z]	Displays CS, CG allocations under VRF.
show bgp ipv4 mvpn route-type 3 detail	Displays BGP S-PMSI route information for IP
show bgp ipv6 mvpn route-type 3 detail	Displays BGP S-PMSI route information for IP
show fabric multicast [ipv4 | ipv6] spmsi-ad-route [*Source Address*] [*Group address*] vrf <*vrf_name*>	Displays fabric multicast SPMSI-AD IPV4/IPv6 tenant VRF.
show ip mroute detail vrf <*vrf_name*>	Displays IP multicast route information for defa
show l2route spmsi {all | topology <*vlan*>}	Displays CS-CG to DS-DG mapping information programming).
show forwarding distribution multicast vxlan mdt-db	Displays MFDM/MFIB data MDT db.
show nve resource multicast	Displays the resource usage of data MDT and a

Configuring IGMP Snooping

Overview of IGMP Snooping Over VXLAN

By default, multicast traffic over VXLAN is flooded in the VNI/VLAN like any broadcast and unknown unicast traffic. With IGMP snooping enabled, each VTEP can snoop IGMP reports and only forward multicast traffic towards interested receivers.

The configuration of IGMP snooping is the same in VXLAN as in the configuration of IGMP snooping in a regular VLAN domain. For more information on IGMP snooping, see the *Configuring IGMP Snooping* section in the Cisco Nexus 9000 Series NX-OS Multicast Routing Configuration Guide, Release 7.x.

Guidelines and Limitations for IGMP Snooping Over VXLAN

See the following guidelines and limitations for IGMP snooping over VXLAN:

- IGMP snooping over VXLAN is not supported on VLANs with FEX member ports.
- IGMP snooping over VXLAN is supported with both IR and multicast underlay.
- IGMP snooping over VXLAN is supported in BGP EVPN topologies, not flood and learn topologies.
- Beginning with NX-OS release 9.3(3), IGMP Snooping over VXLAN is supported on the Cisco Nexus C93600CD-GX.
- Beginning with NX-OS release 9.3(3), IGMP Snooping over VXLAN is supported on the Cisco Nexus C9364C-GX.

Configuring IGMP Snooping Over VXLAN

SUMMARY STEPS

1. switch# **configure terminal**
2. switch(config)#**ip igmp snooping vxlan**
3. switch(config)#**ip igmp snooping disable-nve-static-router-port**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	switch# **configure terminal**	Enters global configuration mode.
Step 2	switch(config)#**ip igmp snooping vxlan**	Enables IGMP snooping for VXLAN VLANs. You have to explicitly configure this command to enable snooping for VXLAN VLANs.

	Command or Action	Purpose
Step 3	switch(config)#**ip igmp snooping disable-nve-static-router-port**	Configures IGMP snooping over VXLAN to not include NVE as static mrouter port using this global CLI command. IGMP snooping over VXLAN has the NVE interface as mrouter port by default.

Verifying VXLAN EVPN and TRM with IPv6 Multicast Underlay

Use the following show command to verify the status of the IPv6 Multicast Underlay configuration:

```
switch(config)# show run interface nve 1

!Command: show running-config interface nve1
!Running configuration last done at: Wed Jul  5 10:03:58 2023
!Time: Wed Jul  5 10:04:01 2023
version 10.3(99x) Bios:version 01.08

interface nve1
  no shutdown
  host-reachability protocol bgp
  source-interface loopback1
  member vni 10501
    mcast-group ff04::40
  member vni 50001 associate-vrf
    mcast-group ff10:0:0:1::1
```

Use the following show commands to verify the PIMv6 ASM configuration:

```
switch(config)# show ipv6 mroute
IPv6 Multicast Routing Table for VRF "default"

(*, ff04::40/128), uptime: 05:20:19, nve pim6 ipv6
  Incoming interface: Ethernet1/36, RPF nbr: fe80::23a:9cff:fe23:8367
  Outgoing interface list: (count: 1)
    nve1, uptime: 05:20:19, nve


(172:172:16:1::1/128, ff04::40/128), uptime: 05:20:19, nve m6rib pim6 ipv6
  Incoming interface: loopback1, RPF nbr: 172:172:16:1::1
  Outgoing interface list: (count: 2)
    Ethernet1/36, uptime: 01:47:03, pim6
    Ethernet1/27, uptime: 04:14:20, pim6


(*, ff10:0:0:1::10/128), uptime: 05:20:18, nve ipv6 pim6
  Incoming interface: Ethernet1/36, RPF nbr: fe80::23a:9cff:fe23:8367
  Outgoing interface list: (count: 1)
    nve1, uptime: 05:20:18, nve


(172:172:16:1::1/128, ff10:0:0:1::10/128), uptime: 05:20:18, nve m6rib ipv6 pim6
  Incoming interface: loopback1, RPF nbr: 172:172:16:1::1
  Outgoing interface list: (count: 2)
    Ethernet1/36, uptime: 04:04:35, pim6
    Ethernet1/27, uptime: 04:13:35, pim6
```

```
switch(config)# show ipv6 pim neighbor
PIM Neighbor Status for VRF "default"
Neighbor                  Interface            Uptime    Expires   DR       Bidir-  BFD
     ECMP Redirect
```

```
                                                                    Priority Capable State
   Capable
fe80::23a:9cff:fe28:5e07     Ethernet1/27         20:23:38  00:01:44  1        yes     n/a
     no
   Secondary addresses:
    27:50:1:1::2
```

```
switch(config)# show ipv6 pim rp
PIM RP Status Information for VRF "default"
BSR disabled
BSR RP Candidate policy: route-map1
BSR RP policy: route-map1

RP: 101:101:101:101::101, (0),
 uptime: 21:30:43   priority: 255,
 RP-source: (local),
 group ranges:
 ff00::/8
```

The following example provides the output for leaf switch BGP neighbor-1:

```
switch(config-if)# show ipv6 bgp neighbors

BGP neighbor is 33:52:1:1::2, remote AS 200, ebgp link, Peer index 3
  BGP version 4, remote router ID 172.17.1.1
  Neighbor previous state = OpenConfirm
  BGP state = Established, up for 00:00:16
  Neighbor vrf: default
  Peer is directly attached, interface Ethernet1/33
  Enable logging neighbor events
  Last read 0.926823, hold time = 3, keepalive interval is 1 seconds
  Last written 0.926319, keepalive timer expiry due 0.073338
  Received 23 messages, 0 notifications, 0 bytes in queue
  Sent 67 messages, 0 notifications, 0(0) bytes in queue
  Enhanced error processing: On
    0 discarded attributes
  Connections established 1, dropped 0
  Last update recd 00:00:15, Last update sent  = 00:00:15
   Last reset by us 00:08:45, due to session closed
  Last error length sent: 0
  Reset error value sent: 0
  Reset error sent major: 104 minor: 0
  Notification data sent:
  Last reset by peer never, due to No error
  Last error length received: 0
  Reset error value received 0
  Reset error received major: 0 minor: 0
  Notification data received:

  Neighbor capabilities:
  Dynamic capability: advertised (mp, refresh, gr) received (mp, refresh, gr)
  Dynamic capability (old): advertised received
  Route refresh capability (new): advertised received
  Route refresh capability (old): advertised received
  4-Byte AS capability: advertised received
  Address family IPv6 Unicast: advertised received
  Graceful Restart capability: advertised received

  Graceful Restart Parameters:
  Address families advertised to peer:
    IPv6 Unicast
  Address families received from peer:
    IPv6 Unicast
  Forwarding state preserved by peer for:
  Restart time advertised to peer: 400 seconds
```

```
  Stale time for routes advertised by peer: 300 seconds
  Restart time advertised by peer: 120 seconds
  Extended Next Hop Encoding Capability: advertised received
  Receive IPv6 next hop encoding Capability for AF:
    IPv4 Unicast  VPNv4 Unicast

  Message statistics:
                              Sent               Rcvd
  Opens:                        46                  1
  Notifications:                 0                  0
  Updates:                       2                  2
  Keepalives:                   18                 18
  Route Refresh:                 0                  0
  Capability:                    2                  2
  Total:                        67                 23
  Total bytes:                 521                538
  Bytes in queue:                0                  0

  For address family: IPv6 Unicast
  BGP table version 10, neighbor version 10
  3 accepted prefixes (3 paths), consuming 864 bytes of memory
  0 received prefixes treated as withdrawn
  2 sent prefixes (2 paths)
  Inbound soft reconfiguration allowed(always)
  Allow my ASN 3 times
  Last End-of-RIB received 00:00:01 after session start
  Last End-of-RIB sent 00:00:01 after session start
  First convergence 00:00:01 after session start with 2 routes sent

  Local host: 33:52:1:1::1, Local port: 179
  Foreign host: 33:52:1:1::2, Foreign port: 17226
  fd = 112
```

The following example provides the output for leaf switch BGP neighbor-2:

```
switch(config-if)# show bgp l2vpn evpn neighbors 172:17:1:1::1

BGP neighbor is 172:17:1:1::1, remote AS 200, ebgp link, Peer index 5
  BGP version 4, remote router ID 172.17.1.1
  Neighbor previous state = OpenConfirm
  BGP state = Established, up for 00:01:33
  Neighbor vrf: default
  Using loopback0 as update source for this peer
  Using iod 65 (loopback0) as update source
  Enable logging neighbor events
  External BGP peer might be up to 5 hops away
  Last read 0.933565, hold time = 3, keepalive interval is 1 seconds
  Last written 0.915927, keepalive timer expiry due 0.083742
  Received 105 messages, 0 notifications, 0 bytes in queue
  Sent 105 messages, 0 notifications, 0(0) bytes in queue
  Enhanced error processing: On
    0 discarded attributes
  Connections established 1, dropped 0
  Last update recd 00:01:32, Last update sent  = 00:01:32
   Last reset by us never, due to No error
  Last error length sent: 0
  Reset error value sent: 0
  Reset error sent major: 0 minor: 0
  Notification data sent:
  Last reset by peer never, due to No error
  Last error length received: 0
  Reset error value received 0
  Reset error received major: 0 minor: 0
  Notification data received:
```

```
Neighbor capabilities:
Dynamic capability: advertised (mp, refresh, gr) received (mp, refresh, gr)
Dynamic capability (old): advertised received
Route refresh capability (new): advertised received
Route refresh capability (old): advertised received
4-Byte AS capability: advertised received
Address family IPv4 MVPN: advertised received
Address family IPv6 MVPN: advertised received
Address family L2VPN EVPN: advertised received
Graceful Restart capability: advertised received

Graceful Restart Parameters:
Address families advertised to peer:
  IPv4 MVPN  IPv6 MVPN  L2VPN EVPN
Address families received from peer:
  IPv4 MVPN  IPv6 MVPN  L2VPN EVPN
Forwarding state preserved by peer for:
Restart time advertised to peer: 400 seconds
Stale time for routes advertised by peer: 300 seconds
Restart time advertised by peer: 120 seconds
Extended Next Hop Encoding Capability: advertised received
Receive IPv6 next hop encoding Capability for AF:
  IPv4 Unicast  VPNv4 Unicast

Message statistics:
                            Sent               Rcvd
Opens:                         1                  1
Notifications:                 0                  0
Updates:                       6                  3
Keepalives:                   95                 95
Route Refresh:                 0                  0
Capability:                    6                  6
Total:                       105                105
Total bytes:                2551               2047
Bytes in queue:                0                  0

For address family: IPv4 MVPN
BGP table version 3, neighbor version 3
0 accepted prefixes (0 paths), consuming 0 bytes of memory
0 received prefixes treated as withdrawn
0 sent prefixes (0 paths)
Community attribute sent to this neighbor
Extended community attribute sent to this neighbor
Allow my ASN 3 times
Outbound route-map configured is RN_NextHop_Unchanged, handle obtained
Last End-of-RIB received 00:00:01 after session start
Last End-of-RIB sent 00:00:01 after session start
First convergence 00:00:01 after session start with 0 routes sent

For address family: IPv6 MVPN
BGP table version 3, neighbor version 3
0 accepted prefixes (0 paths), consuming 0 bytes of memory
0 received prefixes treated as withdrawn
0 sent prefixes (0 paths)
Community attribute sent to this neighbor
Extended community attribute sent to this neighbor
Allow my ASN 3 times
Outbound route-map configured is RN_NextHop_Unchanged, handle obtained
Last End-of-RIB received 00:00:01 after session start
Last End-of-RIB sent 00:00:01 after session start
First convergence 00:00:01 after session start with 0 routes sent

For address family: L2VPN EVPN
BGP table version 7, neighbor version 7
```

```
0 accepted prefixes (0 paths), consuming 0 bytes of memory
0 received prefixes treated as withdrawn
4 sent prefixes (4 paths)
Community attribute sent to this neighbor
Extended community attribute sent to this neighbor
Allow my ASN 3 times
Advertise GW IP is enabled
Outbound route-map configured is RN_NextHop_Unchanged, handle obtained
Last End-of-RIB received 00:00:01 after session start
Last End-of-RIB sent 00:00:01 after session start
First convergence 00:00:01 after session start with 4 routes sent

Local host: 172:16:1:2::1, Local port: 21132
Foreign host: 172:17:1:1::1, Foreign port: 179
fd = 113
```

Example Configuration for VXLAN EVPN and TRM with IPv6 Multicast Underlay

In the following examples, the sample configuration for the leaf, spine, and RP are shown:

- **Leaf - Sample configuration of IPv6 multicast underlay:**
  - NVE Configuration

```
interface nve1
  no shutdown
  host-reachability protocol bgp
  source-interface loopback1
  member vni 10501
    mcast-group ff04::40
  member vni 50001 associate-vrf
    mcast-group ff10:0:0:1::1
```

  - PIMv6 Configuration

```
feature pim6

ipv6 pim rp-address 101:101:101:101::101 group-list ff00::/8

interface loopback1
  ipv6 address 172:172:16:1::1/128
  ipv6 pim sparse-mode

interface Ethernet1/27
  ipv6 address 27:50:1:1::1/64
  ospfv3 hello-interval 1
  ipv6 router ospfv3 v6u area 0.0.0.0
  ipv6 pim sparse-mode
  no shutdown
```

  - BGP Configuration

```
router bgp 100
    router-id 172.16.1.1
    address-family ipv4 unicast
      maximum-paths 64
      maximum-paths ibgp 64
    address-family ipv6 unicast
      maximum-paths 64
      maximum-paths ibgp 64
```

```
    address-family ipv4 mvpn
    address-family l2vpn evpn
    neighbor 172:17:1:1::1
      remote-as 100
      update-source loopback0
      address-family ipv4 mvpn
        send-community
        send-community extended
      address-family ipv6 mvpn
        send-community
        send-community extended
      address-family l2vpn evpn
        send-community
     neighbor 172:17:2:2::1
       remote-as 100
       update-source loopback0
       address-family ipv4 mvpn
         send-community
         send-community extended
       address-family ipv6 mvpn
         send-community
         send-community extended
       address-family l2vpn evpn
         send-community
         send-community extended
     vrf VRF1
       reconnect-interval 1
       address-family ipv4 unicast
         network 150.1.1.1/32
         advertise l2vpn evpn
         redistribute hmm route-map hmmAdv

evpn
  vni 10501 l2
    rd auto
    route-target import auto
    route-target export auto
vrf context VRF1
  vni 50001
rd auto
address-family ipv4 unicast
   route-target both auto
   route-target both auto mvpn
   route-target both auto evpn
address-family ipv6 unicast
   route-target both auto
   route-target both auto mvpn
   route-target both auto evpn

Note: Incase of vPC leafs, you need to configure identical "mvpn vri id" on both
the vPC nodes. For example:

router bgp 100
  mvpn vri id 2001
```



Note MVPN VRI ID must be unique within the network or setup. That is, if the network has three different sets of vPC pairs, each pair must have a different VRI ID.

- **Spine - sample configuration of IPv6 multicast underlay:**

- NVE Configuration

```
nv overlay evpn
```

- PIMv6 Configuration

```
feature pim6

ipv6 pim rp-address 101:101:101:101::101 group-list ff00::/8
ipv6 pim anycast-rp 101:101:101:101::101 102:102:102:102::102
ipv6 pim anycast-rp 101:101:101:101::101 103:103:103:103::103

interface loopback101
  ipv6 address 101:101:101:101::101/128
  ipv6 router ospfv3 v6u area 0.0.0.0
  ipv6 pim sparse-mode

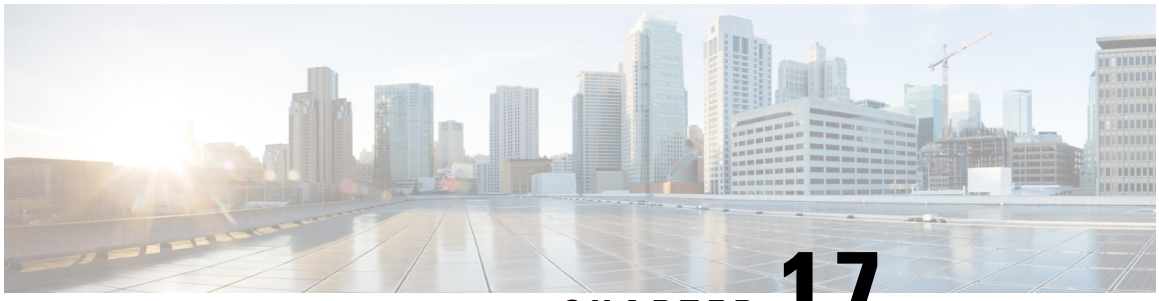
interface loopback102
  ipv6 address 102:102:102:102::102/128
  ipv6 router ospfv3 v6u area 0.0.0.0
  ipv6 pim sparse-mode

interface Ethernet1/50/1
  ipv6 address 27:50:1:1::2/64
  ipv6 pim sparse-mode
  no shutdown
```

- BGP Configuration

```
feature bgp

router bgp 100
        router-id 172.16.40.1
         address-family ipv4 mvpn
        address-family ipv6 mvpn
        address-family l2vpn evpn
        neighbor 172:16:1:1::1
          remote-as 100
          update-source loopback0
          address-family ipv4 mvpn
            send-community
            send-community extended
             route-reflector-client
          address-family ipv6 mvpn
           send-community
            send-community extended
             route-reflector-client
          address-family l2vpn evpn
          send-community
            send-community extended
             route-reflector-client
```



CHAPTER 17

Optimized Layer 2 Overlay Multicast

- Multicast Layer 2 EVPN Type-6 routes, on page 445
- Guidelines and Limitations of Multicast Layer 2 EVPN Type-6 routes, on page 450
- Configure Multicast Layer 2 EVPN Route Type 6, on page 452
- Verification of Multicast Layer 2 EVPN Route Types Configuration, on page 454
- Examples of Multicast Layer 2 EVPN Route Type 6, on page 455

Multicast Layer 2 EVPN Type-6 routes

A multicast Layer 2 EVPN Type-6 route is a BGP route type that:

- enables selective multicast traffic forwarding over a VXLAN fabric,
- advertises interest in specific multicast groups from VTEPs using SMET routes, and
- optimizes bandwidth by only replicating multicast traffic to VTEPs with interested receivers.

In data center topologies, many hosts can be attached to a VLAN or subnet, possibly spanning multiple PODs. In VXLAN-based data centers, VXLAN Tunnel Endpoints (VTEPs) are devices connected to the fabric that create VXLAN tunnels to send Layer 2 traffic between VLAN segments.

Sources located on a VTEP can send IP multicast (S,G) traffic toward receivers by encapsulating packets in VXLAN on the source VTEP and flooding to all remote VTEPs. Receiving VTEPs decapsulate and forward to interested receivers. However, indiscriminate flooding wastes bandwidth if no receivers are interested.

Type 6 EVPN Selective Multicast (SMET) Routes provide a solution to efficiently send IP multicast traffic between sources and receivers over the VXLAN fabric. Hosts attached to VTEPs express interest in multicast groups by sending membership reports on a VLAN; the VTEP snoops the IGMP reports and advertises a Type 6 EVPN SMET route to other remote VTEPs through BGP, enabling all VTEPs attached to the fabric to learn about receivers interested in specified multicast groups. VTEPs (leaf nodes) will only send IP multicast (S,G) traffic originated from locally connected sources over the VXLAN fabric if there is at least one receiver on a remote VTEP that is interested in the traffic, thereby optimizing the use of the fabric by not sending multicast traffic if there are no interested receivers. If ingress-replication is used in the underlay, the traffic can be further optimized by only sending multicast traffic to those VTEPs that have attached receivers.

RFC 9251 describes the procedures and NLRI formats for Type-6 EVPN SMET routes.

Note Cisco NX-OS does not currently support selective multicast delivery over a unicast underlay.

Functionality of Multicast Layer 2 EVPN Type-6 routes

Layer 2 EVPN Route Type-6 (SMET) routes for IPv4 (IGMP) routes have been supported in NX-OS as L2TRM**.

Beginning with Cisco NX-OS Release 10.5(3)F, the Multicast Layer 2 EVPN Route Type-6 will support the following functionality:

- IPv4 IGMP EVPN Type-6 SMET Routes using L2RIB instead of next-generation multicast VPN (NGMVPN).
- IPv6 MLD Snooping EVPN Type-6 SMET Routes.

Generating IGMP EVPN routes using L2RIB

Starting from Cisco NX-OS Release 10.5(3)F, added the following capabilities to Type-6 EVPN SMET route functionality.

- Multicast router or querier behind the EVPN domain, to support (*, *) route advertisement or receipt.
- Multiversion (IGMPv2/v3/v2+v3) advertisement capability

These capabilities are in compliance with RFC 9251.

Generating MLD Snooping SMET route

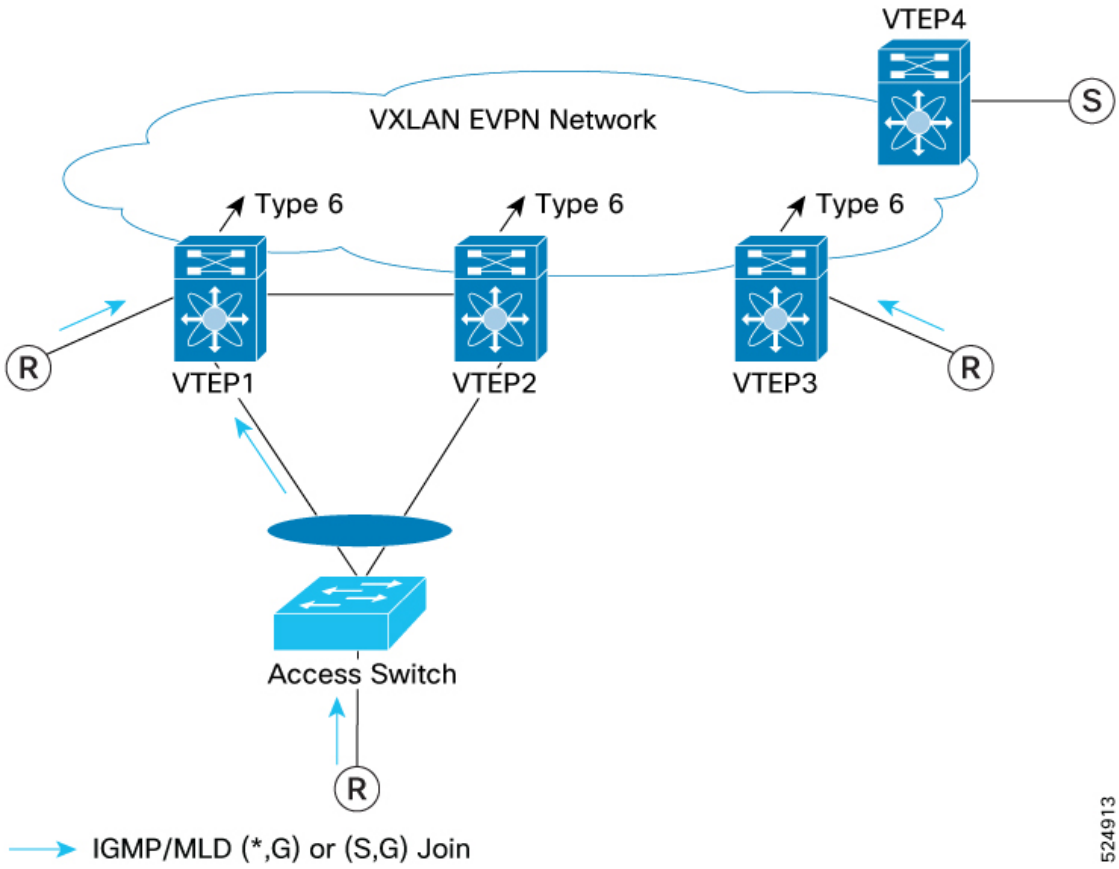
Before Cisco NX-OS Release 10.5(3)F, SMET route generation was limited to IPv4 IGMP snooped entries. Beginning with Cisco NX-OS Release 10.5(3)F, Type-6 SMET route advertisement for MLD snooped routes is supported where L2RIB and BGP will handle IPv4 or IPv6 SMET routes from IGMP/MLD on the transmit side and receive side, filtering and forwarding routes appropriately. Multicast router and multiversion capability (MLDv1/v2/v1+v2) will be supported.

Topology of Multicast Layer 2 EVPN Type-6 routes (Reference)

EVPN Route Type-6 will support vPC, vPC Fabric peering, ESI and non-vPC topologies.

Sample topologies and associated control plane flow for these topologies are as mentioned below.

Layer 2 EVPN Network with vPC



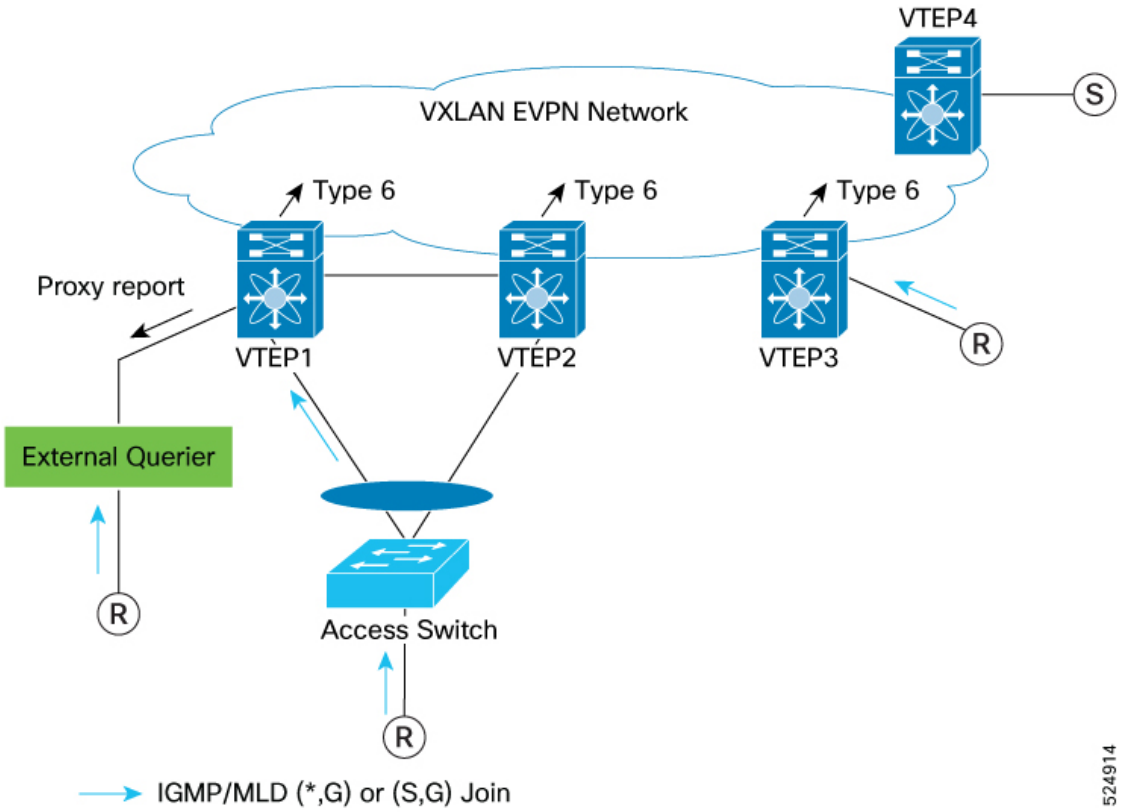
In this example topology, there are 4 VTEPs - a pair of vPC VTEPs (VTEP1 and VTEP2), and two standalone VTEPs (VTEP3 and VTEP4). Receivers (R) are attached to

- an orphan port behind VTEP1,
- access switch behind VPC complex VTEP1 and VTEP2, and
- standalone switch VTEP3.

When these receivers send (*, G) and/or (S,G) IGMP / MLD membership reports that are received by the VTEP1, VTEP2 and VTEP3, each VTEP with attached receiver originates a Type-6 EVPN SMET route type, which is received by all VTEPs, including VTEP4 to which the source is attached.

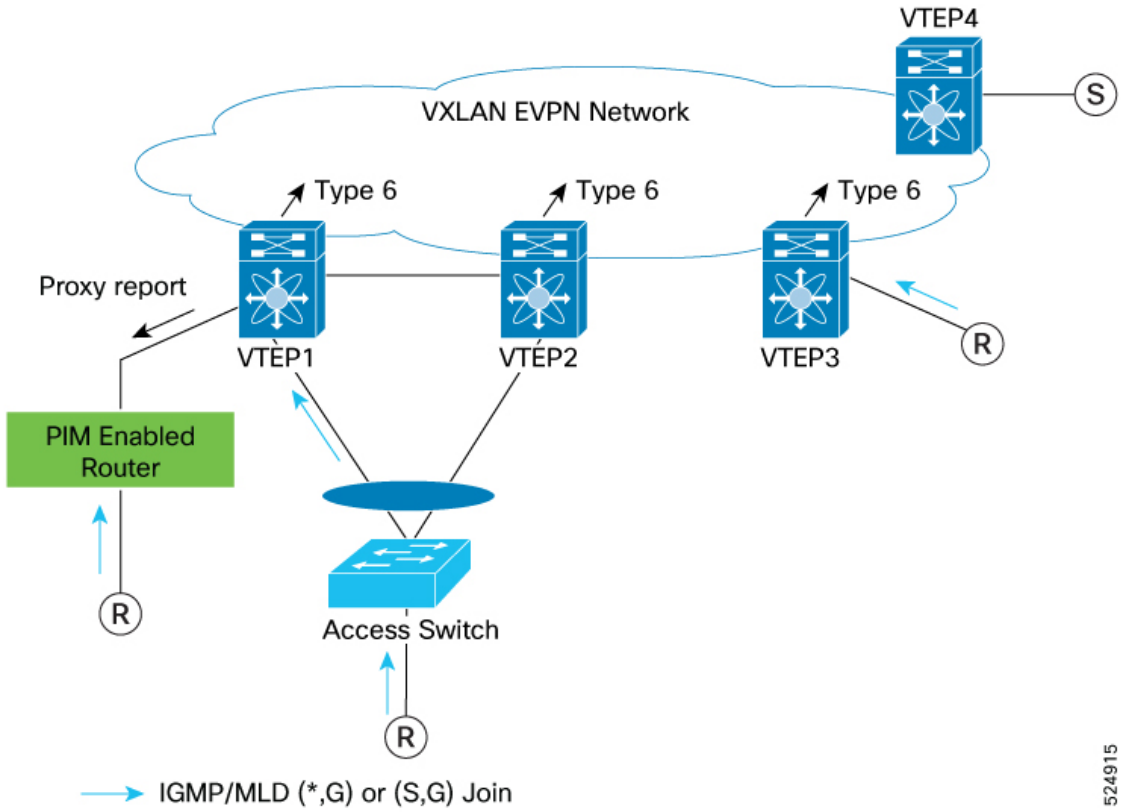
When the source ‘S’ behind VTEP4 sends multicast (S,G) traffic, the traffic is encapsulated in VXLAN by VTEP4 and sent over the fabric toward receivers; receiving VTEPs VTEP1, VTEP2 and VTEP3 decapsulate the VXLAN packet and forward toward local networks, as shown in the attached figure Note that VTEP4 will only forward multicast traffic if it receives at least one EVPN Type-6 EVPN SMET route from other VTEPs in the fabric.

Layer 2 EVPN Network with vPC and External Querier



In this example topology, there is an external querier attached to the orphan port, and an election is done between the VTEP1 and External Querier to determine the Querier for the network. For the case where the external querier wins the election, VTEP1 views the external querier as attached to an ‘mrouter’ port; VTEP1 advertises a (*,*) route toward the EVPN network. Any multicast traffic received from source ‘S’ is forwarded toward the external querier. Furthermore, reports received from VTEP3 over SMET are sent as a proxy report toward the querier. There are variants of the ‘external querier’ topology, where the querier is attached to a standalone node, or attached to the vPC complex, and the case where the VTEP wins the Querier election. These variants are supported, though the specific IGMP / MLD report and traffic flows depend on the specific querier location or querier election winner.

Layer 2 EVPN Network with vPC and PIM Enabled Router



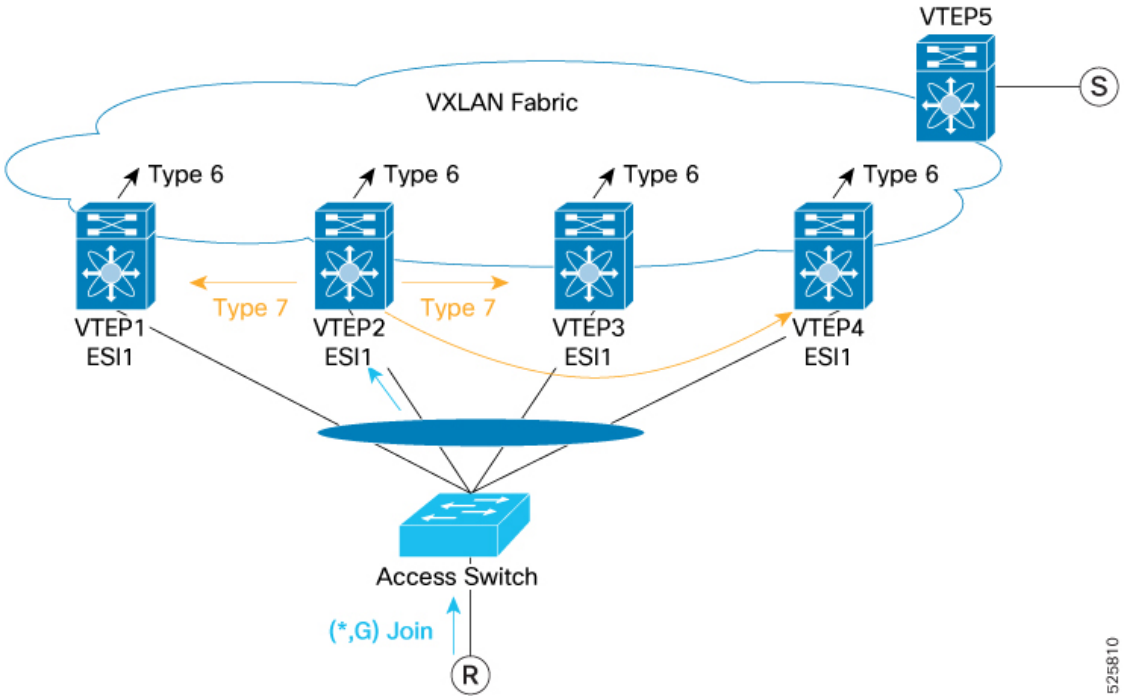
In this example topology, there is a PIM enabled router to the orphan port. In this case, vPC VTEP1 receives PIM hellos from the PIM enabled router, and views the orphan port as a mrouter port; VTEP1 will advertise a (*,*) route toward the EVPN network. Any multicast traffic received from source 'S' is forwarded toward the PIM enabled router. Furthermore, reports received from VTEP3 over SMET are sent as a proxy report toward the PIM enabled router. There are variants of the 'PIM router' topology, where the PIM enabled router is attached to a standalone node; however, the PIM enabled router attached to the vPC complex access switch is not supported.

Note PIM enabled router attached to vPC complex access switch is not supported.

Layer 2 EVPN with ESI

Beginning with Cisco NX-OS Release 10.6(1)F, the support for ESI port-channels has been added. For more information, see Supported L2 multicast topologies and IGMP control flows, on page 854. Correspondingly, support for EVPN Type-6 SMET route generation for a receiver R attached to an ESI port channel is also supported, as shown.

Figure 37: Layer 2 EVPN with ESI



In the ESI case, receivers can send a (*,G) and (S,G) IGMP or MLD membership report over the ESI port-channel. When the report is received by one of the VTEPs, this VTEP will originate a Type-6 EVPN SMET route across the EVPN network.Separately, the VTEP also sends an EVPN Type-7 Report-Sync route to ESI peers. When ESI peers receive the Type-7 routes, these VTEPs originate a Type-6 route to the EVPN network. Note that there is currently no support for the DF only to originate the Type-6 route, as is recommended in RFC 9251.

Guidelines and Limitations of Multicast Layer 2 EVPN Type-6 routes

Multicast Layer 2 EVPN route type 6 has the following configuration guidelines and limitations:

Generic

- If **advertise evpn multicast** is configured, it is recommended to disable NVE as a static router port using the following command:
  - ip igmp snooping disable-nve-static-router-port
  - ipv6 mld snooping disable-nve-static-router-port
- The **advertise evpn multicast** command is required for originating Type-6 SMET routes; IPv4 IGMP Snooped Routes are used for IPv4 and IPv6 MLD Snooped Routes are used for IPv6.

Note If **advertise evpn multicast** is not configured, IPv4 IGMP will flood IGMP control packets in the fabric. However, IPv6 MLD will not flood MLD control packets in the fabric.

- The NGMVPN feature is no longer necessary to originate Type-6 routes, though it can remain without impact. The configuration of "feature ngmvpn" was previously required as part of "L2TRM" for the Type-6 EVPN SMET route advertisement.
- Beginning with Cisco NX-OS Release 10.5(3)F, **show fabric multicast ipv4 l2-mroute** command is deprecated, there are equivalent commands in L2RIB to view the route state in L2RIB component.
- PIM packets will be flooded in the fabric.
- Beginning with Cisco NX-OS Release 10.6(1)F, EVPN route Type-6, 7 and 8 supports interoperability through RFC6625 encoding.

Route Type Advertising and Receiving Support

- Supports advertisement and receipt of IGMPv2, IGMPv3, and combinations of IGMPv2 and IGMPv3.
- Supports advertisement and receipt of MLDv1, MLDv2, and combinations of MLDv1, MLDv2.
- Supports advertisement and receipt of (*,G) IGMPv3 and (*,G) MLDv2 entries with the exclude flag
- Does not support IGMPv1 SMET route origination (as per RFC 9251).
- Does not support advertisement of S,G routes with the exclude flag; these are advertised as (*, G) exclude routes.
- Supports receipt of (S,G) routes with the exclude flag by treating as (*,G) routes.
- Supports advertisement of EVPN Type-6, 7 and 8 routes with multicast source length as 0 for (*, G) routes and both multicast source length and multicast group length set to 0 for (*, *) routes when the **advertise evpn multicast format rfc6625** command is enabled.

Feature Support

- Only supported in a Layer 2 fabric.
- Supports IPv4 multicast and ingress-replication (IR) underlay. However, the IR implementation does not support sending multicast traffic to only VTEP nodes with associated receivers.
- Does not support IPv6 multicast and IR underlay
- Does not support Multisite Border Gateway.
- Does not support MLD Snooping with VXLAN on EoR.
- To ensure membership reports are received and processed by the VTEP, querier functionality must be enabled on VTEPs based on topology.
- For successful downgrade from Cisco NX-OS Release 10.6(1)F to a prior release, ensure that **format rfc6625** option is removed from **advertise evpn multicast format rfc6625** configuration, otherwise downgrade is impacted.

Scale limitations

- In case of vPC or vPC fabric peering, triggers like uplink flap, reload where the peer-link goes down, there can be momentary increase in platform programming resources causing exhaustion. If exhaustion occurs due to such triggers, follow these steps:
  - Use the **clear ip igmp snooping groups * vlan all** command to clear the respective IGMP and MLD snooping tables.
  - Use the **no advertise evpn multicast** command to disable the SMET route generation on the affected node to recover from any mis-programming.
  - Use the **advertise evpn multicast** command to enable the SMET route generation.

Supported Release and Platform of Multicast Layer 2 EVPN Route Type-6

Release	Platform
10.5(3)F and later	Cisco Nexus 9300-FX/FX2/FX3/GX/GX2/H2R/H1 Series switches and Nexus 9500 Series switches with 9700-FX/GX/FX3 line cards

Configure Multicast Layer 2 EVPN Route Type 6

The generation of EVPN Type 6 BGP EVPN SMET routes is done using the **advertise evpn multicast** command. The procedure is similar to Configuring Layer 2 Tenant Routed Multicast, on page 425 section.

IGMP Snooping Querier must be configured per multicast-enabled VXLAN VLAN on all Layer-2 leaf switches.

Configure IPv4 IGMP Snooping

Follow these steps to configure IPv4 IGMP Snooping.

Before you begin

Before you begin, ensure that VXLAN EVPN is configured.

Procedure

Step 1 Run the **advertise evpn multicast** command in global configuration mode, to advertise Layer 2 multicast capability.

Example:

```
switch# configure terminal
switch(config)# advertise evpn multicast
```

Step 2 (Optional) Run the **advertise evpn multicast format rfc6625** command to enable RFC6625 encoding for EVPN route Type 6, 7 and 8, setting multicast source length to 0 for (*, G) routes and both multicast source length and multicast group length to 0 for (*, *) routes.

Example:

```
switch(config)# advertise evpn multicast format rfc6625
```

Note
Use the **format rfc6625** option only if interoperability is required.

Use **advertise evpn multicast** command to revert to the original encoding method with multicast source length as 32/128 and multicast source address as 0.0.0.0/0::0 for (*, G) routes, and both lengths as 32/128 with addresses as 0.0.0.0/0::0 for (*, *) routes.

Step 3 Run the **ip igmp snooping vxlan** command to configure IGMP snooping for VXLANs.

Example:

```
switch(config)# ip igmp snooping vxlan
```

Step 4 Run the **vlan configuration** *vlan-id* command to enter configuration mode for VLAN 101.

Example:

```
switch(config)# vlan configuration 101
switch(config-vlan-config)#
```

Step 5 Run the **ip igmp snooping querier** *querier-ip-address* command to configure IGMP snooping querier for each multicast-enabled VXLAN VLAN.

Example:

```
switch(config-vlan-config)# ip igmp snooping querier 2.2.2.2
```

Querier must be configured on the switch.

Configure IPv6 MLD Snooping

Follow these steps to configure IPv6 MLD Snooping.

Before you begin

Before you begin, ensure that

- VXLAN EVPN is configured.
- The **system mld snooping** command must be configured, followed by a reload.
- Use the **hardware access-list tcam region ing-sup** *size* command to enable the ACL TCAM **ing-sup** region as specified.

  ```
  switch(config)# hardware access-list tcam region ing-sup 768
  switch(config)# hardware access-list tcam region ing-racl 1024
  ```

- Use the **copy running-config startup-config** command and reload the switch or line cards to apply the configuration changes.

Procedure

Step 1 Run the **ipv6 mld snooping** command in global configuration mode, to enable MLD snooping on the switch.

Example:

```
switch(config)# ipv6 mld snooping
```

Step 2 Run the **ipv6 mld snooping vxlan** command to configure MLD snooping for VXLANs.

Example:

```
switch(config)# ipv6 mld snooping vxlan
```

Step 3 Run the **vlan configuration** *vlan-id* command to enter configuration mode for VLAN 101.

Example:

```
switch(config)# vlan configuration 501
switch(config-vlan-config)#
```

Step 4 Run the **ip mld snooping querier** *querier-ip-address* command to configure the MLD snooping querier for each multicast-enabled VXLAN VLAN.

Example:

```
switch(config-vlan-config)# ipv6 mld snooping querier fe80:0:0:1::1
```

Querier must be configured on the switch.

Verification of Multicast Layer 2 EVPN Route Types Configuration

Perform one of the following tasks to display Multicast Layer 2 EVPN route type configuration information.

Command	Purpose
show ip igmp snooping evpn	Displays IGMP snooping information for IPv4 EVPN route type.
show ipv6 mld snooping evpn	Displays MLD snooping information for IPv6 EVPN route type.
show ip igmp snooping evpn smet	Displays IGMP snooping information for IPv4 SMET route type.
show ipv6 mld snooping evpn smet	Displays MLD snooping information for IPv6 SMET route type.
show ip igmp snooping remote groups	Displays IGMP snooping information for remote installed SMET and Joinsync routes.
show ipv6 mld snooping remote groups	Displays MLD snooping information for IPv6 for remote installed SMET and Joinsync routes.

Command	Purpose
show ip igmp snooping mrouter	Displays IGMP mrouter ports added by SMET.
show ipv6 mld snooping mrouter	Displays MLD mrouter ports added by SMET.
show l2route smet {**topology** *topo-id* | **all**} [**detail**]	Displays L2RIB state of the SMET routes.
Show bgp l2vpn evpn route-type 6	Displays BGP route Type 6 state.

Examples of Multicast Layer 2 EVPN Route Type 6

The following examples show outputs for IGMP (IPv4) EVPN, similar commands are also available for MLD (IPv6).

- The following example shows the output of the **show ip igmp snooping evpn** command, where you can see whether the EVPN Route Type 6 SMET feature is configured.

```
switch# show ip igmp snooping evpn
IGMP Snoop EVPN Parameters:
Advertise EVPN : Configured
ESI-MH : Disabled
L2RIB : UP [Initialized 1] [handle 0x42ac020]
EVPN-RIB : UP
Last Trigger : ADD AdvertiseEvpnMcast [id:101]
          : Start:00:02:13 [133914.65048] End:00:02:13 [133914.388186]
```

- The following example shows the output of the **show ip igmp snooping evpn smet** command. This command shows locally originated and remotely received SMET routes and associated versions for this route on a VLAN.

```
switch # show ip igmp snooping evpn smet

VLAN       VNI        Group                Source               Local/Remote
Flags
11         1011       225.0.0.11           0.0.0.0              Local
v2
11         1011       225.0.0.11           0.0.0.0              Remote       E
 v2 v3
11         1011       225.0.0.11           11.1.1.108           Local        I
 v3
11         1011       225.0.0.11           11.1.1.108           Remote       I
 v3
11         1011       225.0.1.11           0.0.0.0              Local
v2
```

- The following example shows the output of the **show ip igmp snooping remote groups** command. This command is used to filter remotely learned groups.

  The **show ip igmp snooping groups** command provides the entire set of ports associated with an IGMP entry. Also, there is a **detail** option to show details of each route and associated ports.

```
switch# show ip igmp snooping remote groups
Type: S - Static, D - Dynamic, R - Router port, F - Fabricpath core port

Vlan  Group Address      Ver  Type  Port list
11    225.0.0.11         v2   S     nve1
        11.1.1.108            S     nve1
11    225.0.1.11         v2   S     nve1
```

- The following example shows the output of the **show ip igmp snooping mrouter** command. This command shows the router ports that are locally advertised or remotely learned. Remotely learned router port through SMET is marked as 'M' with NVE as the router port.



Note

- The "internally created" peer-link router port is not advertised in SMET.
- If "ip igmp snooping disable-nve-router-port" is not configured, then NVE is added as a static router port.

```
switch # show ip igmp snooping mrouter
Type: S - Static, D - Dynamic, V - vPC Peer Link
      I - Internal, F - Fabricpath core port
      C - Co-learned, U - User Configured
      P - learnt by Peer, M - learnt by SMET
Vlan  Router-port   Type      Uptime      Expires
16    nve1          M         05:33:42    never
17    nve1          M         05:49:10    never
18    nve1          M         05:49:10    never
19    nve1          M         05:49:10    never
20    nve1          M         05:49:10    never
```

- The following example shows the output of the **show l2route smet all detail** command. This command is used to provide visibility of the Type 6 SMET route data at the L2RIB level. Remotely learned routes show the associated route-originators for the route, and the flags from each router originator.

```
switch# show l2route smet all detail
V1: MLD version 1, V2: IGMP/MLD version 2, V3: IGMP version 3, I: Include Group Type,
E: Exclude Group Type
Topology    Source Group IP Addresses                Producer
----------- ------------------------------------------------------------------------

11          (0.0.0.0, 225.0.0.11)                        None
            Sent To: IGMPSN
            Total Counters (V1/V2/V3/I/E): 0/1/1/0/1

11          (0.0.0.0, 225.0.0.11)                        BGP
            Total Counters (V1/V2/V3/I/E): 0/2/1/0/1
            Originator router IP: 7.0.0.7, Flags: (V2)
            Originator router IP: 11.0.0.11, Flags: (V2,V3,E)

11          (0.0.0.0, 225.0.0.11)                        IGMP
            Sent To: BGP
            Total Counters (V1/V2/V3/I/E): 0/1/0/0/0
            Originator router IP: Local, Flags: (V2)

11          (11.1.1.108, 225.0.0.11)                     None
            Sent To: IGMPSN
            Total Counters (V1/V2/V3/I/E): 0/0/1/1/0

11          (11.1.1.108, 225.0.0.11)                     BGP
            Total Counters (V1/V2/V3/I/E): 0/0/4/4/0
            Originator router IP: 7.0.0.7, Flags: (V3,I)
            Originator router IP: 11.0.0.11, Flags: (V3,I)
            Originator router IP: 13.0.0.13, Flags: (V3,I)
            Originator router IP: 14.0.0.14, Flags: (V3,I)
```

Table 7: Route and Producer types

Route type	Producer	Description
Locally Advertised Routes	IGMP	These routes are locally advertised
Remotely Learned Routes	BGP	These routes show the associated route-originators for the route.
Routes Sent to IGMP	None	-

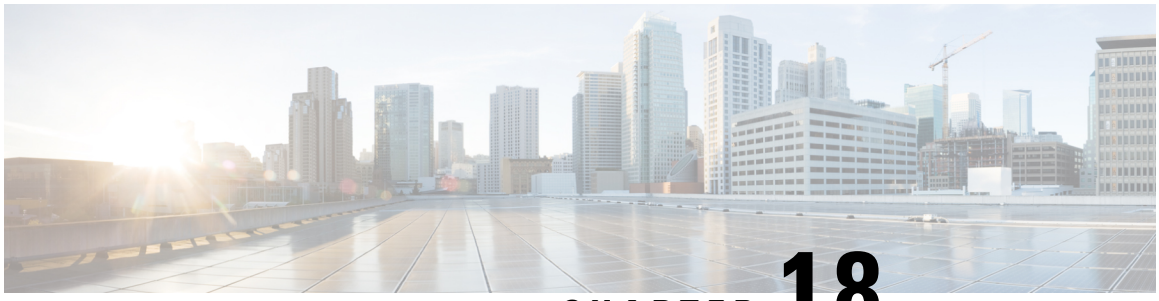
- The following example shows the output of the **show bgp l2vpn evpn route-type 6** command. This command is used to provide details of the received and advertised Type 6 EVPN routes at the BGP level.

```
switch # show bgp l2vpn evpn route-type 6
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 7.0.0.7:32778
BGP routing table entry for
[6]:[0]:[0]:[0.0.0.0]:[32]:[225.0.0.11]:[32]:[7.0.0.7]:[2]/144, version 204525
Paths: (1 available, best #1)
Flags: (0x000002) (high32 00000000) on xmit-list, is not in l2rib/evpn, is not in HW

  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop, is extd
             Imported to 1 destination(s)
             Imported paths list: L2-1011
  AS-Path: NONE, path sourced internal to AS
    7.7.7.7 (metric 45) from 7.0.0.7 (7.0.0.7)
      Origin IGP, MED not set, localpref 100, weight 0
      Extcommunity: RT:100:1011 ENCAP:8

  Path-id 1 not advertised to any peer
BGP routing table entry for
[6]:[0]:[0]:[0.0.0.0]:[32]:[225.0.1.11]:[32]:[7.0.0.7]:[2]/144, version 203912
Paths: (1 available, best #1)
Flags: (0x000002) (high32 00000000) on xmit-list, is not in l2rib/evpn, is not in HW

  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop, is extd
             Imported to 1 destination(s)
             Imported paths list: L2-1011
  AS-Path: NONE, path sourced internal to AS
    7.7.7.7 (metric 45) from 7.0.0.7 (7.0.0.7)
      Origin IGP, MED not set, localpref 100, weight 0
      Extcommunity: RT:100:1011 ENCAP:8
```



CHAPTER 18

VXLAN Path Validation and Verification

- VXLAN OAM or VXLAN NGOAM, on page 459
- Fault Isolation and Verification Tools, on page 462
- VXLAN EVPN Loop Detection and Mitigation , on page 472

VXLAN OAM or VXLAN NGOAM

VXLAN Operations, Administration, and Maintenance (OAM) or Next Generation Operations, Administration, and Maintenance (NGOAM) is a protocol that

- enhance the management of VXLAN overlay networks during installation, monitoring, and troubleshooting, and
- provide troubleshooting tools (VXLAN OAM tools) similar to ping, traceroute, or pathtrace for diagnosing problems in VXLAN networks.

These protocols use diagnostic channels to trace destinations and carry vital information.

VXLAN **OAM**, often referred to as **NGOAM.**

VXLAN OAM Tools

The VXLAN OAM tools are categorized as shown in the table. For more information on OAM tools, see Fault Isolation and Verification Tools, on page 462.

Category	Tools
Fault Verification	Loopback Message
Fault Isolation	Pathtrace Message

VXLAN OAM Payload

The OAM channel is used to identify the type of the VXLAN payload that is present in the OAM packets. The supported types of payloads are:

- **Conventional ICMP channel**: These channels are used to facilitate communication with traditional hosts or switches that do not support the new OAM packet formats.

- **NVO3 draft Tissa channel**: These channels are used to facilitate communication with supported hosts or switches, and they carry important diagnostic information. The term "channels" suggests a mechanism or pathway for delivering data.

  The VXLAN NVO3 draft Tissa OAM messages are identified using

  - **Reserved OAM Ether-Type**: This is a specific identifier used to recognize OAM packets. Ether-Type field indicates which protocol is encapsulated in the payload of an Ethernet frame.
  - **Reserved Source MAC Address**: Alternatively, OAM packets may be identified using a well-known reserved source MAC address. MAC addresses are unique identifiers assigned to network interfaces for communications on the physical network segment.

Guidelines and Limitations for VXLAN NGOAM

VXLANNGOAM has the following guidelines and limitations:

Beginning with Cisco NX-OS Release 10.2(3)F, you do not have to enable the VXLAN feature using the **feature nv overlay** command to use the NGOAM feature on intermediate nodes.

Supported Platform and Release for VXLAN NGOAM

Supported Release	Supported Platform
9.3(3) and later	Cisco Nexus 9300-FX/FX2/GX Series switches
9.3(5) and later	Cisco Nexus 9300-FX3 Series switches
10.2(3)F and later	Cisco Nexus 9300-GX2 Series switches
10.4(1)F and later	Cisco Nexus 9332D-H2R switches
10.4(2)F and later	Cisco Nexus 93400LD-H1 switches
10.4(3)F and later	Cisco Nexus 9364C-H1 switches Cisco Nexus 9800 Series switches
10.5(2)F and later	Cisco Nexus 9500 platform switches with N9K-X9736C-FX3 line card
10.5(3)F and later	Cisco Nexus 9364E-SG2-Q and 9364E-SG2-O switches
10.6(1)F and later	Cisco Nexus 9336C-SE1 switches

Configure VXLAN NGOAM

Follow these steps to configure the VXLAN NGOAM on Cisco Nexus switches.

Before you begin

Before you begin, ensure that the VXLAN configuration is complete.

Procedure

Step 1 Run the **feature ngoam** command in global configuration mode, to enable NGOAM feature.

Example:

```
switch# configure terminal
switch(config)# feature ngoam
```

Step 2 (Optional) Run the **show running-config ngoam** command to verify the ngoam configuration information.

Example:

```
switch# show run ngoam
feature ngoam
```

Configure NGOAM Profile

Follow these steps to configure the VXLAN profile on Cisco Nexus switches.

Before you begin

Before you begin, ensure that the **feature ngoam** configuration is complete.

Procedure

Step 1 Run the **ngoam profile** *profile-id* command in global configuration mode, to enable NGOAM profile.

Example:

```
switch# configure terminal
switch(config)# feature ngoam
switch(config)# ngoam profile 1
switch(config-ng-oam-profile)#
```

Range: 1 to 1023. Default: NA.

Step 2 Use the [**description** | **dot1q** | **flow** | **hop count** | **interface** | **oam-channel 2** | **payload** | **sport**] command to set the required options for configuring NGOAM profile.

Example:

Various options of NGOAM profiles configured under ngoam profile mode.

```
switch(config)# ngoam profile 1
switch(config-ngoam-profile)# oam-channel 2
switch(config-ngoam-profile)#flow forward payload pad 0x2
switch(config-ngoam-profile)#sport 12345, 54321
```

Various options of NGOAM profiles configured under ngoam profile flow sub mode.

```
switch(config)# ngoam profile 1
switch(config-ngoam-profile)# flow forward
switch(config-ng-oam-profile-flow)# oam-channel 2
switch(config-ng-oam-profile-flow)# payload
```

- **description**: Use this option to configure description of the profile.
- **dot1q**: Use this option to specify encapsulation using the dot1q tag.
- **flow**: Use this option to configure ngoam flow.
- **hop**: Use this option to configure ngoam hop count. Range: 1 to 255.
- **interface**: Use this option to configure NGOAM egress interface.
- **oam-channel**: Use this option to set the Oam-channel toNVO3 tissa.
- **payload**: Use this option to configure NGOAM payload.
- **sport**: Use this option to configure NGOAM UDP source port range. Range: 1 to 65535.

Step 3 (Optional) Run the **show running-config ngoam** command to verify the ngoam profile configuration information.

Example:

```
switch# show run ngoam
feature ngoam
ngoam profile 1
oam-channel 2
flow forward payload pad 0x2
```

NGOAM configuration placement in **show running-config** command output has been updated. Previously, NGOAM configurations appeared before interface configurations. Beginning with Cisco NX-OS Release, 10.6(1)F NGOAM configurations appears after interface-level configurations in the **show running-config** command output.

Example before the Change:

```
ngoam profile 1
  oam-channel 2

interface Ethernet1/1
  no switchport
  ip address 60.60.60.1/24
  no shutdown
```

Example after the Change:

```
interface Ethernet1/1
  no switchport
  ip address 60.60.60.1/24
  no shutdown

interface loopback10
  vrf member Org1:vrf1
  ipv6 address 2010::10/128
ngoam profile 1
  oam-channel 2
```

Fault Isolation and Verification Tools

In a VXLAN network, you may want to find the list of switches that a frame traverses to reach its destination. If the loopback test from a source switch to a destination switch fails, you must find out the offending switch in the path.

The Fault Isolation and Verification Tools (VXLAN OAM tools) are used to

- quickly identify problems in IP networks and
- provide reachability information to hosts and VTEPs within a VXLAN network.

Category	Tools
Fault Verification	Loopback (Ping) Message
Fault Isolation	Traceroute and Pathtrace Message

Ping Message

The loopback (ping) message is a utility tool used for fault verification. This utility tool detects various errors and the path failures.

Note: A **Ping**, often referred to as a **Loopback**.

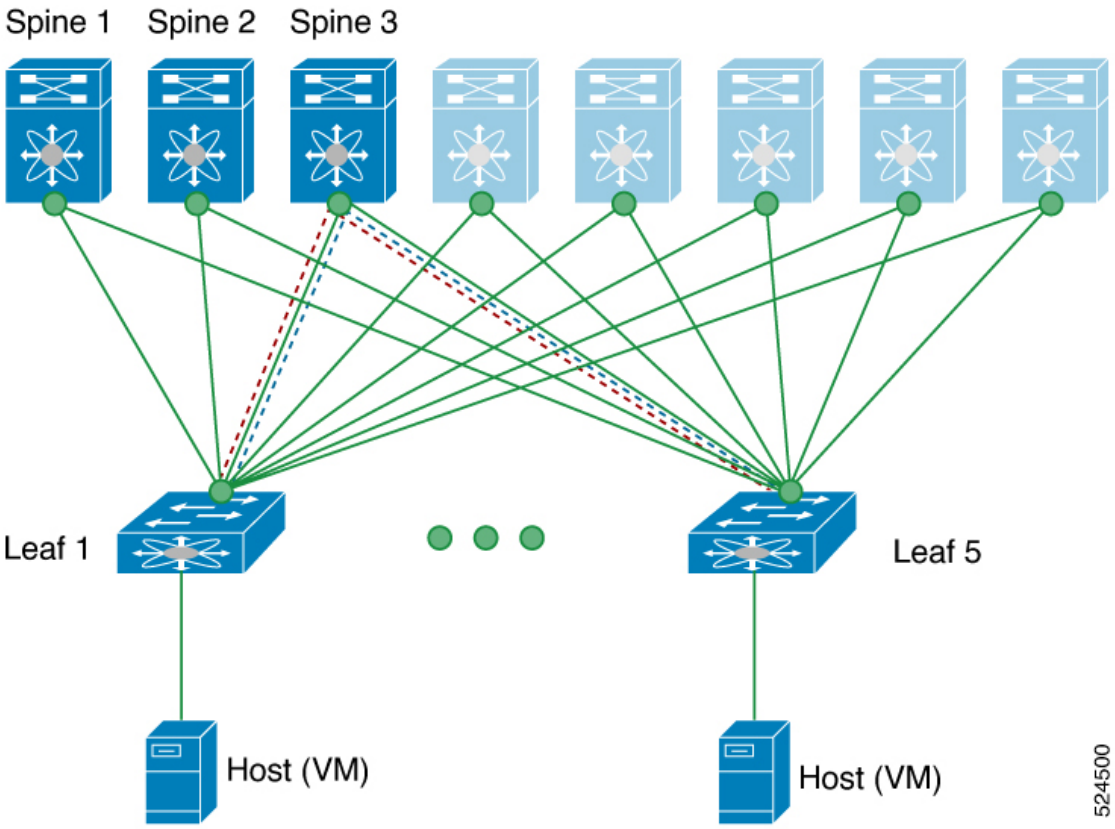
Topology of Ping Functionality

A Clos topology is shown where three core (spine) switches labeled Spine 1, Spine 2, and Spine 3 connect to five leaf switches.

The ping message validates these network reachability options using the **ping** command.

Here's how validation of network reachability takes place from Leaf 1 (VTEP 1) to Leaf 5 (VTEP 2) on different OAM channels

ICMP Channel	NVO3 Draft Tissa Channel
Loopback message is initiated from Leaf 1 to Leaf 5	Loopback message is initiated from Leaf 1 to Leaf 5
Loopback message is forwarded as VXLAN encapsulated data packet based on the outer header from Spine 3	Loopback message is forwarded as VXLAN encapsulated data packet based on the outer header from Spine 3
Loopback is processed and responded in-band from Leaf 5.	• Message is processed, payload is decapsulated and sent to host (VM) • Host generates the reply for the received payload and sends it to leaf 5. • Reply is processed for the response received from remote VTEP and sent in-band.
Loopback response is processed to Leaf 1	Loopback response is processed to Leaf 1.



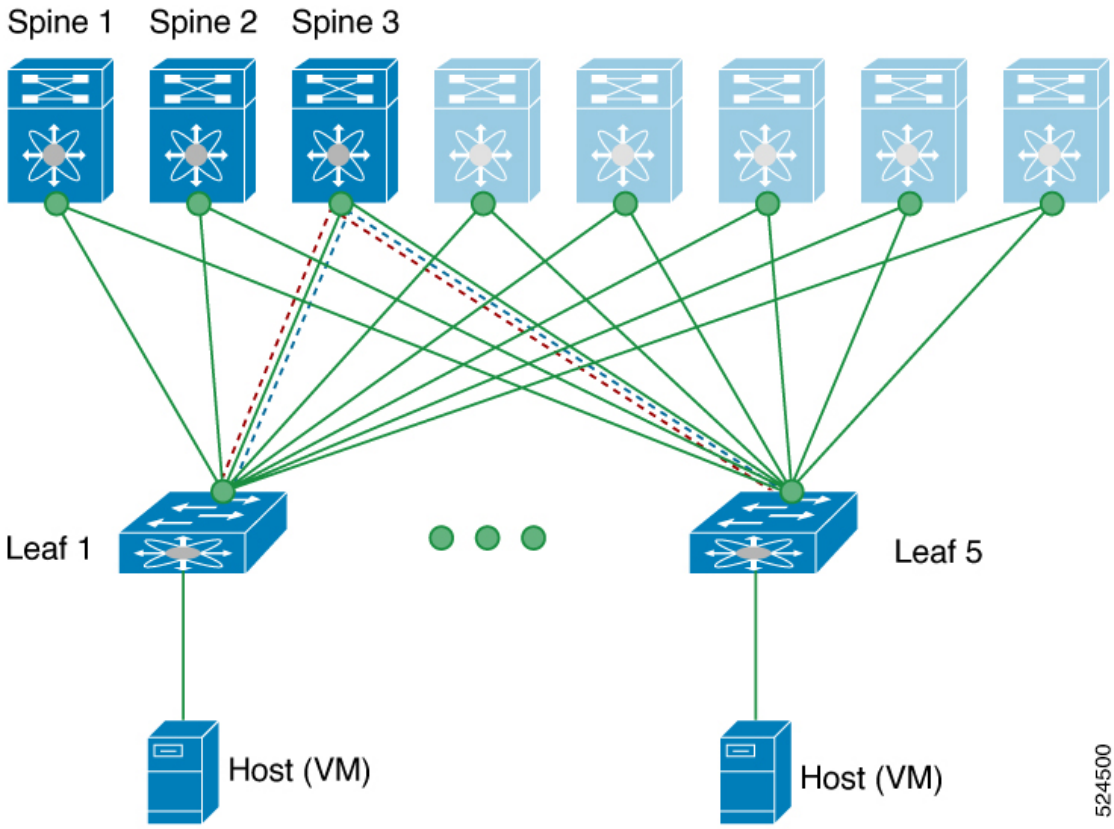
Traceroute Message

The Traceroute message is a utility tool used for the fault Isolation. This utility tool traces various errors and the path failures.

Traceroute behavior differs between single-site and multisite scenarios, with varying outputs. It's important to understand the process of traceroute as mentioned in Traceroute Functionality in Single site, on page 464 and Traceroute Functionality in Multi-Site, on page 465 sections.

Traceroute Functionality in Single site

A Clos topology is shown where three core (spine) switches labeled Spine 1, Spine 2, and Spine 3 connect to five leaf switches.



The Traceroute message validates the path traversed by the packet in the VXLAN overlay using the **Traceroute** command.

Here's how validation of network reachability takes place for Traceroute message from Leaf 1 (VTEP 1) to Leaf 5 (VTEP 2) on ICMP packets (channel-1), encapsulated in VXLAN.

- A traceroute message is initiated from Leaf 1 to Leaf 5 through Spine 3.
- Traceroute message is forwarded as a VXLAN encapsulated data packet based on the outer header from Spine 3.
- Traceroute is processed and responded in-band on Leaf 5.
- Traceroute is processed on Leaf 1.

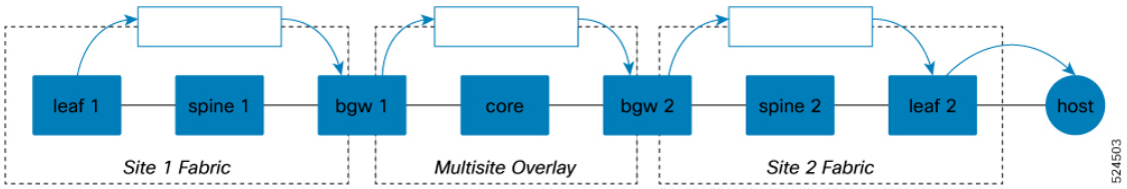
Traceroute Functionality in Multi-Site

Traceroute-IP

As shown in the diagram, the multiple probes are sent with Time To Live (TTL) expiry or ACL hits, where:

- Arrows pointing to a node indicate which hops the trace appears to hit.
- Arrows pointing into a pipe represent the packet being encapsulated in VXLAN.

  When encapsulated, you won't see responses from nodes until it drops out of the pipe, as encapsulation adds a higher TTL to the outer packet, meaning that the TTL expiry upon which a traceroute depends doesn't occur inside the pipe.



IP traceroute Process

The process described for the IP traceroute involves the following steps:

- A regular UDP packet is initiated and encapsulated in VXLAN at the leaf switch.
- The packet travels through the network and is decapsulated at the Border Gateway (BGW) of Site 1.
- The BGW of Site 1 receives the packet and sends a response.
- The packet is then re-encapsulated at BGW 1 and continues its journey through the network.
- The packet exits the tunnel at BGW 2, where another response is received.
- The packet is encapsulated when more and exits at a leaf in Site 2, prompting another response.
- Finally, the packet reaches the leaf, and the final response is seen.

This sequence ensures that the packet is properly encapsulated and decapsulated as it traverses through different sites and network components, allowing for accurate tracing of the packet's path.

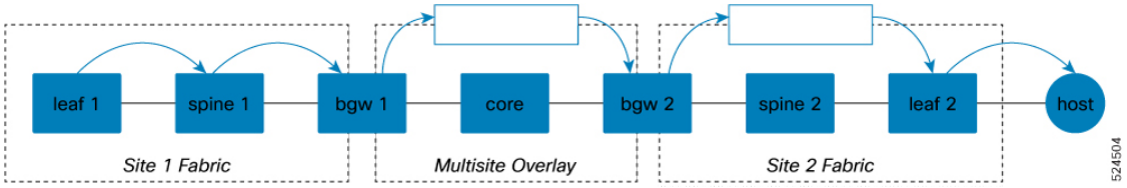
Traceroute - NVE

As shown in the diagram, in an NVE traceroute, NGOAM identifies that the traceroute is generated from a VTEP and involves the following steps:

- Initially traces the underlay network up to the remote VTEP.
- Switches to an overlay traceroute, functioning similarly to an IP traceroute.
- Uses UDP requests for the underlay and ICMP requests for the overlay after the remote VTEP.
- Probes are encapsulated in VXLAN upon reaching the remote VTEP.

Note Probes in the local fabric are not encapsulated in VXLAN, allowing visibility of nodes.

- After the local BGW, the output resembles a normal IP traceroute as probes enter pipes in multi-site fabrics.

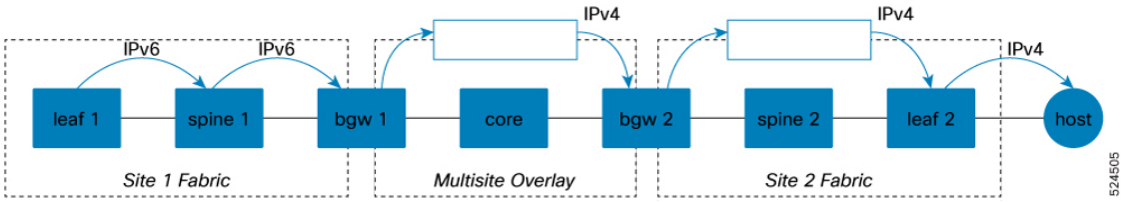


Traceroute (NVE - IPv4 over IPv6)

The hybrid tracing results in a mix of IPv6 and IPv4 responses due to the transition between underlay and overlay networks

As shown in the diagram, in an NVE- IPv4 over IPv6 traceroute, NGOAM generates probes as IPv6 within the local underlay fabric and involves the following steps:

- Receive IPv6 responses from the local spine and Border Gateway (BGW).
- When the trace reaches the BGW, NGOAM switches to overlay tracing.
- As the overlay is IPv4, receive IPv4 responses from the visible nodes beyond the BGW, as the packet effectively becomes IPv4.



Pathtrace Message

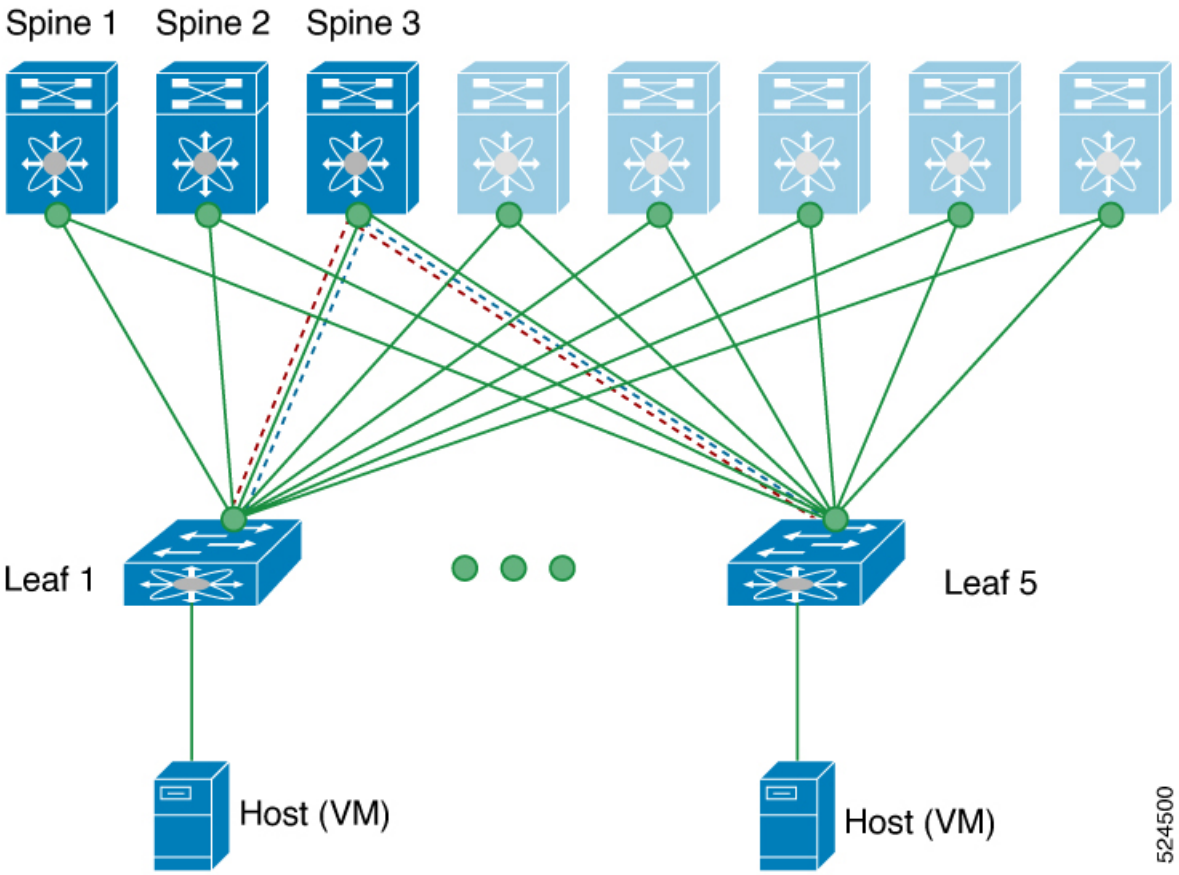
The Pathtrace message is a utility tool used for the fault Isolation. This utility tool traces various errors and path failures.

Pathtrace Functionality in Single site

A Clos topology is shown where three core (spine) switches labeled Spine 1, Spine 2, and Spine 3 connect to five leaf switches.

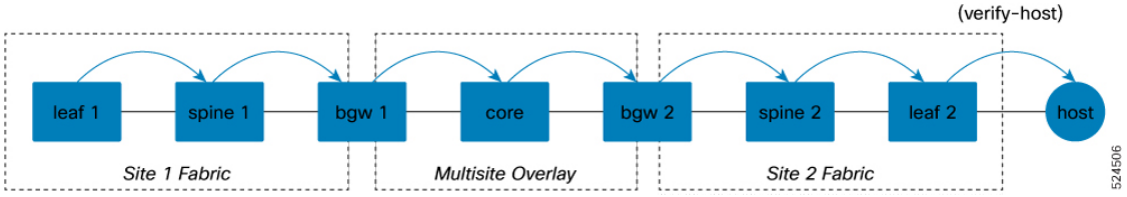
Pathtrace uses the NVO3 draft Tissa channel, encapsulated in the VXLAN encapsulation to reach the host and traces the path that is traversed by the packet in the VXLAN overlay using the **Pathtrace** command.

The NVO3 draft Tissa channel provides additional information about the path, such as ingress interface and egress interface. These packets stop at VTEP, and they do not reach the host. Therefore, only the VTEP responds.



Pathtrace Functionality in Multi-Site

As shown in the figure, pathtrace will generate a response from each node in the fabric. It uses a different channel (NVO3), which allows VXLAN-capable nodes in the fabric to process the packet specially due to ACL hits rather than TTL expiry. This makes it more capable of capturing the packet for processing, as long as the node supports NGOAM. Also, pathtrace receives special handling by NGOAM on the BGW, which adjusts the probe to allow it to continue into the next fabric.



Comparison of Traceroute and Pathtrace Message

Feature	Traceroute Message	Pathtrace Message
Channel Used	ICMP channel	NVO3 draft Tissa channel

Feature	Traceroute Message	Pathtrace Message
Purpose	Fault isolation by discovering the path packets take to reach their destination	Provides additional diagnostic information such as interface load and hop statistics
Behavior on Unsupported Devices	Continues to provide hop information	Behaves as a simple traceroute and provides only hop information

Guidelines for Fault Isolation and Verification Tools

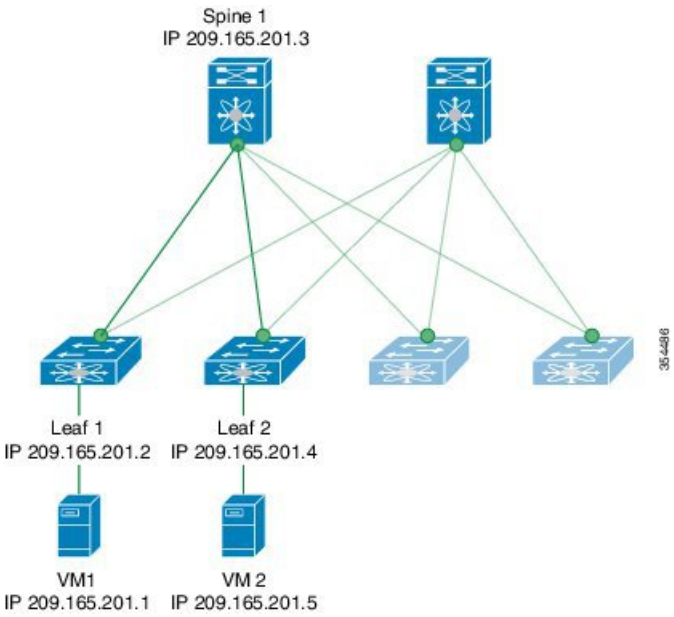
Beginning with NX-OS release 9.3(3), the **Received** field of the **show ngoam pathtrace statistics summary** command indicates all pathtrace requests received by the node on which the command is executed regardless of whether the request was destined to that node.

Examples for Fault Isolation and Verification Tools

Ping Message Examples

VXLAN OAM provides host visibility at the switch level, allowing a leaf to ping the host using the **ping nve** command.

Figure 38: VXLAN Network



The following examples display how to ping from Leaf 1 to VM2 via Spine 1 with channel 1 (unique loopback) and with channel 2 (NVO3 Draft Tissa).

- `switch# ping nve ip 209.165.201.5 vrf vni-31000 source 1.1.1.1 verbose`

  ```
  Codes: '!' - success, 'Q' - request not sent, '.' - timeout, 'D' - Destination
  Unreachable, 'X' - unknown return code, 'm' - malformed request(parameter problem),
  'c' - Corrupted Data/Test, '#' - Duplicate response
  ```

```
Sender handle: 34
! sport 40673 size 39,Reply from 209.165.201.5,time = 3 ms
! sport 40673 size 39,Reply from 209.165.201.5,time = 1 ms
! sport 40673 size 39,Reply from 209.165.201.5,time = 1 ms
! sport 40673 size 39,Reply from 209.165.201.5,time = 1 ms
! sport 40673 size 39,Reply from 209.165.201.5,time = 1 ms
Success rate is 100 percent (5/5), round-trip min/avg/max = 1/4/18 ms Total time elapsed
 49 ms
```



Note Thesource ip-address 1.1.1.1 used in this example is a loopback interface configured on Leaf 1 in the same VRF as the destination IP address. For example, the VRF in this example is vni-31000.

- switch# **ping nve ip unknown vrf vni-31000 payload ip 209.165.201.5 209.165.201.4 payload-end verify-host**

```
<snip>
Sender handle: 34
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 1/4/18 ms Total time elapsed
 49 ms
```

The following example displays the output of the MAC ping from Leaf 2 to Leaf 1 using NVO3 draft Tissa channel.

```
switch# ping nve mac 0050.569a.7418 2901 ethernet 1/51 profile 4 verbose

Codes: '!' - success, 'Q' - request not sent, '.' - timeout, 'D' - Destination Unreachable,
 'X' - unknown return code, 'm' - malformed request(parameter problem),
'c' - Corrupted Data/Test, '#' - Duplicate response

Sender handle: 408
!!!!Success rate is 100 percent (5/5), round-trip min/avg/max = 4/4/5 ms Total time elapsed
 104 ms

switch# show run ngoam
feature ngoam ngoam profile 4
oam-channel 2 ngoam install acl
```

Traceroute Message Examples

The following example displays how to traceroute from Leaf 1 to VM 2 through Spine 1.

```
switch# traceroute nve ip 209.165.201.5 vrf vni-31000 source 1.1.1.1 verbose


Codes: '!' - success, 'Q' - request not sent, '.' - timeout, 'D' - Destination Unreachable,
 'X' - unknown return code, 'm' - malformed request(parameter problem),
'c' - Corrupted Data/Test, '#' - Duplicate response


Traceroute request to peer ip 209.165.201.4 source ip 209.165.201.2 Sender handle: 36
1 !Reply from 209.165.201.3,time = 1 ms
2 !Reply from 209.165.201.4,time = 2 ms
3 !Reply from 209.165.201.5,time = 1 ms
The following example displays the output of the pathtrace from Leaf 2 to Leaf 1.
switch# pathtrace nve ip 209.165.201.4 vni 31000 verbose
Path trace Request to peer ip 209.165.201.4 source ip 209.165.201.2 Sender handle: 42
Hop Code ReplyIP IngressI/f EgressI/f State
```

```
=====================================================================
1 !Reply from 209.165.201.3, Eth5/5/1 Eth5/5/2 UP/UP
2 !Reply from 209.165.201.4, Eth1/3 Unknown UP/DOWN
```

Pathtrace Message Examples

The following example displays how to pathtrace based on a payload from Leaf 2 to Leaf 1.

```
switch# pathtrace nve ip unknown vrf vni-31000 payload mac-addr 0050.569a.d927 0050.569a.a4fa
 ip 209.165.201.5 209.165.201.1 port 15334 12769 proto 17 payload-end
Codes: '!' - success, 'Q' - request not sent, '.' - timeout, 'D' - Destination Unreachable,
 'X' - unknown return code, 'm' - malformed request(parameter problem),
'c' - Corrupted Data/Test, '#' - Duplicate response

Path trace Request to peer ip 209.165.201.4 source ip 209.165.201.2 Sender handle: 46
Hop Code Reply IngressI/f EgressI/f State
=====================================================================
1 !Reply from 209.165.201.3, Eth5/5/1 Eth5/5/2 UP/UP
2 !Reply from 209.165.201.4, Eth1/3 Unknown UP/DOWN
```

Note When the total hop count to the final destination is more than 5, the path trace default TTL value is 5. Use the **max-ttl** option to complete the VXLAN OAM path trace completely.

For example: **pathtracenve ip unknown vrf vrf-vni13001 payload ip 200.1.1.71 200.1.1.23 payload-end verbose max-ttl 10**

The following example displays how to pathtrace NVE MAC

```
switch# pathtrace nve mac 0050.569a.d927 11 payload mac-addr 0050.569a.d927 0050.569a.a4fa
 payload-end vni 31000 verbose

Codes: '!' - success, 'Q' - request not sent, '.' - timeout, 'D' - Destination Unreachable,
 'X' - unknown return code, 'm' - malformed request(parameter problem),
'c' - Corrupted Data/Test, '#' - Duplicate response, 'v' - Other - Use verbose to see the
result

Path trace Request to peer ip 209.165.201.4 source ip 209.165.201.2 Sender handle: 46
Hop Code Reply IngressI/f EgressI/f State
=====================================================================
1 !Reply from 209.165.201.3, Eth5/5/1 Eth5/5/2 UP/UP
2 !Reply from 209.165.201.4, Eth1/3 Unknown UP/DOWN
```

Note When the total hop count to the final destination is more than 5, the path trace default TTL value is 5. Use **max-ttl** option to finish VXLAN OAM path trace completely.

For example: **pathtrace nve ip unknown vrf vrf-vni13001 payload ip 200.1.1.71200.1.1.23 payload-end verbose max-ttl 10**

VXLAN EVPN Loop Detection and Mitigation

Causes and Impacts of Loop

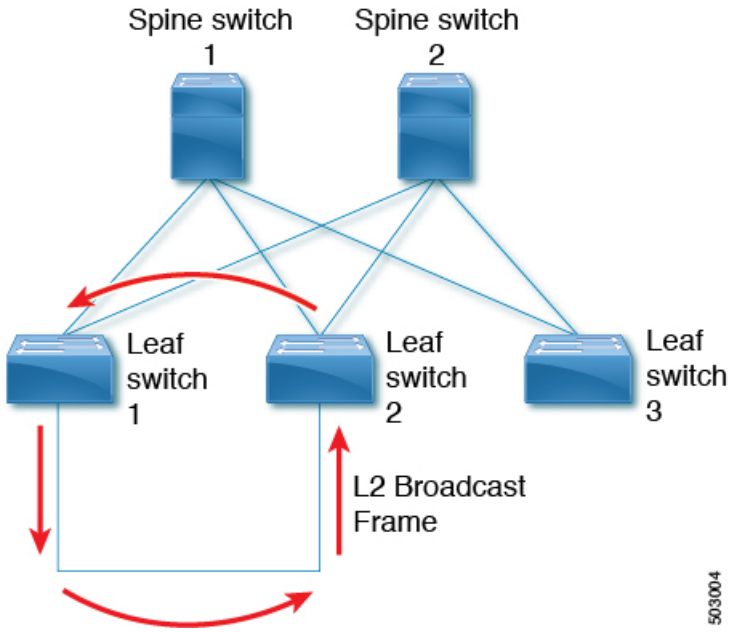
Loops usually occur in a VXLAN EVPN fabric due to incorrect cabling on the south side (access side) of the fabric. When broadcast packets are injected into a network with a loop, the frame remains bridged in the loop. As more broadcast frames enter the loop, they accumulate and can cause a serious disruption of services.

VXLAN EVPN Loop Detection and Mitigation

Cisco NX-OS Release 9.3(5) introduces VXLAN EVPN loop detection and mitigation. This feature detects Layer 2 loops in a single VXLAN EVPN fabric or a Multi-Site environment. It operates at the port/VLAN level and disables the VLAN(s) on each port where a loop is detected. Administrators are also notified (via syslog) about the condition. In this way, the feature ensures that the network remains up and available.

The following figure shows an EVPN fabric in which two leaf devices (Leaf1 and Leaf2) are directly connected on the south side due to incorrect cabling. In this topology, Leaf3 forwards an L2 broadcast frame to Leaf1. Then the broadcast frame is repeatedly forwarded between Leaf1 and Leaf2 through the south side and the fabric. The forwarding continues until the incorrect cabling is fixed.

Figure 39: Two Leaf Nodes Directly Connected



This feature operates in three phases:

1. Loop Detection: Sends a loop detection probe under the following circumstances: when requested by a client, as part of a periodic probe task, and as soon as any port comes up.
2. Loop Mitigation: Blocks the VLANs on a port once a loop has been discovered and displays a syslog message similar to the following:

```
2020 Jan 14 09:58:44 Leaf1 %NGOAM-4-SLD_LOOP_DETECTED: Loop detected - Blocking vlan
1001 :: Eth1/3
```

or

```
2024 Sep  9 15:28:01 Node-11 %ETHPORT-3-IF_ERROR_VLANS_SUSPENDED: VLANs 2704 on Interface
 Ethernet1/49/1 are being suspended. (Reason: SUCCESS)
```

Because loops can lead to incorrect local MAC address learning, this phase also flushes the local and remote MAC addresses. Doing so removes any MAC addresses that are incorrectly learned.

In the previous figure, MAC addresses can be incorrectly learned because packets from hosts sitting behind the remote leaf (Leaf3) can reach both Leaf1 and Leaf2 from the access side. As a result, the hosts incorrectly appear local to Leaf1 and Leaf2, which causes the leafs to learn their MAC addresses.

3. Loop Recovery: Once a loop is detected on a particular port or VLAN and the recovery interval has passed, recovery probes are sent to determine if the loop still exists. When NGOAM recovers from the loop, a syslog message similar to the following appears:

```
2020 Jan 14 09:59:38 Leaf1 %NGOAM-4-SLD_LOOP_GONE: Loop cleared - Enabling vlan 1001
:: Eth1/3
```

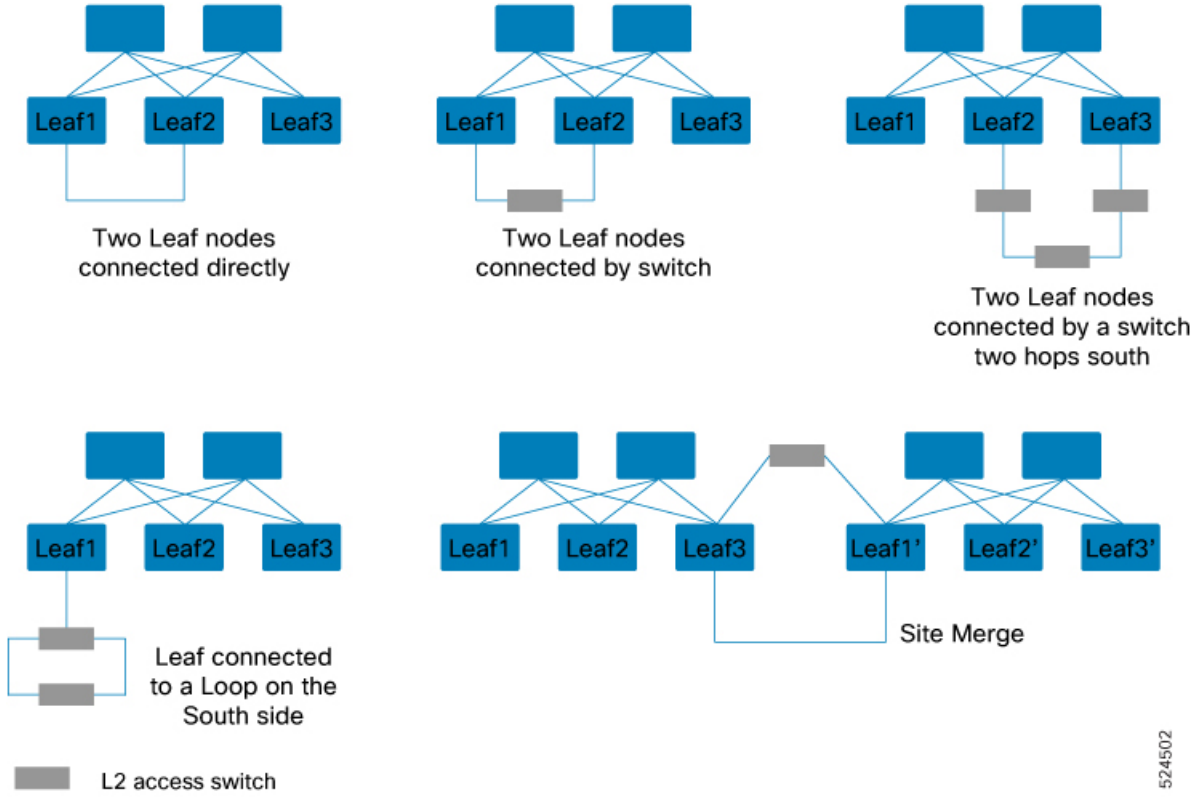
or

```
2024 Sep  9 15:24:23 Node-11 %ETHPORT-3-IF_ERROR_VLANS_REMOVED: VLANs 384 on Interface
Ethernet1/49/1 are removed from suspended state.
```



Note The default logging level for NGOAM does not generate a syslog message. Modifying the logging level of NGOAM to 5 with "logging level ngoam 5" will result in a syslog message being generated when a loop is detected.

Various Loop Scenarios



Southbound Loop Detection on Layer-3 Interface

Beginning with NX-OS release 10.4(3)F, Cisco Nexus switches support Southbound Loop Detection (SLD) on a Layer-3 (L3) Ethernet and L3 port-channel interfaces in a single VXLAN EVPN fabric or a Multi-Site environment. Before this release, the SLD feature was supported only on Layer-2 interfaces.

This feature detects loops in southbound side (L2 access switches) that are connected to a single leaf switch through an L3 interface or port channel.

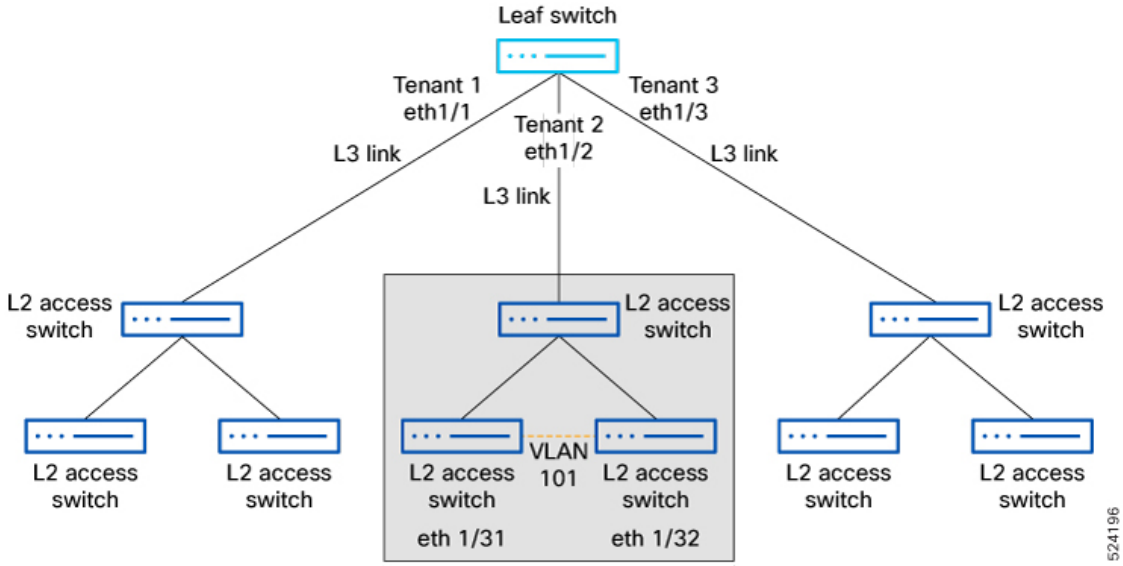
When the SLD feature is enabled on the L3 interface, it sends periodic SLD probes to detect loops in a downstream tenant's Layer-2 domain. It continues to monitor for loops and blocks the L3 interface on detection until the user takes action to correct the condition in the downstream L2 domain.

Functionalities of SLD on Layer-3 Interface

- Isolates a single L3 attached tenant to prevent the impact of a storm from propagating beyond a single L3 boundary due to control-plane policing congestion.
- Detects downstream L2 loops and blocks attached L3 interface or L3 port-channel if a loop is detected by receipt of an originated NGOAM probe.
- Unblocks the L3 port if the originated NGOAM probes are no longer detected.

Topology Overview of SLD on Layer-3 Interface

The following figure shows an EVPN fabric with a leaf switch configured with three VRFs (Tenant 1, Tenant 2, and Tenant 3). These VRFs are connected to L2 access switches on the south side using different L3 ports and their respective L3 interfaces.



This feature operates in three phases:

- **Loop Detection**: The SLD L3 feature sends periodic probes to detect loops in the downstream tenant's Layer-2 domain (L2 access switches).

  SLD sends a loop detection probe under the following circumstances: when requested by a client, as part of a periodic probe task, and when any port comes up.

  For example: Tenant 2 accidentally creates a bridging loop due to a cabling error while disabling STP on local VLAN 101. This triggers an ARP storm toward Eth1/2, consuming the entire CoPP Class Normal policer, which causes CoPP policer saturation in Tenant 1 and Tenant 3.

  ```
  2024 Jun 27 02:34:39 tenant2 %L2FM-2-L2FM_CONTINUOUS_MAC_MOVE: Mac
  Address (f80f.6f96.a127) in Vlan 101 is moving continuously. Mac moved
  between Eth1/32 to Eth1/31. Please enable 'logging level l2fm 4' for
  verbose output.
  ```

- **Loop Mitigation**: Blocks the L3 port when a loop has been discovered and displays a syslog message like the following indicating the loop detection and port status changes.

  ```
  2024 Jun 27 02:37:50 leaf %ETHPORT-5-IF_DOWN_NONE: Interface Ethernet1/2 is down (None)
  2024 Jun 27 02:37:50 leaf %ETHPORT-5-IF_DOWN_ERROR_DISABLED: Interface Ethernet1/2 is
  down (Error disabled. Reason:error)
  2024 Jun 27 02:38:52 leaf %ETHPORT-5-IF_ERRDIS_RECOVERY: Interface Ethernet1/2 is being
   recovered from error disabled state (Last Reason:error)
  2024 Jun 27 02:38:54 leaf %ETHPORT-5-IF_DOWN_ERROR_DISABLED: Interface Ethernet1/2 is
  down (Error disabled. Reason:error)
  !
  leaf# show ngoam loop-detection status l3
  Port       Status       NumLoops      DetectionTime             ClearedTime
  ================================================================================
  ```

```
Eth1/2     BLOCKED      2             Tue Jun 27 02:38:54 2024  Tue Jun 27 02:38:52
2024
```

After each probe error recovery interval, the blocked L3 port is brought up to send probes and recheck for the loop. Now, the Eth1/2 L3 interface is moved from the **Blocked** state to the **Forwarding** state. The probe checks for the loop, and if the loop still exists, it moves the eth1/2 L3 interface back to the **Blocked** state. This process continues until the user corrects the bridging loop within the L2 domain.

The following sample output displays the state (blocking and unblocking) based on the probe generated:

```
2024 Jun 27 20:26:56 leaf %NGOAM-4-SLD_L3_LOOP_DETECTED: Loop detected - Blocking port
 Eth1/2
2024 Jun 27 20:26:56 leaf %ETHPORT-5-IF_DOWN_NONE: Interface Ethernet1/2 is down (None)
2024 Jun 27 20:26:56 leaf %ETHPORT-5-IF_DOWN_ERROR_DISABLED: Interface Ethernet1/2 is
down (Error disabled. Reason:error)
2024 Jun 27 20:27:58 leaf %ETHPORT-5-IF_ERRDIS_RECOVERY: Interface Ethernet1/2 is being
 recovered from error disabled state (Last Reason:error)
2024 Jun 27 20:27:58 leaf %NGOAM-4-SLD_L3_LOOP_GONE: Loop cleared - Enabling port Eth1/2
2024 Jun 27 20:28:00 leaf %NGOAM-4-SLD_L3_LOOP_DETECTED: Loop detected - Blocking port
 Eth1/2
2024 Jun 27 20:28:01 leaf %ETHPORT-5-IF_DOWN_ERROR_DISABLED: Interface Ethernet1/2 is
down (Error disabled. Reason:error)
```

- **Loop Recovery**: When the cabling error is fixed, the loops on the southbound side will be removed. After the recovery interval has passed, recovery probes will be sent from the L3 interface on the leaf switch to determine whether a loop exists. If the loop is resolved, the port will remain in the forwarding state, and the following syslog message will be generated.

```
2024 Jun 27 22:39:26 tenant2 %ETHPORT-5-IF_DOWN_ADMIN_DOWN: Interface Ethernet1/32 is
down (Administratively down)
2024 Jun 27 22:39:56 tenant2 %ETHPORT-5-SPEED: Interface Ethernet1/2, operational speed
 changed to 10 Gbps
2024 Jun 27 22:39:56 tenant2 %ETHPORT-5-IF_DUPLEX: Interface Ethernet1/2, operational
duplex mode changed to Full
2024 Jun 27 22:39:56 tenant2 %ETHPORT-5-IF_RX_FLOW_CONTROL: Interface Ethernet1/2,
operational Receive Flow Control state changed to off
2024 Jun 27 22:39:56 tenant2 %ETHPORT-5-IF_TX_FLOW_CONTROL: Interface Ethernet1/2,
operational Transmit Flow Control state changed to off
2024 Jun 27 22:39:56 tenant2 %ETHPORT-5-IF_UP: Interface Ethernet1/17 is up
2024 Jun 27 22:41:03 tenant2 %VSHD-5-VSHD_SYSLOG_CONFIG_I: Configured from vty by admin
 on 10.82.195.201@pts/2
```



Note The default logging level for NGOAM does not generate a syslog message. Modifying the logging level of NGOAM to 5 with "logging level ngoam 5" will result in a syslog message being generated when a loop is detected.

L2 and L3 SLD Feature Functionality Comparison

Features	SLD on L2 Interface	SLD on L3 Interface
Operation level	Port and VLAN level	Ethernet and L3 port-channel
Environment	Single-Site and Multi-Site	Single-Site and Multi-Site
Loop detection	Detects the loop of a particular port or VLAN	Detects downstream L2 loops and blocks the L3 interfaces or L3 port channels

Features	SLD on L2 Interface	SLD on L3 Interface
Loop mitigation	Blocks the VLANs on a port once a loop has been discovered and displays a syslog message	Isolates a single L3 attached tenant to prevent the impact of a storm from propagating beyond a single L3 boundary by consuming shared CoPP policer resources
Loop blocking	Breaks the southbound loops	Isolates detected loops from impacting the control plane by shedding storm-related traffic
Loop recovery	Sends recovery probes, re-enables VLANs, and logs syslog messages once the loop is cleared	Sends recovery probes, re-enables the port or ethernet interfaces if the NGOAM process no longer sees NGOAM probes, and logs syslog messages once the loop is cleared

Guidelines and Limitations for VXLAN EVPN Loop Detection and Mitigation

VXLAN EVPN loop detection and mitigation has the following guidelines and limitations:

- VXLAN EVPN loop detection and mitigation is supported in both STP and STP-less environments.
- To be able to detect loops across sites for VXLAN EVPN Multi-Site deployments, the **ngoam loop-detection** command needs to be configured on all border gateways in the site where the feature is being deployed.
- VXLAN EVPN loop detection and mitigation isn't supported with the following features:
  - Private VLANs
  - VLAN translation
  - ESI-based multihoming
  - VXLAN Cross Connect
  - Q-in-VNI
  - EVPN segment routing (Layer 2)



Note Ports or VLANs configured with these features must be excluded from VXLAN EVPN loop detection and mitigation. You can use the **disable** {**vlan** *vlan-range*} [**port** *port-range*] command to exclude them.

Supported Platform and Release for VXLAN EVPN Loop Detection and Mitigation

Supported Release	Supported Platform
9.3(5) and later	Cisco Nexus 9300-FX/FX2 Series switches Cisco Nexus 9500 platform switches with X97160YC-EX and 9700-FX line cards
10.1(1) and later	Cisco Nexus 9300-FX3/GX Series switches
10.2(3)F and later	Cisco Nexus 9300-GX2 Series switches
10.4(1)F and later	Cisco Nexus 9332D-H2R Series switches
10.4(2)F and later	Cisco Nexus 93400LD-H1 Series switches
10.4(3)F and later	Cisco Nexus 9364C-H1 Series switches
10.5(2)F and later	Cisco Nexus 9500 platform switches with 9700-FX3 line cards

Guidelines and Limitations for SLD on L3 Interface

- SLD is supported only on L3 ethernet and L3 port-channel interfaces. It is not supported on L3 sub-interfaces.

Supported Platform and Release for SLD on L3 Interface

Release	Platform
10.4(3)F and later	Cisco Nexus 9300-FX/FX2/GX/GX2/H2R/H1 Series switches Cisco Nexus 9500 platform switches with 9700-FX/GX line cards
10.5(2)F and later	Cisco Nexus 9500 platform switches with 9700-FX3 line cards

Prerequisites of NGOAM Southbound Loop Detection

Before you begin, ensure to

- Enable the NGOAM feature.
- Use the **hardware access-list tcam region ing-sup 768** command to create space for the TCAM ing-sup region.

Note

- Ensure that additional TCAM entries are freed up before increasing the allocation for the ing-sup region.
- Configuring the TCAM region requires the node to be rebooted.

Configure NGOAM Southbound Loop Detection on Layer-2 Interfaces

Follow these steps to configure NGOAM Southbound loop detection and mitigation.

Procedure

Step 1 Run the [**no**] **ngoam loop-detection** command in global configuration mode, to enable NGOAM Southbound loop detection and mitigation for all VLANs or ports.

Example:

```
switch# configure terminal
switch(config)# ngoam loop-detection
switch(config-ng-oam-loop-detection)#
```

This feature is disabled by default.

The **no** form of this command disable the NGOAM Southbound loop detection and mitigation.

Step 2 (Optional) Run the [**no**] **disable** {**vlan** *vlan-range*} [**port** *port-range*] command to disable NGOAM Southbound loop detection and mitigation for specific VLANs or ports and brings up any loop-detected ports.

Example:

Disables on specific VLAN ports:

```
switch(config-ng-oam-loop-detection)# disable vlan 1200 port ethernet 1/1
```

Disables on specific VLANs:

```
switch(config-ng-oam-loop-detection)# disable vlan 1300
```

The **no** form of this command resumes active monitoring of these VLANs or ports.

Step 3 (Optional) Run the [**no**] **periodic-probe-interval** *value* command to specify how often periodic loop-detection probes are sent.

Example:

```
switch(config-ng-oam-loop-detection)# periodic-probe-interval 200
```

Range: 60 seconds to 3600 seconds (60 minutes). Default: 300 seconds (5 minutes).

Step 4 (Optional) Run the [**no**] **port-recovery-interval** *value* command to specify how often recovery probes are sent when a port or VLAN is shut down.

Example:

```
switch(config-ng-oam-loop-detection)# port-recovery-interval 300
```

Range: 300 seconds to 3600 seconds (60 minutes). Default value: 600 seconds (10 minutes).

Step 5 (Optional) Run the **show ngoam loop-detection summary** command to verify the loop-detection configuration and current loop summary.

Example:

```
switch# show ngoam loop-detection summary
Loop detection:enabled
Periodic probe interval: 200
Port recovery interval: 300
Number of vlans: 1
Number of ports: 1
Number of loops: 1
Number of ports blocked: 1
Number of vlans disabled: 0
Number of ports disabled: 0
Total number of probes sent: 214
Total number of probes received: 102
Next probe window start: Thu May 14 15:14:23 2020 (0 seconds)
Next recovery window start: Thu May 14 15:54:23 2020 (126 seconds)
```

What to do next

Configure a QoS policy on the spine. (For configuration example, see Configuration Examples for NGOAM Southbound Loop Detection and Mitigation, on page 482).

Configure NGOAM Southbound Loop Detection on Layer-3 Interfaces

Follow these steps to enable NGOAM Southbound Loop Detection on Ethernet and L3 port-channel interfaces.

Procedure

Step 1 Run the [**no**] **ngoam loop-detection** command in global configuration mode, to enable NGOAM Southbound loop detection and mitigation for all VLANs or ports.

Example:

```
switch# config terminal
switch(config)# ngoam loop-detection
switch(config-ng-oam-loop-detection)#
```

This feature is disabled by default.

Step 2 Run the [**no**] **l3 ethernet port** *port-range*command to enable the L3 loop-detection on ethernet interfaces.

Example:

```
switch(config-ng-oam-loop-detection)# l3 ethernet port Eth1/49
```

Use the **no** form of this command to disable the L3 loop-detection on ethernet interfaces.

Step 3 Run the [**no**] **l3 port-channel port** *port-range* command to enable the L3 loop-detection on port-channel interfaces.

Example:

```
switch(config-ng-oam-loop-detection)# l3 port-channel port port-channel1
```

Use the **no** form of this command to disable the L3 loop-detection on port-channel interfaces.

Step 4 (Optional) Run the **show ngoam loop-detection status l3** command to verify the loops detected on L3 interfaces.

Example:

```
switch# show ngoam loop-detection status l3
Port       Status       NumLoops      DetectionTime             ClearedTime
================================================================================
 Eth1/2     BLOCKED      2             Tue Jun 25 02:38:54 2024  Tue Jun 25 02:38:52 2024
```

Step 5 (Optional) Run the **show run ngoam** command to verify the loop-detection configuration and current loop summary.

Example:

```
switch# show run ngoam
ngoam loop-detection
  periodic-probe-interval 60
  port-recovery-interval 600
  l3 ethernet port Ethernet1/1-3
!
2024 Jun 25 02:37:50 switch %ETHPORT-5-IF_DOWN_NONE: Interface Ethernet1/2 is down (None)
2024 Jun 25 02:37:50 switch %ETHPORT-5-IF_DOWN_ERROR_DISABLED: Interface Ethernet1/2 is down (Error
 disabled. Reason:error)
2024 Jun 25 02:38:52 switch %ETHPORT-5-IF_ERRDIS_RECOVERY: Interface Ethernet1/2 is being recovered
 from error disabled state (Last Reason:error)
2024 Jun 25 02:38:54 switch %ETHPORT-5-IF_DOWN_ERROR_DISABLED: Interface Ethernet1/2 is down (Error
 disabled. Reason:error)
```

Detect Loops and Bring Up Ports On Demand

Follow the steps in this section to detect loops or bring up blocked ports on demand.

Procedure

Step 1 (Optional) Run the **ngoam loop-detection probe** {**vlan** *vlan-range*} [**port** *port-range*] command to sends a loop-detection probe on the specified VLAN or port.

Example:

```
switch# ngoam loop-detection probe vlan 1200 port ethernet 1/1
```

This command also send a notification to check whether the probe was successfully sent.

Step 2 (Optional) Run the **ngoam loop-detection bringup** {**vlan** *vlan-range*} [**port** *port-range*] command to bring UP the VLANs or ports that were blocked earlier.

Example:

```
switch# ngoam loop-detection bringup vlan 1200 port ethernet 1/1
```

This command also clears any entries stuck in the NGOAM.

Note
It can take up to two port-recovery intervals for the ports to come up after a loop is cleared. You can speed up the recovery by manually overriding the timer with the **ngoam loop-detection bringup vlan** {**vlan** *vlan-range*} [**port** *port-range*] command.

Step 3 (Optional) Run the **show ngoam loop-detection status** [**history**] [**vlan** *vlan-range*] [**port** *port-range*] command to verify the loop-detection status for the VLAN or port with and without the **history** option..

Example:

Without **history** option

```
switch# show ngoam loop-detection status
VlanId Port   Status     NumLoops  Detection Time                 ClearedTime
====== ====== ========== ========= ============================== ===============
100    Eth1/3 BLOCKED     1        Tue Apr 14 20:07:50.313 2020  Never
```

With **history** option

```
switch# show ngoam loop-detection status history
VlanId Port   Status     NumLoops  Detection Time                 ClearedTime
====== ====== ========== ========= ============================== ===============
100    Eth1/3 BLOCKED     1        Tue Apr 14 20:07:50.313 2020  Never
200    Eth1/2 FORWARDING  1        Tue Apr 14 21:19:52.215 2020  May 11 21:30:54.830 2020
```

The status can be one of the following:

- **BLOCKED**: The VLAN or port is shut down because a loop has been detected.
- **FORWARDING**: A loop has not been detected, and the VLAN or port is operational.
- **RECOVERING**: Recovery probes are being sent to determine if a previously detected loop still exists.

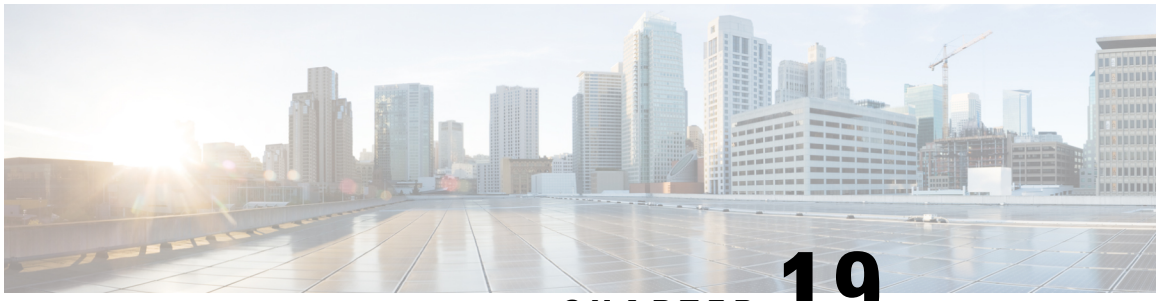
The **history** option displays blocked, forwarding, and recovering ports. Without the **history** option, the command displays only blocked and recovering ports.

Configuration Examples for NGOAM Southbound Loop Detection and Mitigation

The following example hows to configure a QoS policy on the spine and apply it to all of the spine interfaces to which the loop-detection-enabled leaf is connected:

```
class-map type qos match-any Spine-DSCP56
match dscp 56
policy-map type qos Spine-DSCP56
class Spine-DSCP56
set qos-group 7

interface Ethernet1/31
mtu 9216
no link dfe adaptive-tuning
service-policy type qos input Spine-DSCP5663
no ip redirects
ip address 27.4.1.2/24
ip router ospf 200 area 0.0.0.0
ip pim sparse-mode
no shutdown
```



CHAPTER 19

Configuring VXLAN QoS

This chapter contains these sections:

- Information About VXLAN QoS, on page 483
- Guidelines and Limitations for VXLAN QoS, on page 493
- Default Settings for VXLAN QoS, on page 496
- Configuring VXLAN QoS, on page 497
- Verifying the VXLAN QoS Configuration, on page 500
- VXLAN QoS Configuration Examples, on page 500

Information About VXLAN QoS

VXLAN QoS enables you to provide Quality of Service (QoS) capabilities to traffic that is tunneled in VXLAN.

Traffic in the VXLAN overlay can be assigned to different QoS properties:

- Classification traffic to assign different properties.
- Including traffic marking with different priorities.
- Queuing traffic to enable priority for the protected traffic.
- Policing for misbehaving traffic.
- Shaping for traffic that limits speed per interface.
- Properties traffic sensitive to traffic drops.

Note QoS allows you to classify the network traffic, police and prioritize the traffic flow, and provide congestion avoidance. For more information about QoS, see the Cisco Nexus 9000 Series NX-OS Quality of Service Configuration Guide, Release 7.x.

Note QoS allows you to classify the network traffic, police and prioritize the traffic flow, and provide congestion avoidance. For more information about QoS, see the Cisco Nexus 9000 Series NX-OS Quality of Service Configuration Guide, Release 9.2(x).

This section contains the following topics:

VXLAN QoS Terminology

This section defines VXLAN QoS terminology.

Table 8: VXLAN QoS Terminology

Term	Definition
Frames	Carries traffic at Layer 2. Layer 2 frames carry Layer 3 packets.
Packets	Carries traffic at Layer 3.
VXLAN packet	Carries original frame, encapsulated in VXLAN IP/UDP header.
Original frame	A Layer 2 or Layer 2 frame that carries the Layer 3 packet before encapsulation in a VXLAN header.
Decapsulated frame	A Layer 2 or a Layer 2 frame that carries a Layer 3 packet after the VXLAN header is decapsulated.
Ingress VTEP	The point where traffic is encapsulated in the VXLAN header and enters the VXLAN tunnel.
Egress VTEP	The point where traffic is decapsulated from the VXLAN header and exits the VXLAN tunnel.
Class of Service (CoS)	Refers to the three bits in an 802.1Q header that are used to indicate the priority of the Ethernet frame as it passes through a switched network. The CoS bits in the 802.1Q header are commonly referred to as the 802.1p bits. 802.1Q is discarded prior to frame encapsulation in a VXLAN header, where CoS value is not present in VXLAN tunnel. To maintain QoS when a packet enters the VXLAN tunnel, the type of service (ToS) and CoS values map to each other.
IP precedence	The 3 most significant bits of the ToS byte in the IP header.
Differentiated Services Code Point (DSCP)	The first six bits of the ToS byte in the IP header. DSCP is only present in an IP packet.
Explicit Congestion Notification (ECN)	The last two bits of the ToS byte in the IP header. ECN is only present in an IP packet.

Term	Definition
QoS tags	Prioritization values carried in Layer 3 packets and Layer 2 frames. A Layer 2 CoS label can have a value ranging between zero for low priority and seven for high priority. A Layer 3 IP precedence label can have a value ranging between zero for low priority and seven for high priority. IP precedence values are defined by the three most significant bits of the 1-byte ToS byte. A Layer 3 DSCP label can have a value between 0 and 63. DSCP values are defined by the six most significant bits of the 1-byte IP ToS field.
Classification	The process used for selecting traffic for QoS
Marking	The process of setting: a Layer 2 COS value in a frame, Layer 3 DSCP value in a packet, and Layer 3 ECN value in a packet. Marking is also the process of choosing different values for the CoS, DSCP, ECN field to mark packets so that they have the priority that they require during periods of congestion.
Policing	Limiting bandwidth used by a flow of traffic. Policing can mark or drop traffic.
MQC	The Cisco Modular QoS command line interface (MQC) framework, which is a modular and highly extensible framework for deploying QoS.

VXLAN QoS Features

The following topics describe the VXLAN QoS features that are supported in a VXLAN network:

Trust Boundaries

The trust boundary forms a perimeter on your network. Your network trusts (and does not override) the markings on your switch. The existing ToS values are trusted when received on in the VXLAN fabric.

Classification

You use classification to partition traffic into classes. You classify the traffic based on the port characteristics or the packet header fields that include IP precedence, differentiated services code point (DSCP), Layer 3 to Layer 4 parameters, and the packet length.

The values used to classify traffic are called match criteria. When you define a traffic class, you can specify multiple match criteria, you can choose to not match on a particular criterion, or you can determine the traffic class by matching any or all criteria.

Traffic that fails to match any class is assigned to a default class of traffic called class-default.

Marking

Marking is the setting of QoS information that is related to a packet. Packet marking allows you to partition your network into multiple priority levels or classes of service. You can set the value of a standard QoS field for COS, IP precedence, and DSCP. You can also set the QoS field for internal labels (such as QoS groups) that can be used in subsequent actions. Marking QoS groups is used to identify the traffic type for queuing and scheduling traffic.

Policing

Policing causes traffic that exceeds the configured rate to be discarded or marked down to a higher drop precedence.

Single-rate policers monitor the specified committed information rate (CIR) of traffic. Dual-rate policers monitor both CIR and peak information rate (PIR) of traffic.

Queuing and Scheduling

The queuing and scheduling process allows you to control the queue usage and the bandwidth that is allocated to traffic classes. You can then achieve the desired trade-off between throughput and latency.

You can limit the size of the queues for a particular class of traffic by applying either static or dynamic limits.

You can apply weighted random early detection (WRED) to a class of traffic, which allows packets to be dropped based on the QoS group. The WRED algorithm allows you to perform proactive queue management to avoid traffic congestion.

ECN can be enabled along with WRED on a particular class of traffic to mark the congestion state instead of dropping the packets. ECN marking in the VXLAN tunnel is performed in the outer header, and at the Egress VTEP is copied to decapsulated frame.

Traffic Shaping

You can shape traffic by imposing a maximum data rate on a class of traffic so that excess packets are retained in a queue to smooth (constrain) the output rate. In addition, minimum bandwidth shaping can be configured to provide a minimum guaranteed bandwidth for a class of traffic.

Traffic shaping regulates and smooths out the packet flow by imposing a maximum traffic rate for each port's egress queue. Packets that exceed the threshold are placed in the queue and are transmitted later. Traffic shaping is similar to Traffic Policing, but the packets are not dropped. Because packets are buffered, traffic shaping minimizes packet loss (based on the queue length), which provides better traffic behavior for TCP traffic.

By using traffic shaping, you can control the following:

- Access to available bandwidth.
- Ensure that traffic conforms to the policies established for it.
- Regulate the flow of traffic to avoid congestion that can occur when the egress traffic exceeds the access speed of its remote, target interface.

For example, you can control access to the bandwidth when the policy dictates that the rate of a given interface must not, on average, exceed a certain rate. Despite the access rate exceeding the speed.

Network QoS

The network QoS policy defines the characteristics of each CoS value, which are applicable network wide across switches. With a network QoS policy, you can configure the following:

- Pause behavior—You can decide whether a CoS requires the lossless behavior which is provided by using a priority flow control (PFC) mechanism that prevents packet loss during congestion) or not. You can configure drop (frames with this CoS value can be dropped) and no drop (frames with this CoS value cannot be dropped). For the drop and no drop configuration, you must also enable PFC per port. For more information about PFC, see “Configuring Priority Flow Control".

Pause behavior can be achieved in the VXLAN tunnel for a specific queue-group.

VXLAN Priority Tunneling

In the VXLAN tunnel, DSCP values in the outer header are used to provide QoS transparency in end-to-end of the tunnel. The outer header DSCP value is derived from the DSCP value with Layer 3 packets or the CoS value for Layer 2 frames. At the VXLAN tunnel egress point, the priority of the decapsulated traffic is chosen based on the mode. For more information, see Decapsulated Packet Priority Selection, on page 491.

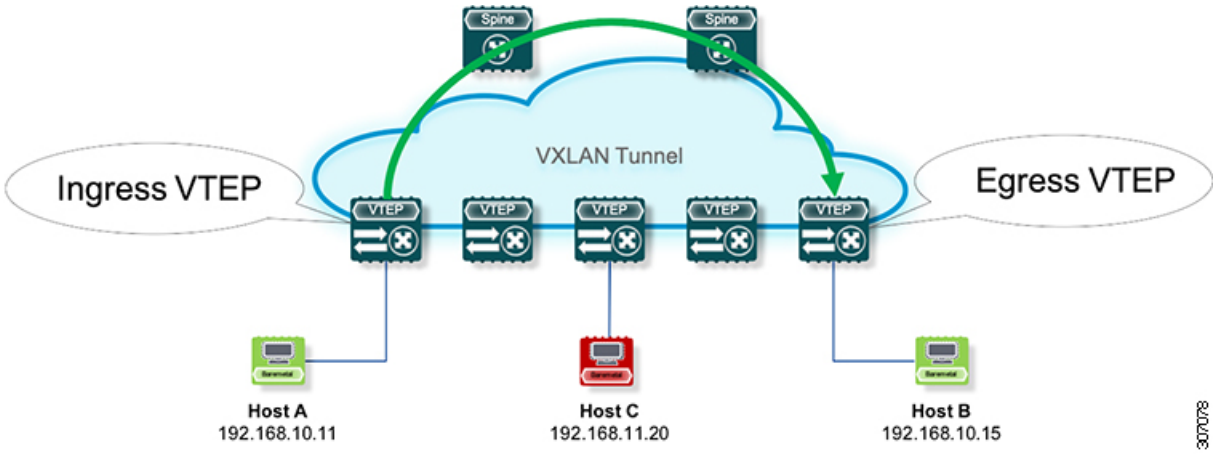
MQC CLI

All available QoS features for VXLAN QoS are managed from the modular QoS command-line interface (CLI). The Modular QoS CLI (MQC) allows you to define traffic classes (class maps), create and configure traffic policies (policy maps), and perform actions that are defined in the policy maps to interface (service policy).

VXLAN QoS Topology and Roles

This section describes the roles of network devices in implementing VXLAN QoS.

Figure 40: VXLAN Network



The network is bidirectional, but in the previous image, traffic is moving left to right.

In the VXLAN network, points of interest are ingress VTEPs where the original traffic is encapsulated in a VXLAN header. Spines are transporting hops that connect ingress and egress VTEPs. An egress VTEP is the point where VXLAN encapsulated traffic is decapsulated and egresses the VTEP as classical Ethernet traffic.



Note Ingress and egress VTEPs are the boundary between the VXLAN tunnel and the IP network.

This section contains the following topics:

Ingress VTEP and Encapsulation in the VXLAN Tunnel

At the ingress VTEP, the VTEP processes packets as follows:

Procedure

Step 1 Layer 2 or Layer 3 traffic enters the edge of the VXLAN network.

Step 2 The switch receives the traffic from the input interface and uses the 802.1p bits or the DSCP value to perform any classification, marking, and policing. It also derives the outer DSCP value in the VXLAN header. For classification of incoming IP packets, the input service policy can also use access control lists (ACLs).

Step 3 For each incoming packet, the switch performs a lookup of the IP address to determine the next hop.

Step 4 The packet is encapsulated in the VXLAN header. The encapsulated packet's VXLAN header is assigned a DSCP value that is based on QoS rules.

Step 5 The switch forwards the encapsulated packets to the appropriate output interface for processing.

Step 6 The encapsulated packets, marked by the DSCP value, are sent to the VXLAN tunnel output interface.

Transport Through the VXLAN Tunnel

In the transport through a VXLAN tunnel, the switch processes the VXLAN packets as follows:

Procedure

Step 1 The VXLAN encapsulated packets are received on an input interface of a transport switch. The switch uses the outer header to perform classification, marking, and policing.

Step 2 The switch performs a lookup on the IP address in the outer header to determine the next hop.

Step 3 The switch forwards the encapsulated packets to the appropriate output interface for processing.

Step 4 VXLAN sends encapsulated packets through the output interface.

Egress VTEP and Decapsulation of the VXLAN Tunnel

At the egress VTEP boundary of the VXLAN tunnel, the VTEP processes packets as follows:

Procedure

Step 1 Packets encapsulated in VXLAN are received at the NVE interface of an egress VTEP, where the switch uses the inner header DSCP value to perform classification, marking, and policing.

Step 2 The switch removes the VXLAN header from the packet, and does a lookup that is based on the decapsulated packet's headers.

Step 3 The switch forwards the decapsulated packets to the appropriate output interface for processing.

Step 4 Before the packet is sent out, a DSCP value is assigned to a Layer 3 packet based on the decapsulation priority or based on marking Layer 2 frames.

Step 5 The decapsulated packets are sent through the outgoing interface to the IP network.

Classification at the Ingress VTEP, Spine, and Egress VTEP

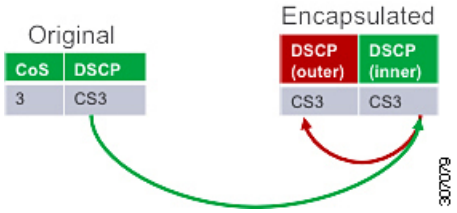
This section includes the following topics:

IP to VXLAN

At the ingress VTEP, the ingress point of the VXLAN tunnel, traffic is encapsulated in the VXLAN header. Traffic on an ingress VTEP is classified based on the priority in the original header. Classification can be performed by matching the CoS, DSCP, and IP precedence values or by matching traffic with the ACL based on the original frame data.

When traffic is encapsulated in the VXLAN, the Layer 3 packet's DSCP value is copied from the original header to the outer header of the VXLAN encapsulated packet. This behavior is illustrated in the following figure:

Figure 41: Copy of Priority from Layer-3 Packet to VXLAN Outer Header



For Layer 2 frames without the IP header, the DSCP value of the outer header is derived from the CoS-to-DSCP mapping present in the hardware illustrated in Default Settings for VXLAN QoS, on page 496. In this way, the original QoS attributes are preserved in the VXLAN tunnel. This behavior is illustrated in the following figure:

Figure 42: Copy of Priority from Layer-2 Frame to VXLAN Outer Header



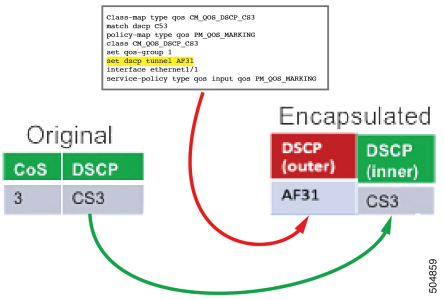
A Layer 2 frame, does not have a DSCP value present because the IP header is not present in the frame. After a Layer 2 frame is encapsulated, the original CoS value is not preserved in the VXLAN tunnel.

IP to VXLAN with Outer DSCP

Beginning with Cisco NX-OS Release 10.4(1)F, the policy with set outer DSCP action can be applied to the access interface in the ingress direction.

When traffic is encapsulated in the VXLAN, for Layer-3 packets, DSCP value from the original packet is copied to the inner header and the user configured DSCP value is set in the outer header of the VXLAN encapsulated packet. This behavior is illustrated in the following figure:

Figure 43: VXLAN Outer DSCP Value Applied from Set Configuration

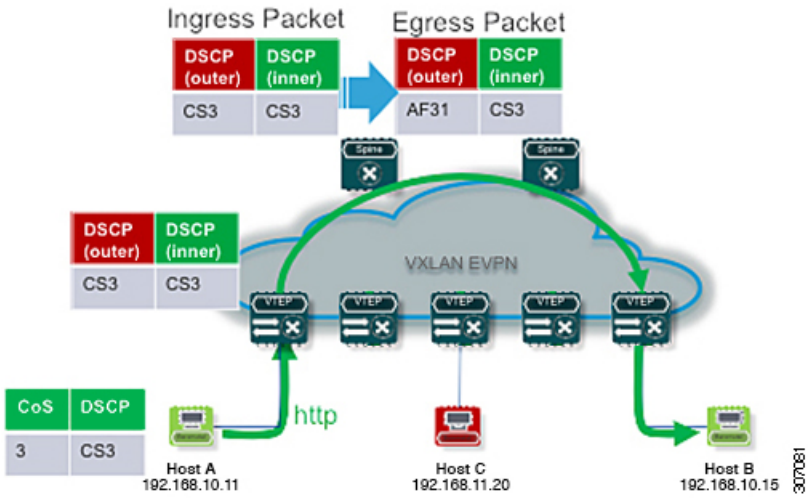


Inside the VXLAN Tunnel

Inside the VXLAN tunnel, traffic classification is based on the outer header DSCP value. Classification can be done matching the DCSP value or using ACLs for classification.

If VXLAN encapsulated traffic is crossing the trust boundary, marking can be changed in the packet to match QoS behavior in the tunnel. Marking can be performed inside of the VXLAN tunnel, where a new DSCP value is applied only on the outer header. The new DSCP value can influence different QoS behaviors inside the VXLAN tunnel. The original DSCP value is preserved in the inner header.

Figure 44: Marking Inside of the VXLAN Tunnel



VXLAN to IP

Classification at the egress VTEP is performed for traffic leaving the VXLAN tunnel. For classification at the egress VTEP, the inner header and outer DSCP values are used. The inner or outer DSCP value is used for priority-based classification. Classification can be performed using ACLs.

Classification is performed on the NVE interface for all VXLAN tunneled traffic.

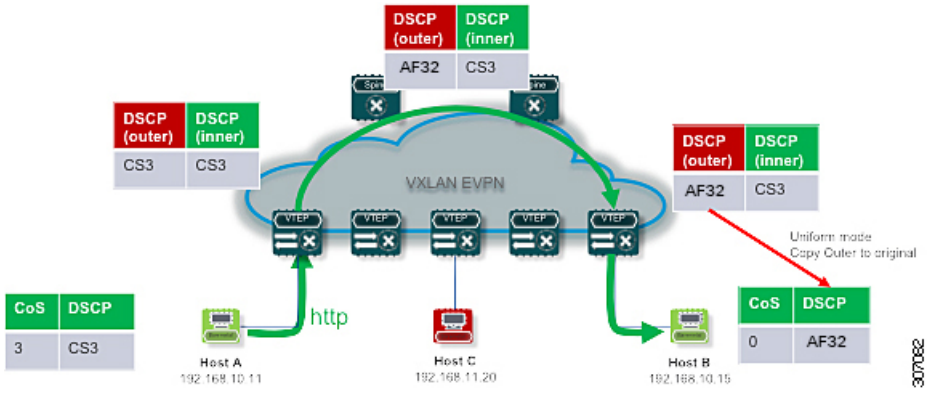
Marking and policing can be performed on the NVE interface for tunneled traffic. If marking is configured, newly marked values are present in the decapsulated packet. Because the original CoS value is not preserved in the encapsulated packet, marking can be performed for decapsulated packets for any devices that expect an 802.1p field for QoS in the rest of the network.

Decapsulated Packet Priority Selection

At the egress VTEP, the VXLAN header is removed from the packet and the decapsulated packet egresses the switch with the DSCP value. The switch assigns the DSCP value of the decapsulated packet based on two modes:

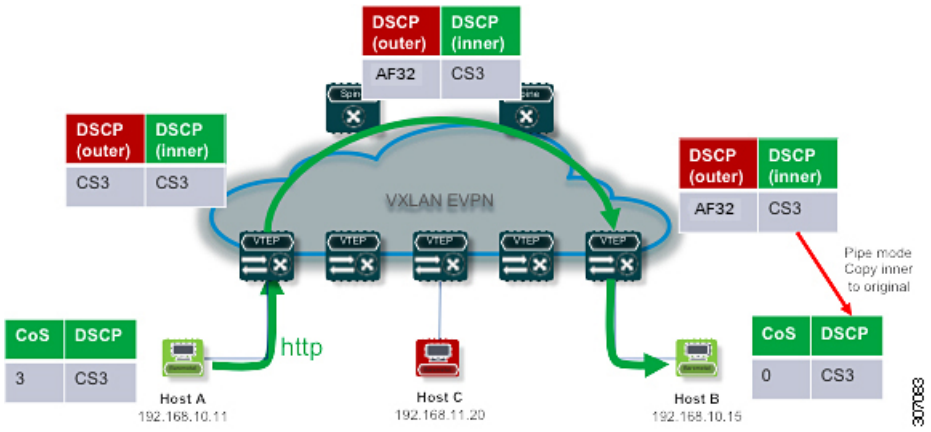
- Uniform mode – the DSCP value from the outer header of the VXLAN packet is copied to the decapsulated packet. Any change of the DSCP value in the VXLAN tunnel is preserved and present in the decapsulated packet. Uniform mode is the default mode of decapsulated packet priority selection.

Figure 45: Uniform Mode Outer DSCP Value is Copied to Decapsulated Packet DSCP Value for a Layer-3 Packet



- Pipe mode – the original DSCP value is preserved at the VXLAN tunnel end. At the egress VTEP, the system copies the inner DSCP value to the decapsulated packet DSCP value. In this way, the original DSCP value is preserved at the end of the VXLAN tunnel.

Figure 46: Pipe Mode Inner DSCP Value is Copied to Decapsulated Packet DSCP Value for Layer-3 Packet



CoS Preservation

Beginning with Cisco NX-OS Release 10.4(1)F, the **default-vxlan-in-tnl-dscp-policy** QoS policy-map template is added to provide CoS preservation for non-IP packets.

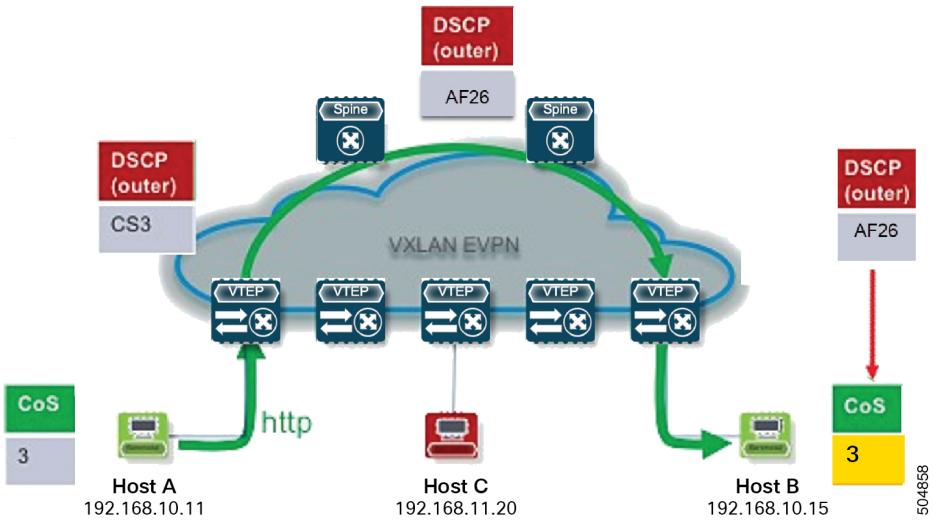
When this template is enabled on the NVE interface, the switch performs a match on the outer DSCP of the VXLAN packet and rewrites the CoS in the decapsulated ethernet packet on the egress VTEP based on a fixed outer DSCP to CoS mapping.

The following table lists the default Outer DSCP-to-CoS mapping in the Egress VTEP for Layer 2 frames:

Table 9: Default Outer DSCP-to-CoS Mapping

DSCP of Outer VXLAN Header	CoS of Original Layer 2 Frame
0	0
8	1
16	2
26	3
32	4
46	5
48	6
56	7

Figure 47: Non-IP CoS Value Restored in Decapsulated Packet



Guidelines and Limitations for VXLAN QoS

Note The QoS policy must be configured end-to-end for this feature to work as designed.

VXLAN QoS has the following configuration guidelines and limitations:

- Cisco Nexus 9300-FX/FX2/FX3 platform switches and Cisco Nexus 9500 platform switches with X97160YC-EX, 9700-FX or -R/RX line cards support VXLAN QoS.
- Beginning with Cisco NX-OS Release 9.3(3), Cisco Nexus 9300-GX platform switches support VXLAN QoS in default mode.
- Beginning with Cisco NX-OS Release 10.2(3)F, VXLAN QoS in default mode is supported on Cisco Nexus 9300-GX2 platform switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, VXLAN QoS in default mode is supported on Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, VXLAN QoS in default mode is supported on Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, VXLAN QoS in default mode is supported on Cisco Nexus 9364C-H1 switches.
- Beginning with Cisco NX-OS Release 10.5(2)F, VXLAN QoS is supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.
- The following features are supported on Cisco Nexus 9504 and 9508 platform switches with -R/RX line cards:
  - Physical interface level queuing should work as normal L2/L3 queuing/QoS

    - IPv4 bridged case works in terms of copying inner ToS to outer VXLAN ToS

- The following features are not supported on Cisco Nexus 9504 and 9508 platform switches with -R and -RX line cards:
    - Policies on the NVE interface
    - IPv6 type of service (ToS) from inner to VXLAN outer copying
    - IPv4 routed cases for QoS. ToS from inner is not copied to outer VXLAN header
- For Cisco Nexus 9504 and 9508 platform switches with -RX line cards, the default mode is pipe for VXLAN decapsulation (inner packet DSCP not modified based on outer IP header DSCP value). This is a difference in behavior from other line cards types. If -RX line cards and other line cards are used in the same network, the **qos-mode pipe** command can be used in switches where non-RX line cards are present in order to have the same behavior. For details on the configuration command, see Configuring Type QoS on the Egress VTEP, on page 497.
- VXLAN QoS is supported in the EVPN fabric.
- The original IEEE 802.1Q header is not preserved in the VXLAN tunnel. The CoS value is not present in the inner header of the VXLAN-encapsulated packet.
- Statistics (counters) are present for the NVE interface.
- Egress policing is not supported on outgoing interface (uplink connecting to spine) of the encap (ingress) VXLAN VTEP.
- In a vPC, configure the change of the decapsulated packet priority selection on both peers.
- The service policy on an NVE interface can attach only in the input direction.
- A classification policy applied to an NVE interface applies only on VXLAN-encapsulated traffic. For all other traffic, the classification policy must be applied on the incoming interface.
- To mark the decapsulated packet with a CoS value, a marking policy must be attached to the NVE interface to mark the CoS value to packets where the VLAN header is present.
- The following guidelines and limitations apply to VXLAN QoS configuration on the DCI handoff node:
    - Beginning with Cisco NX-OS Release 9.3(5), Cisco Nexus 9300-GX platform switches support VXLAN QoS configuration on the DCI handoff node.
    - Beginning with Cisco NX-OS Release 10.2(3)F, Cisco Nexus 9300-GX2 platform switches support VXLAN QoS configuration on the DCI handoff node.
    - Beginning with Cisco NX-OS Release 10.4(1)F, Cisco Nexus 9332D-H2R switches support VXLAN QoS configuration on the DCI handoff node.
    - Beginning with Cisco NX-OS Release 10.4(2)F, Cisco Nexus 93400LD-H1 switches support VXLAN QoS configuration on the DCI handoff node.
    - Beginning with Cisco NX-OS Release 10.4(3)F, Cisco Nexus 9364C-H1 switches support VXLAN QoS configuration on the DCI handoff node.
    - VXLAN QoS configuration on the DCI handoff node does not support end-to-end priority flow control (PFC) for Cisco Nexus 9336C-FX2, 93240YC-FX2, and 9300-GX platform switches.

- Microburst, dynamic packet prioritization (DPP), and approximate fair-drop (AFD) are supported on VXLAN-encapsulated packets.

- The following guidelines and limitations apply to the outer DSCP based VXLAN QoS Policy feature:
  - Beginning with Cisco NX-OS Release 10.4(1)F, the outer DSCP based VXLAN QoS Policy feature is supported on Cisco Nexus 9300-FX2/FX3/GX/GX2 platform switches and 9500 switch with N9K-X9716D-GX line card.
  - Beginning with Cisco NX-OS Release 10.4(2)F, the outer DSCP based VXLAN QoS Policy feature is supported on Cisco Nexus 9332D-H2R, and 93400LD-H1 switches.
  - Beginning with Cisco NX-OS Release 10.4(3)F, the outer DSCP based VXLAN QoS Policy feature is supported on Cisco Nexus 9364C-H1 switches.
  - Beginning with Cisco NX-OS Release 10.5(2)F, the outer DSCP based VXLAN QoS Policy feature is supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.
  - In VXLAN QoS policy, the **match dscp tunnel** command can only be applied on the NVE interface and in ingress direction.
  - In VXLAN QoS Policy, both inner and outer DSCP match rules is not supported. However, the match criteria like **ip access-lists** or **mac access-list** in the same policy applied on the NVE interface would always match on the inner header.
  - For non-IP packets, the outer header QoS policy under the NVE interface only supports L2 rewrite and traffic class assignment or outgoing queue. Action like policer is not supported.
  - In VXLAN QoS policy, the **match dscp tunnel** command on the NVE interface performs match on the VXLAN packets destined to the current VTEP, where tunnel termination happens, and packets are decapsulated.
  - In VXLAN QoS policy, the **match dscp tunnel** command does not support non-IP packets, due to this the CoS preservation will not work on IPv6 underlay.
  - In VXLAN QoS policy, the **set dscp tunnel** command does not support non-IP packets. In case of non-IP packets, the outer DSCP value is applied based on the default CoS to DSCP mapping information on the switch.
  - In VXLAN QoS policy, the **set dscp tunnel** command cannot be applied to the NVE interface, as this command is applicable to encapsulation packets.
  - If the **set dscp tunnel** command is applied on the ingress VTEP of VXLAN multisite, the outer DSCP value could be replaced with the inner DSCP if pipe mode is configured on the border gateway. We recommend to configure uniform mode on the border gateway to carry the new outer DSCP header to the remote site.
  - The outer DSCP based VXLAN QoS Policy feature is not supported on the VXLAN multisite deployments.

- The following limitations apply to the VXLAN QoS policies when using a Border Gateway (BGW) Spine:
  - If QoS policies are needed for intra-site BUM traffic for VNI with multicast underlay, and that multicast underlay group is also owned by a VNI defined on the BGW Spine, then the QoS policy must be applied to the NVE interface. QoS policies applied to fabric interfaces will not modify these flows since the NVE interface acts as an incoming interface.

- If QoS policies are needed for intra-site BUM traffic for VNI with multicast underlay, and that multicast group is not owned by a VNI defined on the BGW Spine, then the QoS policy must be applied to a fabric interface. QoS policies applied to the NVE interface will not modify these flows since the NVE is not considered an incoming interface.
- If the NVE interface of the BGW Spine owns a multicast group used for BUM traffic within the local fabric, QoS policies cannot be applied to both the fabric interfaces and NVE interface to differentiate treatment of intra-site and inter-site flows for that multicast group.

- Beginning with Cisco NX-OS Release 10.4(3)F, Cisco Nexus 9808/9804 switches with X9836DM-A and X98900CD-A line cards support VXLAN QoS policies when using a BGW spine with the following limitations:
  - Physical ingress QoS policy and system-level QoS policy are supported.
  - QoS policy on NVE is not supported.
  - Explicit Congestion Notification (ECN) or ECN-Capable Transport (ECT) marking is not retained.
- Beginning with Cisco NX-OS Release 10.5(3)F, on the Cisco Nexus 9364E-SG2-Q and 9364E-SG2-O switches, VXLAN QoS is supported. Below are the functionalities and limitations:
  - You can configure options to manage DSCP values for VXLAN tunnel traffic, ensuring appropriate traffic handling and prioritization.
  - 2 VXLAN QoS modes are being introduced:
    - **Uniform Mode**: The default mode that propagates DSCP values from inner to outer headers, maintaining consistent traffic classification.
    - **Pipe Mode**: An alternative mode where DSCP values are maintained independently between inner and outer headers, allowing for differentiated traffic management.
  - VXLAN QoS configurations are specific to DSCP management and do not extend to other QoS features beyond tunnel traffic handling.
  - System QoS policies are capable of managing traffic based on the DSCP value in the outer IP header of a VXLAN packet during decapsulation.
  - QoS policy on NVE is not supported.
- Beginning with Cisco NX-OS Release 10.5(3)F, on the Cisco Nexus 9364E-SG2-Q and 9364E-SG2-O switches, RDMA over Converged Ethernet version 2 (RoCEv2) over VXLAN is supported.

Default Settings for VXLAN QoS

The following table lists the default CoS-to-DSCP mapping in the ingress VTEP for Layer 2 frames:

Table 10: Default CoS-to-DSCP Mapping

CoS of Original Layer 2 Frame	DSCP of Outer VXLAN Header
0	0

CoS of Original Layer 2 Frame	DSCP of Outer VXLAN Header
1	8
2	16
3	26
4	32
5	46
6	48
7	56

Configuring VXLAN QoS

Configuration of VXLAN QoS is done using the MQC model. The same configuration that is used for the QoS configuration applies to VXLAN QoS. For more information about configuring QoS, see the Cisco Nexus 9000 Series NX-OS Quality of Service Configuration Guide, Release 9.2(x).

VXLAN QoS introduces a new service-policy attachment point which is NVE – Network Virtual Interface. At the egress VTEP, the NVE interface is the point where traffic is decapsulated. To account for all VXLAN traffic, the service policy must be attached to an NVE interface.

The next section describes the configuration of the classification at the egress VTEP, and **service-policy type qos** attachment to an NVE interface.

Configuring Type QoS on the Egress VTEP

Configuration of VXLAN QoS is done by using the MQC model. The same configuration is used for QoS configuration for VXLAN QoS. For more information about configuring QoS, see the Cisco Nexus 9000 Series NX-OS Quality of Service Configuration Guide, Release 9.2(x).

VXLAN QoS introduces a new service-policy attachment point which is the Network Virtual Interface (NVE). At the egress VTEP, the NVE interface points where traffic is decapsulated. To account for all VXLAN traffic, the service policy must be attached to an NVE interface.

This procedure describes the configuration of classification at the egress VTEP, and **service-policy type qos** attachment to an NVE interface.

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.

	Command or Action	Purpose
Step 2	**[no] class-map [type [qos]] | [match-all] | [match-any]** *class-map-name* **Example:** `switch(config)# class-map type qos class1`	Creates or accesses the class map *class--map-name* and enters **class-map** mode. The *class--map-name* argument can contain alphabetic, hyphen, or underscore characters, and can be up to 40 characters. (match-any is the default when the **no** option is selected and multiple match statements are entered.)
Step 3	**[no] match [access-group | cos | dscp [tunnel] | precedence]** *{name | 0-7 | 0-63 | 0-7}* **Example:** `switch(config-cmap-qos)# match dscp tunnel 26`	Configures the traffic class by matching packets based on access-list, **cos** value, **dscp** values, or IP **precedence** value Beginning with Cisco NX-OS Release 10.4(1)F, the **tunnel** option is provided to match the DSCP value on the outer VXLAN header of the ingress packet. **Note** The **match dscp tunnel** command will be used in an ingress service policy that will be applied to the NVE interface on the egress VTEP.
Step 4	**[no] policy-map type qos** *policy-map-name* **Example:** `switch(config-cmap-qos)# policy-map type qos policy`	Creates or accesses the policy map that is named *policy-map-name* and then enters policy-map mode. The policy-map name can contain alphabetic, hyphen, or underscore characters, is case sensitive, and can be up to 40 characters.
Step 5	**[no] class** *class-name* **Example:** `switch(config-pmap-qos)# class class1`	Creates a reference to class-name and enters policy-map class configuration mode. The class is added to the end of the policy map unless insert-before is used to specify the class to insert before. Use the class-default keyword to select all traffic that is not currently matched by classes in the policy map.
Step 6	**[no] set qos-group** *qos-group-value* **Example:** `switch(config-pmap-c-qos)# set qos-group 1`	Sets the QoS group value to *qos-group-value*. The value can range from 1 through 126. The **qos-group** is referenced in type queuing and type network-qos as matching criteria.
Step 7	**exit** **Example:** `switch(config-pmap-c-qos)# exit`	Exits class-map mode.
Step 8	**[no] interface nve** *nve-interface-number* **Example:** `switch(config)# interface nve 1`	Enters interface mode to configure the NVE interface.
Step 9	**[no] service-policy type qos input** *policy-map-name* **Example:** `switch(config-if-nve)# service-policy type qos input policy`	Adds a service-policy *policy-map-name* to the interface in the input direction. You can attach only one input policy to an NVE interface.

	Command or Action	Purpose
Step 10	(Optional) **[no] qos-mode [pipe]** **Example:** `switch(config-if-nve)# qos-mode pipe`	Selecting decapsulated packet priority selection and using pipe mode. Entering the **no** form of this command negates pipe mode and defaults to uniform mode.

Setting Outer DSCP on the Ingress VTEP

VXLAN QoS policy introduces a new outer DSCP set action for all VXLAN traffic, the service policy must be attached to the access (ingress) interface of the ingress VTEP.

SUMMARY STEPS

1. **configure terminal**
2. **[no] class-map [type qos] [match-all]| [match-any]** *class-map-name*
3. **[no] policy-map type qos** *policy-map-name*
4. **[no] class** *class-name*
5. **[no] set dscp [tunnel]** *dscp-val*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**[no] class-map [type qos] [match-all]| [match-any]** *class-map-name* **Example:** `switch(config)# class-map type qos class1`	Creates or accesses the class map *class--map-name* and enters **class-map** mode. The *class--map-name* argument can contain alphabetic, hyphen, or underscore characters, and can be up to 40 characters. (match-any is the default when the **no** option is selected and multiple match statements are entered.)
Step 3	**[no] policy-map type qos** *policy-map-name* **Example:** `switch(config-cmap-qos)# policy-map type qos policy`	Creates or accesses the policy map that is named *policy-map-name* and then enters policy-map mode. The policy-map name can contain alphabetic, hyphen, or underscore characters, is case sensitive, and can be up to 40 characters.
Step 4	**[no] class** *class-name* **Example:** `switch(config-pmap-qos)# class class1`	Creates a reference to class-name and enters policy-map class configuration mode. The class is added to the end of the policy map unless insert-before is used to specify the class to insert before. Use the class-default keyword to select all traffic that is not currently matched by classes in the policy map.

	Command or Action	Purpose
Step 5	**[no] set dscp [tunnel]** *dscp-val* **Example:** `switch(config-pmap-c-qos)# set dscp tunnel 32`	Sets the DSCP value in the outer VXLAN header of the Ingress packet.

Verifying the VXLAN QoS Configuration

Table 11: VXLAN QoS Verification Commands

Command	Purpose
show class map	Displays information about all configured class maps.
show policy-map	Displays information about all configured policy maps.
show running ipqos	Displays configured QoS configuration on the switch.

VXLAN QoS Configuration Examples

Ingress VTEP Classification and Marking

This example shows how to configure the **class-map type qos** command for classification matching traffic with an ACL. Enter the **policy-map type qos** command to put traffic in qos-group 1 and set the DSCP value. Enter the **service-policy type qos** command to attach to the ingress interface in the input direction to classify traffic matching the ACL.

```
access-list ACL_QOS_DSCP_CS3 permit ip any any eq 80

class-map type qos CM_QOS_DSCP_CS3
 match access-group name ACL_QOS_DSCP_CS3

policy-map type qos PM_QOS_MARKING
 class CM_QOS_DSCP_CS3
 set qos-group 1
 set dscp 24

interface ethernet1/1
 service-policy type qos input PM_QOS_MARKING
```

Transit Switch – Spine Classification

This example shows how to configure the **class-map type qos** command for classification matching DSCP 24 set on the ingress VTEP. Enter the **policy-map type qos** command to put traffic in qos-group 1. Enter the **service-policy type qos** command to attach to the ingress interface in the input direction to classify traffic matching criteria.

```
class-map type qos CM_QOS_DSCP_CS3
 match dscp 24
```

```
policy-map type qos PM_QOS_CLASS
 class CM_QOS_DSCP_CS3
 set qos-group 1

interface Ethernet 1/1
 service-policy type qos input PM_QOS_CLASS
```

Egress VTEP Classification and Marking

This example shows how to configure the **class-map type qos** command for classification matching traffic by DSCP value. Enter the **policy-map type qos** to place traffic in qos-group 1 and mark CoS value in outgoing frames. The **service-policy type qos** command is applied to the NVE interface in the input direction to classify traffic coming out of the VXLAN tunnel.

```
class-map type qos CM_QOS_DSCP_CS3
 match dscp 24

policy-map type qos PM_QOS_MARKING
 class CM_QOS_DSCP_CS3
 set qos-group 1
 set cos 3

interface nve 1
 service-policy type qos input PM_QOS_MARKING
```

Queuing

This example shows how to configure the **policy-map type queueing** command for traffic in qos-group 1. Assigning 50% of the available bandwidth to q1 mapped to qos-group 1 and attaching policy in the output direction to all ports using the **system qos** command.

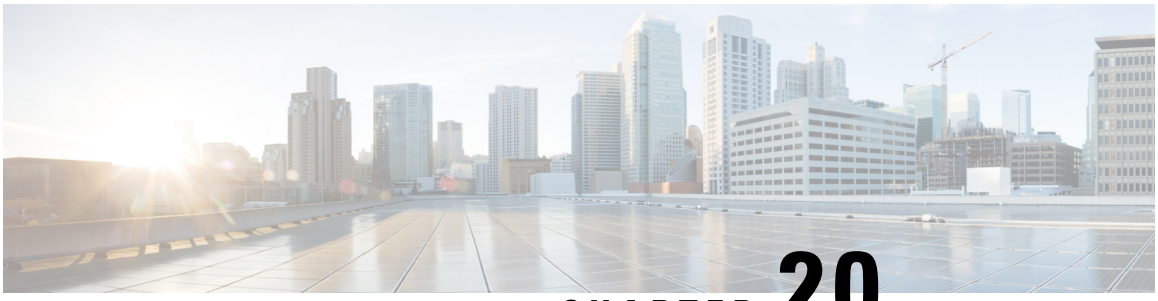
```
policy-map type queuing PM_QUEUING
class type queuing c-out-8q-q7
      priority level 1
    class type queuing c-out-8q-q6
      bandwidth remaining percent 0
    class type queuing c-out-8q-q5
      bandwidth remaining percent 0
    class type queuing c-out-8q-q4
      bandwidth remaining percent 0
    class type queuing c-out-8q-q3
      bandwidth remaining percent 0
    class type queuing c-out-8q-q2
      bandwidth remaining percent 0
    class type queuing c-out-8q-q1
      bandwidth remaining percent 50
    class type queuing c-out-8q-q-default
      bandwidth remaining percent 50

system qos
 service-policy type queueing output PM_QUEUING
```

Preserve CoS Configuration

This example shows how to configure the CoS preservation on NVE interface:

```
interface nve 1
   service-policy type qos input default-vxlan-in-tnl-dscp-policy
```



CHAPTER 20

Configuring BGP EVPN Filtering

This chapter contains these sections:

- About BGP EVPN Filtering, on page 503
- Guidelines and Limitations for BGP EVPN Filtering, on page 504
- Configuring BGP EVPN Filtering, on page 504
- Verifying BGP EVPN Filtering, on page 522

About BGP EVPN Filtering

This feature describes the requirements for route filtering and attributes handling, arising from the implementation of BGP NLRIs of address family L2VPN EVPN.

EVPN routes are quite different from regular IPv4 and IPv6 routes in NLRI format. They contain many fields and carry attributes specific to EVPN. Using route maps, we can filter routes on the basis of these attributes. The following route-filtering options are available for the routes belonging to the EVPN address family:

- Matching based on the EVPN route type: Six types of NLRIs are available in EVPN. Matching is based on the type specified in the route-map match statement.
- Matching based on the IP address in the NLRI: This option can be used to filter based on IP address information in type-2 and type-5 NLRIs.
- Matching based on the MAC address in the NLRI: This option is similar to matching based on the IP address embedded in the NLRI. EVPN type-2 routes contain a MAC address along with an IP address. This option can be used to filter such routes.
- Matching based on the RMAC extended community: EVPN type-2 and type-5 routes carry the router MAC (RMAC) extended community, which carries a MAC address. The RMAC is advertised as part of the update message to the neighbor along with other extended community information. It specifies the MAC address of the remote next hop of a route. This option allows matching against this RMAC extended community.
- Setting the RMAC extended community: This option allows you to change the RMAC extended community value of an EVPN NLRI.
- Setting the EVPN next-hop IP address: This option sets the next-hop IP address of the EVPN route once the match condition has been met. Setting the next-hop IP address for EVPN routes should be accompanied by setting the RMAC extended community to ensure correctness in forwarding.

- Setting the gateway IP address for route type-5: The gateway IP address encodes an overlay IP index for the IP prefixes that form the type-5 EVPN routes. It gets advertised as part of the EVPN NLRI in the update message. The default value is 0.0.0.0. When it's set to any other value, the next hop on the route in the VRF context changes to the gateway IP address specified.
- Using table maps: You can configure table maps to filter MAC routes downloaded to the Layer 2 Routing Information Base (L2RIB).

The rest of this chapter provides information on configuring and applying these options.

Guidelines and Limitations for BGP EVPN Filtering

The following are the guidelines and limitations for BGP EVPN filtering:

Cisco Nexus 9000 Series switches support BGP EVPN filtering.

The following match and set options are available for filtering an EVPN address family of routes:

- Matching based on the route type
- Matching based on the MAC address in the NLRI
- Matching based on the RMAC extended community
- Setting the RMAC extended community
- Setting the EVPN next-hop IP address—If more than one next-hop IP address is configured, only the first one is used and processed if using for EVPN. IPv4 and IPv6 can be used as next-hop addresses.
- Setting the gateway IP address for a route type-5—You can set an IPv4 gateway IP address using the **route-map** command.
- Using table maps—A table map for filtering MAC routes is downloaded to the Layer 2 Routing Information Base (L2RIB).

Configuring BGP EVPN Filtering

To perform route filtering for the EVPN address-family routes, you can perform the following tasks:

- Configuring the Route Map with Match and Set Clauses, on page 504
- Applying the Route Map at the Inbound or Outbound Level, on page 509

To configure the table map, you can perform the following tasks:

- Configuring a MAC List and a Route Map that Matches the MAC List, on page 518
- Applying the Table Map, on page 519

Configuring the Route Map with Match and Set Clauses

You can use the existing route-map configuration along with the match and set clauses to decide the kind of filtering that you need.

• Matching Based on EVPN Route Type, on page 505
• Matching Based on MAC Address in the NLRI, on page 505
• Matching Based on RMAC Extended Community, on page 506
• Setting the RMAC Extended Community, on page 507
• Setting the EVPN Next-Hop IP Address, on page 507
• Setting the Gateway IP Address for Route Type-5, on page 508

Matching Based on EVPN Route Type

SUMMARY STEPS

1. **configure terminal**
2. **route-map** *route-map-name*
3. **match evpn route-type {1 | 2 | 2-mac-ip | 2-mac-only | 3 | 4 | 5 | 6}**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.
Step 2	**route-map** *route-map-name* **Example:** `switch(config)# route-map ROUTE_MAP_1`	Create a route map.
Step 3	**match evpn route-type {1 | 2 | 2-mac-ip | 2-mac-only | 3 | 4 | 5 | 6}** **Example:** `switch(config-route-map)# match evpn route-type 6`	Match BGP EVPN routes.

Matching Based on MAC Address in the NLRI

SUMMARY STEPS

1. **configure terminal**
2. **mac-list** *list-name* **[seq** *seq-number***] {deny | permit}** *mac-address* **[mac-mask]**
3. **route-map** *route-map-name*
4. **match mac-list** *mac-list-name*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.
Step 2	**mac-list** *list-name* **[seq** *seq-number***] {deny | permit}** *mac-address* **[mac-mask]** **Example:** `switch(config)# mac-list MAC_LIST_1 permit E:E:E`	Build a MAC list.
Step 3	**route-map** *route-map-name* **Example:** `switch(config)# route-map ROUTE_MAP_1`	Create a route map.
Step 4	**match mac-list** *mac-list-name* **Example:** `switch(config-route-map)# match mac-list MAC_LIST_1`	Match entries of MAC lists. The maximum length is 63 characters.

Matching Based on RMAC Extended Community

SUMMARY STEPS

1. **configure terminal**
2. **ip extcommunity-list standard** *list-name* **seq** *5* **{deny | permit} rmac** *mac-addr*
3. **route-map** *route-map-name*
4. **match extcommunity** *list-name*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.
Step 2	**ip extcommunity-list standard** *list-name* **seq** *5* **{deny | permit} rmac** *mac-addr* **Example:** `switch(config)# ip extcommunity-list standard EXTCOMM_LIST_RMAC seq 5 permit rmac a8b4.56e4.7edf`	Add an extcommunity list entry. The *list-name* argument must not exceed 63 characters.

	Command or Action	Purpose
Step 3	**route-map** *route-map-name* **Example:** `switch(config)# route-map ROUTE_MAP_1`	Create a route map.
Step 4	**match extcommunity** *list-name* **Example:** `switch(config-route-map)# match extcommunity EXTCOMM_LIST_RMAC`	Match the extended community list name.

Setting the RMAC Extended Community

SUMMARY STEPS

1. **configure terminal**
2. **route-map** *route-map-name*
3. **set extcommunity evpn rmac** *mac-address*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.
Step 2	**route-map** *route-map-name* **Example:** `switch(config)# route-map ROUTE_MAP_1`	Create a route map.
Step 3	**set extcommunity evpn rmac** *mac-address* **Example:** `switch(config-route-map)# set extcommunity evpn rmac EEEE.EEEE.EEEE`	Set the BGP RMAC extcommunity attribute.

Setting the EVPN Next-Hop IP Address

SUMMARY STEPS

1. **configure terminal**
2. **route-map** *route-map-name*
3. **set ip next-hop** *next-hop*
4. **set ipv6 next-hop** *next-hop*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.
Step 2	**route-map** *route-map-name* **Example:** `switch(config)# route-map ROUTE_MAP_1`	Create a route map.
Step 3	**set ip next-hop** *next-hop* **Example:** `switch(config-route-map)# set ip next-hop 209.165.200.226`	Set the IP address of the EVPN IP next hop.
Step 4	**set ipv6 next-hop** *next-hop* **Example:** `switch(config-route-map)# set ipv6 next-hop 2001:0DB8::1`	Set the IPv6 next-hop address.

Setting the Gateway IP Address for Route Type-5

SUMMARY STEPS

1. **configure terminal**
2. **route-map** *route-map-name*
3. **set evpn gateway-ip** *gw-ip-address*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.
Step 2	**route-map** *route-map-name* **Example:** `switch(config)# route-map ROUTE_MAP_1`	Create a route map.

	Command or Action	Purpose
Step 3	**set evpn gateway-ip** *gw-ip-address* **Example:** `switch(config-route-map)# set evpn gateway-ip 209.165.200.227`	Set the gateway IP address.

Applying the Route Map at the Inbound or Outbound Level

Once you've configured the route map with match and set clauses based on your requirements, use this procedure to apply the route map at the inbound or outbound level.

SUMMARY STEPS

1. **configure terminal**
2. **router bgp** *as-num*
3. **neighbor** *address*
4. **address-family l2vpn evpn**
5. **route-map** *route-map* **{in | out}**

DETAILED STEPS

Procedure

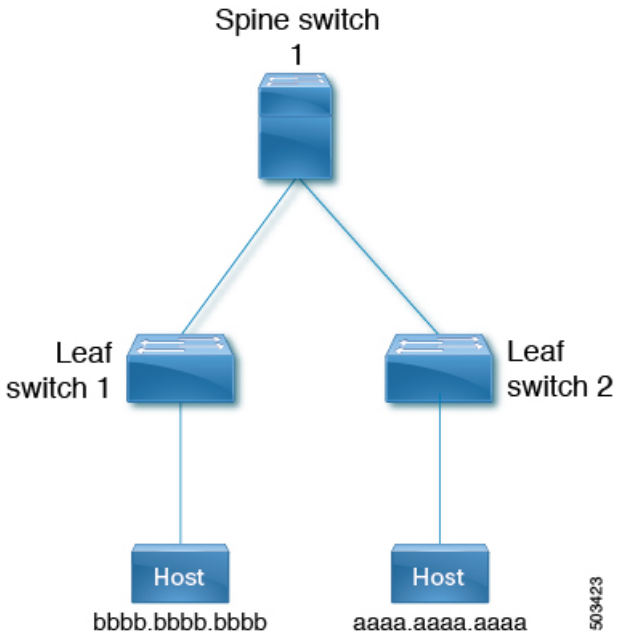
	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.
Step 2	**router bgp** *as-num* **Example:** `switch(config)# router bgp 100`	Enables a routing process. The range of *as-num* is from 1 to 65535.
Step 3	**neighbor** *address* **Example:** `switch(config-router)# neighbor 1.1.1.1`	Configure a BGP neighbor.
Step 4	**address-family l2vpn evpn** **Example:** `switch(config-router-neighbor)# address-family l2vpn evpn`	Configure the L2VPN address family.
Step 5	**route-map** *route-map* **{in | out}** **Example:** `switch(config-router-neighbor-af)# route-map ROUTE_MAP_1 in`	Apply the route map to the neighbor.

BGP EVPN Filtering Configuration Examples

This section provides example configurations for filtering EVPN routes.

Example 1

The following example shows how to filter EVPN type-2 routes and set the RMAC extended community as 52fc.c310.2e80.



1. The following output shows the routes in the EVPN table and a type-2 EVPN MAC route before the route map is applied.

```
leaf1(config)# show bgp l2vpn evpn
BGP routing table information for VRF default, address family L2VPN EVPN
BGP table version is 12, Local Router ID is 1.1.1.1
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-injected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup, 2 - best2

Network            Next Hop            Metric     LocPrf     Weight Path
Route Distinguisher: 1.1.1.1:32868    (L2VNI 101)
*>i[2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:[101.0.0.3]/272
33.33.33.33                         100          0 i

Route Distinguisher: 3.3.3.3:3
*>i[2]:[0]:[0]:[48]:[52fc.d83a.1b08]:[0]:[0.0.0.0]/216
33.33.33.33                         100          0 i
*>i[5]:[0]:[0]:[24]:[101.0.0.0]/224
3.3.3.3                  0          100          0 ?

Route Distinguisher: 3.3.3.3:32868
*>i[2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:[101.0.0.3]/272
33.33.33.33                         100          0 i

Route Distinguisher: 1.1.1.1:3    (L3VNI 100)
*>i[2]:[0]:[0]:[48]:[52fc.d83a.1b08]:[0]:[0.0.0.0]/216
33.33.33.33                         100          0 i
```

```
*>i[2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:[101.0.0.3]/272
33.33.33.33                       100          0 i
*>l[5]:[0]:[0]:[24]:[10.0.0.0]/224
1.1.1.1                  0        100      32768 ?
*>l[5]:[0]:[0]:[24]:[100.0.0.0]/224
1.1.1.1                  0        100      32768 ?
*>i[5]:[0]:[0]:[24]:[101.0.0.0]/224
3.3.3.3                  0        100          0 ?

leaf1(config)# show bgp l2vpn evpn aaaa.aaaa.aaaa
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 1.1.1.1:32868    (L2VNI 101)
BGP routing table entry for [2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:[101.0.0.3]/2
72, version 12
Paths: (1 available, best #1)
Flags: (0x000212) (high32 00000000) on xmit-list, is in l2rib/evpn, is not in HW

Advertised path-id 1
Path type: internal, path is valid, is best path, no labeled nexthop, in rib
Imported from 3.3.3.3:32868:[2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:
[101.0.0.3]/272
AS-Path: NONE, path sourced internal to AS
33.33.33.33 (metric 81) from 101.101.101.101 (101.101.101.101)
Origin IGP, MED not set, localpref 100, weight 0
Received label 101 100
Extcommunity: RT:100:100 RT:100:101 SOO:33.33.33.33:0 ENCAP:8
Router MAC:52fc.d83a.1b08
Originator: 3.3.3.3 Cluster list: 101.101.101.101

Path-id 1 not advertised to any peer

Route Distinguisher: 3.3.3.3:32868
BGP routing table entry for [2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:[101.0.0.3]/2
72, version 8
Paths: (1 available, best #1)
Flags: (0x000202) (high32 00000000) on xmit-list, is not in l2rib/evpn, is not in HW

Advertised path-id 1
Path type: internal, path is valid, is best path, no labeled nexthop
Imported to 3 destination(s)
Imported paths list: vni100 default default
AS-Path: NONE, path sourced internal to AS
33.33.33.33 (metric 81) from 101.101.101.101 (101.101.101.101)
Origin IGP, MED not set, localpref 100, weight 0
Received label 101 100
Extcommunity: RT:100:100 RT:100:101 SOO:33.33.33.33:0 ENCAP:8
Router MAC:52fc.d83a.1b08
Originator: 3.3.3.3 Cluster list: 101.101.101.101

Path-id 1 not advertised to any peer

Route Distinguisher: 1.1.1.1:3    (L3VNI 100)
BGP routing table entry for [2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:[101.0.0.3]/2
72, version 11
Paths: (1 available, best #1)
Flags: (0x000202) (high32 00000000) on xmit-list, is not in l2rib/evpn, is not in HW

Advertised path-id 1
Path type: internal, path is valid, is best path, no labeled nexthop
Imported from 3.3.3.3:32868:[2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:
[101.0.0.3]/272
AS-Path: NONE, path sourced internal to AS
33.33.33.33 (metric 81) from 101.101.101.101 (101.101.101.101)
Origin IGP, MED not set, localpref 100, weight 0
```

```
Received label 101 100
Extcommunity: RT:100:100 RT:100:101 SOO:33.33.33.33:0 ENCAP:8
Router MAC:52fc.d83a.1b08
Originator: 3.3.3.3 Cluster list: 101.101.101.101

Path-id 1 not advertised to any peer
```

2. The following example shows the route-map configuration.

```
leaf1(config)# show run rpm

!Command: show running-config rpm
!Running configuration last done at: Thu Sep  3 22:32:23 2020
!Time: Thu Sep  3 22:32:31 2020

version 9.3(5) Bios:version
route-map FILTER_EVPN_TYPE2 permit 10
    match evpn route-type 2
    set extcommunity evpn rmac 52fc.c310.2e80
route-map allow permit 10
```

3. The following example shows how to apply the route map to the EVPN peer as an inbound route map.

```
leaf1(config-router-neighbor-af)# show run bgp

!Command: show running-config bgp
!Running configuration last done at: Mon Aug  3 18:08:24 2020
!Time: Mon Aug  3 18:08:28 2020

version 9.3(5) Bios:version
feature bgp

router bgp 100
  event-history detail size large
  neighbor 101.101.101.101
    remote-as 100
    update-source loopback0
    address-family l2vpn evpn
      send-community extended
      route-map FILTER_EVPN_TYPE2 in
  vrf vni100
    address-family ipv4 unicast
      advertise l2vpn evpn
      redistribute direct route-map allow
```

4. The following output shows the routes in the EVPN table and a type-2 EVPN MAC route after the route map is applied.

```
leaf1(config)# show bgp l2vpn evpn
BGP routing table information for VRF default, address family L2VPN EVPN
BGP table version is 19, Local Router ID is 1.1.1.1
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-injected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup, 2 - best2

Network            Next Hop            Metric     LocPrf     Weight Path
Route Distinguisher: 1.1.1.1:32868    (L2VNI 101)
*>i[2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:[101.0.0.3]/272
                      33.33.33.33                       100          0 i

Route Distinguisher: 3.3.3.3:3
*>i[2]:[0]:[0]:[48]:[52fc.d83a.1b08]:[0]:[0.0.0.0]/216
```

```
                      33.33.33.33                       100          0 i

Route Distinguisher: 3.3.3.3:32868
*>i[2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:[101.0.0.3]/272
                      33.33.33.33                       100          0 i

Route Distinguisher: 1.1.1.1:3    (L3VNI 100)
*>i[2]:[0]:[0]:[48]:[52fc.d83a.1b08]:[0]:[0.0.0.0]/216
                      33.33.33.33                       100          0 i
*>i[2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:[101.0.0.3]/272
                      33.33.33.33                       100          0 i
*>l[5]:[0]:[0]:[24]:[10.0.0.0]/224
                      1.1.1.1                 0         100      32768 ?
*>l[5]:[0]:[0]:[24]:[100.0.0.0]/224
                      1.1.1.1                 0         100      32768 ?

leaf1(config)# show bgp l2vpn evpn aaaa.aaaa.aaaa
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 1.1.1.1:32868    (L2VNI 101)
BGP routing table entry for [2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:[101.0.0.3]/2
72, version 19
Paths: (1 available, best #1)
Flags: (0x000212) (high32 00000000) on xmit-list, is in l2rib/evpn, is not in HW

Advertised path-id 1
Path type: internal, path is valid, is best path, no labeled nexthop, in rib
Imported from 3.3.3.3:32868:[2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:
[101.0.0.3]/272
AS-Path: NONE, path sourced internal to AS
33.33.33.33 (metric 81) from 101.101.101.101 (101.101.101.101)
Origin IGP, MED not set, localpref 100, weight 0
Received label 101 100
Extcommunity: RT:100:100 RT:100:101 SOO:33.33.33.33:0 ENCAP:8
Router MAC:52fc.c310.2e80
Originator: 3.3.3.3 Cluster list: 101.101.101.101
Path-id 1 not advertised to any peer

Route Distinguisher: 3.3.3.3:32868
BGP routing table entry for [2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:[101.0.0.3]/2
72, version 15
Paths: (1 available, best #1)
Flags: (0x000202) (high32 00000000) on xmit-list, is not in l2rib/evpn, is not in HW

Advertised path-id 1
Path type: internal, path is valid, is best path, no labeled nexthop
Imported to 3 destination(s)
Imported paths list: vni100 default default
AS-Path: NONE, path sourced internal to AS
33.33.33.33 (metric 81) from 101.101.101.101 (101.101.101.101)
Origin IGP, MED not set, localpref 100, weight 0
Received label 101 100
Extcommunity: RT:100:100 RT:100:101 SOO:33.33.33.33:0 ENCAP:8
Router MAC:52fc.c310.2e80
Originator: 3.3.3.3 Cluster list: 101.101.101.101

Path-id 1 not advertised to any peer

Route Distinguisher: 1.1.1.1:3    (L3VNI 100)
BGP routing table entry for [2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:[101.0.0.3]/2
72, version 18
Paths: (1 available, best #1)
Flags: (0x000202) (high32 00000000) on xmit-list, is not in l2rib/evpn, is not in HW

Advertised path-id 1
```

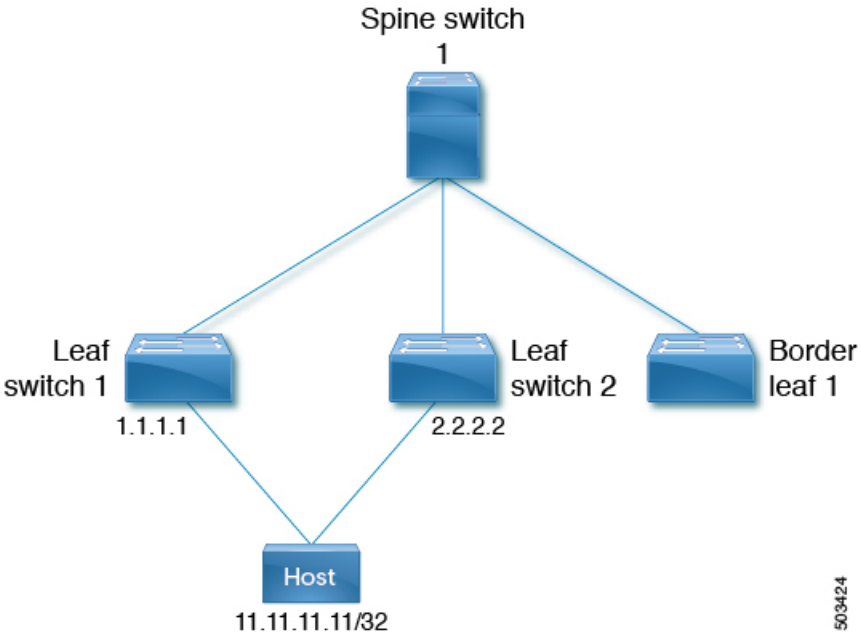
```
Path type: internal, path is valid, is best path, no labeled nexthop
Imported from 3.3.3.3:32868:[2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:
[101.0.0.3]/272
AS-Path: NONE, path sourced internal to AS
33.33.33.33 (metric 81) from 101.101.101.101 (101.101.101.101)
Origin IGP, MED not set, localpref 100, weight 0
Received label 101 100
Extcommunity: RT:100:100 RT:100:101 SOO:33.33.33.33:0 ENCAP:8
Router MAC:52fc.c310.2e80
Originator: 3.3.3.3 Cluster list: 101.101.101.101

Path-id 1 not advertised to any peer
```

In a similar manner, you can use the other EVPN-specific match and set clauses with existing route-map options to filter EVPN routes as required.

Example 2

The following example shows how EVPN route filtering can be used to redirect traffic to a different VTEP than the one from which the EVPN route was learned. It involves setting the next-hop IP address and the RMAC of the route to the one corresponding to the other VTEP.



This example demonstrates the following:

- Host 1 belongs to VRF evpn-tenant-0002 and VLAN 3002, and is connected to Leaf 1 and Leaf 2.
- Reachability to Host1 is advertised by Leaf 1 and Leaf 2 to BL1.

At BL1, both routes to 11.11.11.11/32 are received as follows:

- One from 1.1.1.1, which is Leaf 1
- One from 2.2.2.2, which is Leaf 2

1. Initially the best path to reach 11.11.11.11 is through 1.1.1.1:

```
bl1(config)# show bgp l2 e 11.11.11.11
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 1.1.1.1:3
BGP routing table entry for [5]:[0]:[0]:[32]:[11.11.11.11]/224, version 15
Paths: (1 available, best #1)
Flags: (0x000002) (high32 00000000) on xmit-list, is not in l2rib/evpn, is not in HW

Advertised path-id 1
Path type: internal, path is valid, is best path, no labeled nexthop
    Imported to 2 destination(s)
    Imported paths list: evpn-tenant-0002 default
Gateway IP: 0.0.0.0
AS-Path: 150 , path sourced external to AS
  1.1.1.1 (metric 81) from 101.101.101.101 (101.101.101.101)
    Origin incomplete, MED 0, localpref 100, weight 0
    Received label 3003002
    Extcommunity: RT:1:3003002 ENCAP:8 Router MAC:5254.0074.caf5
    Originator: 1.1.1.1 Cluster list: 101.101.101.101

Path-id 1 not advertised to any peer

Route Distinguisher: 2.2.2.2:4
BGP routing table entry for [5]:[0]:[0]:[32]:[11.11.11.11]/224, version 79
Paths: (1 available, best #1)
Flags: (0x000002) (high32 00000000) on xmit-list, is not in l2rib/evpn, is not in HW

Advertised path-id 1
Path type: internal, path is valid, is best path, no labeled nexthop
    Imported to 2 destination(s)
    Imported paths list: evpn-tenant-0002 default
Gateway IP: 0.0.0.0
AS-Path: 150 , path sourced external to AS
  2.2.2.2 (metric 81) from 101.101.101.101 (101.101.101.101)
    Origin incomplete, MED 0, localpref 100, weight 0
    Received label 3003002
    Extcommunity: RT:1:3003002 ENCAP:8 Router MAC:5254.0090.433e
    Originator: 2.2.2.2 Cluster list: 101.101.101.101

Path-id 1 not advertised to any peer

Route Distinguisher: 3.3.3.3:3    (L3VNI 3003002)
BGP routing table entry for [5]:[0]:[0]:[32]:[11.11.11.11]/224, version 80
Paths: (2 available, best #2)Flags: (0x000002) (high32 00000000) on xmit-list, is not
in l2rib/evpn, is not in HW

Path type: internal, path is valid, not best reason: Router Id, no labeled nexthop
    Imported from 2.2.2.2:4:[5]:[0]:[0]:[32]:[11.11.11.11]/224
Gateway IP: 0.0.0.0
AS-Path: 150 , path sourced external to AS
  2.2.2.2 (metric 81) from 101.101.101.101 (101.101.101.101)
    Origin incomplete, MED 0, localpref 100, weight 0
    Received label 3003002
    Extcommunity: RT:1:3003002 ENCAP:8 Router MAC:5254.0090.433e
    Originator: 2.2.2.2 Cluster list: 101.101.101.101

Advertised path-id 1
Path type: internal, path is valid, is best path, no labeled nexthop
    Imported from 1.1.1.1:3:[5]:[0]:[0]:[32]:[11.11.11.11]/224
Gateway IP: 0.0.0.0
AS-Path: 150 , path sourced external to AS
  1.1.1.1 (metric 81) from 101.101.101.101 (101.101.101.101)
    Origin incomplete, MED 0, localpref 100, weight 0
    Received label 3003002
    Extcommunity: RT:1:3003002 ENCAP:8 Router MAC:5254.0074.caf5
```

```
    Originator: 1.1.1.1 Cluster list: 101.101.101.101

Path-id 1 not advertised to any peer

Route Distinguisher: 3.3.3.3:4    (L3VNI 3003003)
BGP routing table entry for [5]:[0]:[0]:[32]:[11.11.11.11]/224, version 24
Paths: (1 available, best #1)
Flags: (0x000002) (high32 00000000) on xmit-list, is not in l2rib/evpn

Advertised path-id 1
Path type: local, path is valid, is best path, no labeled nexthop
Gateway IP: 0.0.0.0
AS-Path: 150 , path sourced external to AS
  3.3.3.3 (metric 0) from 0.0.0.0 (3.3.3.3)
    Origin incomplete, MED 0, localpref 100, weight 0
    Received label 3003003
    Extcommunity: RT:1:3003003 ENCAP:8 Router MAC:5254.006a.435b
    Originator: 1.1.1.1 Cluster list: 101.101.101.101

Path-id 1 advertised to peers:
101.101.101.101

bl1(config)# show ip route 11.11.11.11
IP Route Table for VRF "default"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

11.11.11.11/32, ubest/mbest: 1/0
*via 1.1.1.1, [200/0], 00:02:51, bgp-1, internal, tag 150 (evpn) segid: 3003
002 tunnelid: 0x1010101 encap: VXLAN
```

2. To redirect traffic to the other VTEP leaf-2, you can set the next hop and RMAC on the 11.11.11.11/32 route with a route-map configuration.

```
bl1(config-route-map)# show run rpm

Command: show running-config rpm
!Running configuration last done at: Wed Mar 27 00:12:14 2019
!Time: Wed Mar 27 00:12:17 2019

version 9.2(3) Bios:version
ip prefix-list PFX_LIST1_1 seq 5 permit 11.11.11.11/32
route-map TEST_SET_IP_NEXTHOP permit 10
    match ip address prefix-list PFX_LIST1_1
    set ip next-hop 2.2.2.2
    set extcommunity evpn rmac 5254.0090.433e
```

3. After applying the route map at the inbound level at BL1, the following are the route outputs for route 11.11.11.11/32.

```
bl1(config-router-neighbor-af)# show bgp l2 e 11.11.11.11
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 1.1.1.1:3
BGP routing table entry for [5]:[0]:[0]:[32]:[11.11.11.11]/224, version 81
Paths: (1 available, best #1)
Flags: (0x000002) (high32 00000000) on xmit-list, is not in l2rib/evpn, is not in HW

Advertised path-id 1
Path type: internal, path is valid, is best path, no labeled nexthop
    Imported to 2 destination(s)
```

```
    Imported paths list: evpn-tenant-0002 default
Gateway IP: 0.0.0.0
AS-Path: 150 , path sourced external to AS
  2.2.2.2 (metric 81) from 101.101.101.101 (101.101.101.101)
    Origin incomplete, MED 0, localpref 100, weight 0
    Received label 3003002
    Extcommunity: RT:1:3003002 ENCAP:8 Router MAC:5254.0090.433e
    Originator: 1.1.1.1 Cluster list: 101.101.101.101

Path-id 1 not advertised to any peer

Route Distinguisher: 2.2.2.2:4
BGP routing table entry for [5]:[0]:[0]:[32]:[11.11.11.11]/224, version 79
Paths: (1 available, best #1)
Flags: (0x000002) (high32 00000000) on xmit-list, is not in l2rib/evpn, is not in HW

Advertised path-id 1
Path type: internal, path is valid, is best path, no labeled nexthop
    Imported to 2 destination(s)
    Imported paths list: evpn-tenant-0002 default
Gateway IP: 0.0.0.0
AS-Path: 150 , path sourced external to AS
  2.2.2.2 (metric 81) from 101.101.101.101 (101.101.101.101)
    Origin incomplete, MED 0, localpref 100, weight 0
    Received label 3003002
    Extcommunity: RT:1:3003002 ENCAP:8 Router MAC:5254.0090.433e
    Originator: 2.2.2.2 Cluster list: 101.101.101.101

Path-id 1 not advertised to any peer

Route Distinguisher: 3.3.3.3:3    (L3VNI 3003002)
BGP routing table entry for [5]:[0]:[0]:[32]:[11.11.11.11]/224, version 82
Paths: (2 available, best #2)
Flags: (0x000002) (high32 00000000) on xmit-list, is not in l2rib/evpn, is not in HW

Path type: internal, path is valid, not best reason: Router Id, no labeled nexthop
    Imported from 2.2.2.2:4:[5]:[0]:[0]:[32]:[11.11.11.11]/224
Gateway IP: 0.0.0.0
AS-Path: 150 , path sourced external to AS
  2.2.2.2 (metric 81) from 101.101.101.101 (101.101.101.101)
    Origin incomplete, MED 0, localpref 100, weight 0
    Received label 3003002
    Extcommunity: RT:1:3003002 ENCAP:8 Router MAC:5254.0090.433e
    Originator: 2.2.2.2 Cluster list: 101.101.101.101

Advertised path-id 1
Path type: internal, path is valid, is best path, no labeled nexthop
    Imported from 1.1.1.1:3:[5]:[0]:[0]:[32]:[11.11.11.11]/224
Gateway IP: 0.0.0.0
AS-Path: 150 , path sourced external to AS
  2.2.2.2 (metric 81) from 101.101.101.101 (101.101.101.101)
    Origin incomplete, MED 0, localpref 100, weight 0
    Received label 3003002
    Extcommunity: RT:1:3003002 ENCAP:8 Router MAC:5254.0090.433e
    Originator: 1.1.1.1 Cluster list: 101.101.101.101

Path-id 1 not advertised to any peer

Route Distinguisher: 3.3.3.3:4    (L3VNI 3003003)
BGP routing table entry for [5]:[0]:[0]:[32]:[11.11.11.11]/224, version 24
Paths: (1 available, best #1)
Flags: (0x000002) (high32 00000000) on xmit-list, is not in l2rib/evpn

Advertised path-id 1
```

```
Path type: local, path is valid, is best path, no labeled nexthop
Gateway IP: 0.0.0.0
AS-Path: 150 , path sourced external to AS
  3.3.3.3 (metric 0) from 0.0.0.0 (3.3.3.3)
    Origin incomplete, MED 0, localpref 100, weight 0
    Received label 3003003
    Extcommunity: RT:1:3003003 ENCAP:8 Router MAC:5254.006a.435b
    Originator: 1.1.1.1 Cluster list: 101.101.101.101

Path-id 1 advertised to peers:
101.101.101.101


bl1(config-router-neighbor-af)# show ip route 11.11.11.11
IP Route Table for VRF "default"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

11.11.11.11/32, ubest/mbest: 1/0
*via 2.2.2.2, [200/0], 00:02:37, bgp-1, internal, tag 150 (evpn) segid: 3003
002 tunnelid: 0x2020202 encap: VXLAN
```

After the next hop and RMAC value are set using the route map, the traffic that was earlier directed through 1.1.1.1 is now directed through 2.2.2.2.

Configuring a Table Map

Perform these tasks to configure and apply a table map:

- Configuring a MAC List and a Route Map that Matches the MAC List, on page 518
- Applying the Table Map, on page 519

Configuring a MAC List and a Route Map that Matches the MAC List

SUMMARY STEPS

1. **configure terminal**
2. **mac-list** *list-name* **[seq** *seq-number***] {deny | permit}** *mac-address* **[mac-mask]**
3. **route-map** *route-map-name*
4. **match mac-list** *mac-list-name*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.

	Command or Action	Purpose
Step 2	**mac-list** *list-name* **[seq** *seq-number***] {deny | permit}** *mac-address* **[mac-mask]** **Example:** `switch(config)# mac-list MAC_LIST_1 permit E:E:E`	Build a MAC list.
Step 3	**route-map** *route-map-name* **Example:** `switch(config)# route-map ROUTE_MAP_1`	Create a route map.
Step 4	**match mac-list** *mac-list-name* **Example:** `switch(config-route-map)# match mac-list MAC_LIST_1`	Match entries of MAC lists. The maximum length is 63 characters.

Applying the Table Map

SUMMARY STEPS

1. **configure terminal**
2. **evpn**
3. **vni** *vni-id* **l2**
4. **table-map** *route-map-name* **[filter]**

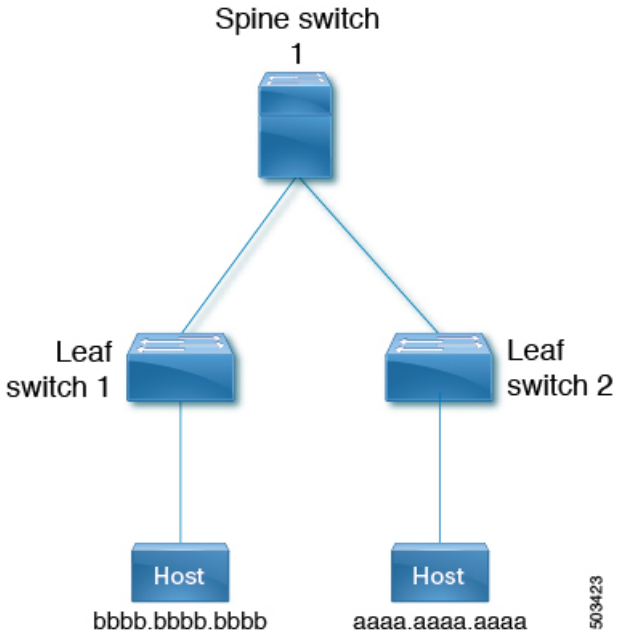
DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.
Step 2	**evpn** **Example:** `switch(config)# evpn`	Enter EVPN configuration mode.
Step 3	**vni** *vni-id* **l2** **Example:** `switch(config-evpn)# vni 101 l2`	Configure the Ethernet VPN ID. The range of *vni-id* is from 1 to 16777214.
Step 4	**table-map** *route-map-name* **[filter]** **Example:** `switch(config-evpn-evi)# table-map ROUTE_MAP_1 filter`	Apply table maps at the EVPN VNI configuration level. If the **filter** option is specified, any route that gets denied by the route-map validation isn't downloaded into the L2RIB.

Table Map Configuration Example

The following table-map configuration example shows how to filter MAC route aaaa.aaaa.aaaa from being downloaded into the L2RIB.



1. The following example shows the output for routes in the EVPN table and MAC routes in the L2RIB before the route map is applied.

```
leaf1(config)# show bgp l2vpn evpn
BGP routing table information for VRF default, address family L2VPN EVPN
BGP table version is 25, Local Router ID is 1.1.1.1
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-injected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup, 2 - best2

Network            Next Hop            Metric     LocPrf     Weight Path
Route Distinguisher: 1.1.1.1:32868    (L2VNI 101)
*>i[2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:[101.0.0.3]/272
33.33.33.33                          100          0 i

Route Distinguisher: 3.3.3.3:3
*>i[2]:[0]:[0]:[48]:[52fc.d83a.1b08]:[0]:[0.0.0.0]/216
33.33.33.33                          100          0 i

Route Distinguisher: 3.3.3.3:32868
*>i[2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:[101.0.0.3]/272
33.33.33.33                          100          0 i

Route Distinguisher: 1.1.1.1:3    (L3VNI 100)
*>i[2]:[0]:[0]:[48]:[52fc.d83a.1b08]:[0]:[0.0.0.0]/216
33.33.33.33                          100          0 i
*>i[2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:[101.0.0.3]/272
33.33.33.33                          100          0 i
*>l[5]:[0]:[0]:[24]:[10.0.0.0]/224
1.1.1.1                   0          100      32768 ?
*>l[5]:[0]:[0]:[24]:[100.0.0.0]/224
1.1.1.1                   0          100      32768 ?
```

```
leaf1(config)# show l2route evpn mac all

Flags -(Rmac):Router MAC (Stt):Static (L):Local (R):Remote (V):vPC link
(Dup):Duplicate (Spl):Split (Rcv):Recv (AD):Auto-Delete (D):Del Pending
(S):Stale (C):Clear, (Ps):Peer Sync (O):Re-Originated (Nho):NH-Override
(Pf):Permanently-Frozen, (Orp): Orphan

Topology  Mac Address     Prod  Flags  Seq No  Next-Hops
--------- -----------     ----  -----  ------- ----------
100       52fc.d83a.1b08  VXLAN  Rmac  0       33.33.33.33
101       aaaa.aaaa.aaaa  BGP    Spl   0       33.33.33.33 (Label: 101)

leaf1(config-evpn-evi)# show mac address-table vlan 101
Legend: * - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC
age - seconds since last seen,+ - primary entry using vPC Peer-Link,
(T) - True, (F) - False, C - ControlPlane MAC, ~ - vsanVLAN     MAC Address
Type      age     Secure NTFY Ports
---------+-----------------+--------+---------+------+----+------------------
C  101     aaaa.aaaa.aaaa   dynamic  0         F      F    nve1(33.33.33.33)
G  101     521d.7cef.1b08   static   -         F      F    sup-eth1(R)
```

2. The following example shows how to configure the route map to filter MAC route aaaa.aaaa.aaaa.

```
leaf1(config)# show run rpm

!Command: show running-config rpm
!Running configuration last done at: Thu Sep  3 21:47:48 2020
!Time: Thu Sep  3 22:27:57 2020

version 9.4(1) Bios:version
mac-list FILTER_MAC_AAA seq 5 deny aaaa.aaaa.aaaa ffff.ffff.ffff
route-map TABLE_MAP_FILTER permit 10
  match mac-list FILTER_MAC_AAA
```

3. The following example shows how to apply the route map at the BGP EVPN level.

```
leaf1(config-evpn-evi)# show run bgp | section evpn
evpn
    vni 101 l2
      table-map TABLE_MAP_FILTER filter
      rd auto
      route-target import auto
      route-target export auto
      route-target both auto evpn
```

4. The following example shows the output for routes in the EVPN table and MAC routes in the L2RIB after the table map is configured.

```
leaf1(config-evpn-evi)# show bgp l2vpn evpn
BGP routing table information for VRF default, address family L2VPN EVPN
BGP table version is 26, Local Router ID is 1.1.1.1
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-injected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup, 2 - best2
Network            Next Hop            Metric     LocPrf     Weight Path
Route Distinguisher: 1.1.1.1:32868    (L2VNI 101)
*>i[2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:[101.0.0.3]/272
33.33.33.33                       100          0 i

Route Distinguisher: 3.3.3.3:3
*>i[2]:[0]:[0]:[48]:[52fc.d83a.1b08]:[0]:[0.0.0.0]/216
```

```
33.33.33.33                       100          0 i

Route Distinguisher: 3.3.3.3:32868
*>i[2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:[101.0.0.3]/272
33.33.33.33                       100          0 i

Route Distinguisher: 1.1.1.1:3    (L3VNI 100)
*>i[2]:[0]:[0]:[48]:[52fc.d83a.1b08]:[0]:[0.0.0.0]/216
33.33.33.33                       100          0 i
*>i[2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:[101.0.0.3]/272
33.33.33.33                       100          0 i
*>l[5]:[0]:[0]:[24]:[10.0.0.0]/224
1.1.1.1                  0        100      32768 ?
*>l[5]:[0]:[0]:[24]:[100.0.0.0]/224
1.1.1.1                  0        100      32768 ?

leaf1(config-evpn-evi)# show l2route evpn mac all

Flags -(Rmac):Router MAC (Stt):Static (L):Local (R):Remote (V):vPC link
(Dup):Duplicate (Spl):Split (Rcv):Recv (AD):Auto-Delete (D):Del Pending
(S):Stale (C):Clear, (Ps):Peer Sync (O):Re-Originated (Nho):NH-Override
(Pf):Permanently-Frozen, (Orp): Orphan

Topology    Mac Address    Prod   Flags  Seq No  Next-Hops
----------- -------------- ------ ------ ------- ------------
100         52fc.d83a.1b08 VXLAN  Rmac   0       33.33.33.33

leaf1(config-evpn-evi)# show mac address-table vlan 101
Legend:
* - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC
age - seconds since last seen,+ - primary entry using vPC Peer-Link,
(T) - True, (F) - False, C - ControlPlane MAC, ~ - vsan
VLAN     MAC Address      Type      age     Secure NTFY Ports
---------+-----------------+--------+---------+------+----+------------------
G  101     521d.7cef.1b08   static   -         F      F    sup-eth1(R)
```

Verifying BGP EVPN Filtering

To display the status of the BGP EVPN Filtering configuration, enter the following command:

Table 12: Display BGP EVPN Filtering

Command	Purpose
show mac-list	Displays MAC Lists.
show route-map *name*	Displays information about a route map.
show running-config bgp	Displays the BGP configuration.
show running-config rpm	Displays all Route Policy Manager (RPM) information.
show bgp l2vpn evpn	Displays routes in BRIB.

Example of the **show mac-list** command:

```
switch(config)# show mac-list
mac-list list1: 5 entries
   seq 5 deny 0000.836d.f8b7 ffff.ffff.ffff
   seq 6 deny 0000.836d.f8b5 ffff.ffff.ffff
   seq 7 permit 0000.0422.6811 ffff.ffff.ffff
   seq 8 deny 0000.836d.f8b1 ffff.ffff.ffff
   seq 10 permit 0000.0000.0000 0000.0000.0000
mac-list list2: 3 entries
   seq 5 deny 0000.836e.f8b6 ffff.ffff.ffff
   seq 8 deny 0000.0421.6818 ffff.ffff.ffff
   seq 10 permit 0000.0000.0000 0000.0000.0000
mac-list list3: 2 entries
   seq 5 deny 0000.836d.f8b6 ffff.ffff.ffff
   seq 10 permit 0000.836d.f8b7 ffff.ffff.ffff
```

Example of the **show route-map** command:

```
switch# show route-map pol10
route-map pol10, permit, sequence 10
  Match clauses:
    mac-list: list2
  Set clauses:
    ip next-hop 6.6.6.1 3.3.3.10
    ipv6 next-hop 303:304::1
```

Example of the **show running-config bgp** command:

```
switch# show running-config bgp | beg "5000"
vni 5000 l2
table-map pol1 filter
rd auto
route-target import auto
route-target export auto
vni 5001 l2
rd auto
route-target import auto
route-target export auto
```

Example of the **show running-config rpm** command:

```
switch# show running-config rpm
!Running configuration last done at: Thu May 23 13:58:31 2019
!Time: Thu May 23 13:58:47 2019

version 9.3(1) Bios:version 07.65
feature pbr

mac-list list1 seq 5 permit 0001.0001.0001 ffff.ffff.ffff
mac-list mclist seq 5 permit 0001.0001.0001 ffff.ffff.ffff
route-map test permit 10
match evpn route-type 5
set evpn gateway-ip 1.1.1.2
```

Example of the **show bgp l2vpn evpn aaaa.aaaa.aaaa** command to view detailed information about EVPN route aaaa.aaaa.aaaa:

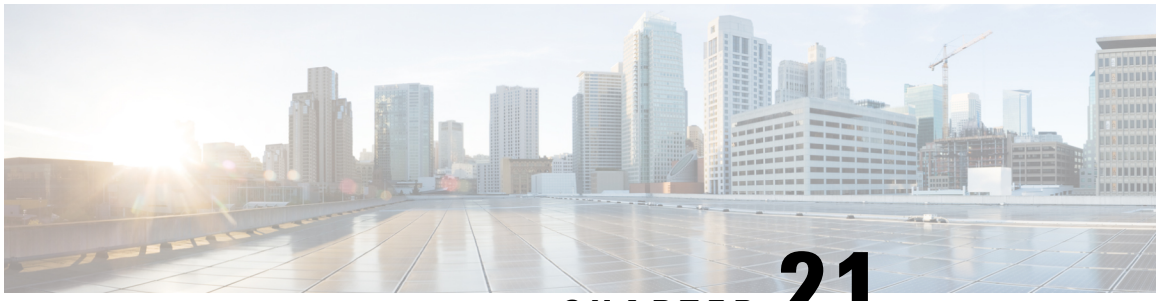
```
switch(config-evpn-evi)# show bgp l2 e aaaa.aaaa.aaaa

BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 1.1.1.1:32868 (L2VNI 101)
BGP routing table entry for [2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:[101.0.0.3]/2
72, version 11
```

```
Paths: (1 available, best #1)
Flags: (0x000202) (high32 00000000) on xmit-list, is not in l2rib/evpn, table-ma
p filtered, is not in HW

Advertised path-id 1
Path type: internal, path is valid, is best path, remote nh not installed, no
labeled nexthop
Imported from 3.3.3.3:32868:[2]:[0]:[0]:[48]:[aaaa.aaaa.aaaa]:[32]:
[101.0.0.3]/272
AS-Path: NONE, path sourced internal to AS
33.33.33.33 (metric 81) from 101.101.101.101 (101.101.101.101)
Origin IGP, MED not set, localpref 100, weight 0
Received label 101 100
Extcommunity: RT:100:100 RT:100:101 SOO:33.33.33.33:0 ENCAP:8
Router MAC:5254.009b.4275
Originator: 3.3.3.3 Cluster list: 101.101.101.101

Path-id 1 not advertised to any peer
```



CHAPTER 21

Configuring VXLAN BGP-EVPN Null Route

This chapter contains these sections:

- About EVPN Null Route, on page 525
- Guidelines and Limitations for VXLAN BGP-EVPN Null Route, on page 526
- Configuring Static MAC, on page 527
- Configuring ARP/ND, on page 527
- Configuring Prefix-Null Route on Local VTEP, on page 529
- Configuring RPM Route-Map on Remote VTEP, on page 531
- Configuration Example for Null Route, on page 533
- Verifying EVPN Null Route Configuration, on page 534

About EVPN Null Route

A Distributed Denial of Service (DDoS) attack on a host in an EVPN Fabric consumes the network bandwidth resources and in turn impacts legitimate traffic to other hosts.

The DDoS attack can be from any of the following setups:

- Host connected to a leaf switch within the local site
- Host connected to a leaf switch in a remote site
- External networks such as WAN

The DDoS attack can be intra-subnets (MAC based) or inter-subnets (Host-based – IPv4/IPv6)

Null route filtering has been traditionally used in mitigating DDoS attacks especially in service provider networks.

A null route is a network route (routing table entry) that goes nowhere. Matching packets are dropped (ignored or redirected) rather than forwarded, acting as a kind of limited firewall. The act of using null routes is often called null route filtering.

NX-OS already has mechanisms to configure the null/drop route for IPv4/IPv6/MAC. The null route will be required to be configured on all VTEPs in the fabric.

For IPv4/IPv6 based attacks, use the following commands to configure an IPv4/IPv6 static route with null interface:

- **ip route x.x.x.x/y Null0**

- **ipv6 route X:X:X::X/Y Null0**

For MAC-based attacks, use the following command to configure MAC address with drop adjacency to drop the packets:

- **mac address-table static xxxx.yyyy.zzzz vlan <VLAN-ID> drop**

In a fabric with large number of VTEPs and across multiple sites, manually configuring and administering the drop route on all VTEPs is difficult task in the absence of Nexus Dashboard Fabric Controller (NDFC) or other Orchestrator.

The EVPN null routing feature is used when you do not have a way to configure and inject a null route from a central location such as with NDFC or other Orchestrators.

EVPN null routing feature enables a VTEP within the network to send Type-2 and Type-5 routes tagged with a specific community.

Other VTEPs (Borders and Leafs) in the single-site and multi-site can install an entry in MAC or IP (IPv4/IPv6) table such that any traffic destined to MAC or IP respectively is dropped at the Edge or leaf switch which prevents the usage of bandwidth within the site and across the site.

The programmed null route entry can be a Host IP (/32 or /128), a Prefix (VLSM) or a MAC.

Guidelines and Limitations for VXLAN BGP-EVPN Null Route

- A null route (static) MAC configuration must have matching static ARP/ND configuration which means you must not have a dynamic ARP/ND with MACs configured as null route MACs.
- If you use only L2-services (and has no configuration that can lead to dynamic ARP/ND learning) then a “mac drop” configuration alone is allowed. In all other cases, we require static ARP/ND configuration also along with the “mac drop” configuration.
- In case of vPC, the null route (MAC, mac-ip, prefix) must be configured on both vPC boxes (VMCT and PMCT). The behavior is undefined if this is not configured on both boxes. The same holds good during unconfiguring the null route. The vPC consistency checker for this feature is not supported.
- The route-map must be applied on the remote VTEPs. This ingress Route-Map is important for Type-5 routes.
- No feature interaction with multicast traffic.
- When remote static is seen on a VTEP and if you want to configure the same MAC as a local static (static MAC with a valid interface or MAC set to drop/null route MAC), a syslog will be generated to warn about the duplicate configuration in the fabric that must be corrected. However, the configuration will not be rejected. The local static configuration holds precedence over a remote static configuration on that VTEP.
- If local static MAC with a valid interface is configured on a VTEP, and you want to convert this static MAC to a null route MAC on the same VTEP, the null route MAC takes effect.
- Though the remote dynamic MAC route permits any remote MAC route derived from MAC-IP route split to overwrite its entry, and propagate to MAC manager the remote static MAC route will no longer honor these derived MACs to overwrite its entry. As a result, the MAC entry remains unchanged until the remote static MAC is deleted.

- The null route MAC is another form of static MAC configuration only.

Configuring Static MAC

Before you begin

You can configure static drop MAC addresses. These static MAC addresses override dynamically learned MAC addresses on any interfaces.

SUMMARY STEPS

1. **configure terminal**
2. **mac address-table static** *mac-address* **vlan** *vlan-id* {[**drop**| **interface**{*type slot/port*} | **port-channel** *number*]}
3. **exit**

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# configure terminal`
	`switch(config)#`
		Enters global configuration mode.

Step 2	**mac address-table static** *mac-address* **vlan** *vlan-id* {[**drop**| **interface**{*type slot/port*}	**port-channel** *number*]}
		Example:
		`switch(config)# mac address-table static 3001.3010.99aa vlan 3001 drop`
		`switch(config)#`
			Specifies a static MAC address to add to the Layer 2 MAC address table.

Step 3	**exit**
	Example:
	`switch# exit`
	`switch#`
		Exits the configuration mode.

Configuring ARP/ND

You can configure ARP/ND host on IPv4/IPv6 route for the corresponding SVI.

Before you begin

Ensure to configure static MAC-IP configuration on the switch where MAC is configured as drop entry. This will avoid MAC-IP mobility and ensures both DROP MAC and MAC-IP are originated from same VTEP.

SUMMARY STEPS

1. **configure terminal**
2. **interface** *vlan-number*
3. **vrf member** *vrf-name*
4. **no ip redirects**
5. **ip address** *address*
6. **ipv6 address** *address*
7. **ipv6 neighbor address** *ipv6address mac_addr*
8. **no ipv6 redirects**
9. **ip arp address** *ipaddr mac_addr*
10. **fabric forwarding mode anycast-gateway**

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# configure terminal`
	`switch(config)#`
		Enters global configuration mode.

Step 2	**interface** *vlan-number*
	Example:
	`switch(config)# interface Vlan 3001`
	`switch(config-if)#`
		Specifies the VLAN interface.

Step 3	**vrf member** *vrf-name*
	Example:
	`switch(config-if)# vrf member cgw_3001_3050`
	`switch(config-if)#`
		Assigns the VLAN interface to the tenant VRF.

Step 4	**no ip redirects**
	Example:
	`switch(config-if)# no ip redirects`
	`switch(config-if)#`
		Disables the IPv4 redirects.

Step 5	**ip address** *address*
	Example:
	`switch(config-if)# ip address 30.1.0.1/16`
	`switch(config-if)#`
		Specifies the IP address.

	Command or Action	Purpose

Step 6	**ipv6 address** *address*
	Example:
	`switch(config-if)# ipv6 address 2001:3001::1/64`
	`switch(config-if)#`
		Specifies the IPv6 address.

Step 7	**ipv6 neighbor address** *ipv6address mac_addr*
	Example:
	`switch(config-if)# ipv6 neighbor 2001:3001::99 3001.3010.99aa`
	`switch(config-if)#`
		Configures static IPv6 neighbor.

Step 8	**no ipv6 redirects**
	Example:
	`switch(config-if)# no ipv6 redirects`
	`switch(config-if)#`
		Disables the IPv6 redirects.

Step 9	**ip arp address** *ipaddr mac_addr*
	Example:
	`switch(config-if)# ip arp 30.1.0.99 3001.3010.99aa`
	`switch(config-if)#`
		Associates an IP address with a MAC address as a static entry.

Step 10	**fabric forwarding mode anycast-gateway**
	Example:
	`switch# fabric forwarding mode anycast-gateway`
	`switch#`
		Associates SVI with anycast gateway under VLAN configuration mode.

Configuring Prefix-Null Route on Local VTEP

On a local VTEP where the Null route is configured, configure route-map to set blackhole community on static route and redistribute into BGP.

SUMMARY STEPS

1. **configure terminal**
2. **vrf context** *vrf-name*
3. **ip route** {<*ip*>/*mask*} **Null0 tag** <*tag-number*> or **ip route** {<*ipv6*>/*mask*} **Null0 tag** <*tag-number*>
4. **route-map** *map-name* [**permit** | **deny**] [*seq*]
5. **match tag** <*tag-number*>
6. **set weight** *value*
7. **set community blackhole**
8. **router bgp** *as-number*
9. **vrf** *vrf-name*
10. **address-family ipv4/ipv6 unicast**
11. **redistribute static route-map** *route-map name*

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# configure terminal`
	`switch(config)#`
		Enters global configuration mode.

Step 2	**vrf context** *vrf-name*
	Example:
	`switch(config)# vrf context tenant-0001`
	`switch(config-vrf)#`
		Configures the tenant VRF.

Step 3	**ip route** {<*ip*>/*mask*} **Null0 tag** <*tag-number*> or **ip route** {<*ipv6*>/*mask*} **Null0 tag** <*tag-number*>
	Example:
	For IPv4
	`switch(config-vrf)# ip route 50.1.0.0/24 Null0 tag 6666`
	`switch(config-vrf)#`
	For IPv6
	`switch(config-vrf)# ipv6 route 50::1:0/120 Null0 tag 6666`
	`switch(config-vrf)#`
		Configures static-route for destination prefix with Null0 nexthop and matching tag.

Step 4	**route-map** *map-name* [**permit** | **deny**] [*seq*]
	Example:
	`switch(config)# route-map SET_BHC permit 10`
	`switch(config-route-map)#`
		Creates a route map or enters route-map configuration mode for an existing route map. Use seq to order the entries in a route map.

Step 5	**match tag** <*tag-number*>
	Example:
	`switch(config-route-map)# match tag 6666`
	`switch(config-route-map)#`
		Matches the routes with the configured tag.

Step 6	**set weight** *value*
	Example:
	`switch (config-route-map)# set weight 65535`
	`switch(config-route-map)#`
		Sets the weight for the incoming route with blackhole community. we recommend to set the **set weight** value to maximum value, to give the highest precedence to the null routes. The maximum value of **set weight** is 65535.

Step 7	**set community blackhole**
	Example:
	`switch(config-route-map)# set community blackhole`
	`switch(config-route-map)#`
		Sets the community as Blackhole (well-known community).

	Command or Action	Purpose

Step 8	**router bgp** *as-number*
	Example:
	`switch(config)# router bgp 100`
	`switch(config-router)#`
		Enables a routing process. The range of as-num is 1–65535.

Step 9	**vrf** *vrf-name*
	Example:
	`switch(config-router)# vrf tenant-0001`
	`switch(config-router-vrf)#`
		Configures the tenant VRF.

Step 10	**address-family ipv4/ipv6 unicast**
	Example:
	`switch(config-router-vrf)# address-family ipv4 unicast`
	`switch(config-router-vrf-af)#`
		Configure the IPv4/IPv6 address family. This configuration is required for IPv4/IPv6 over VXLAN with IPv4/IPv6 underlay.

Step 11	**redistribute static route-map** *route-map name*
	Example:
	`switch(config-router-vrf-af)# redistribute static route-map SET_BHC`
	`switch(config-router-vrf-af)#`
		Redistributes the prefix-null static route into BGP using the configured route-map.

Configuring RPM Route-Map on Remote VTEP

Before you begin

On remote VTEP, use a community-list and route-map to give precedence to the null routes:

SUMMARY STEPS

1. **configure terminal**
2. **ip community-list standard** <*community-list-name*> **seq** <*seq-number*> **permit blackhole**
3. **route-map** *map-name***[permit | deny]** <*seq-number*>
4. **match community** <*community-list*>
5. **set weight** *value*
6. **route-map** *map-name***permit** <*seq-number*>
7. **router bgp** *as-number*
8. **route-map** *route-map* {**in** | **out**}

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# configure terminal`
	`switch(config)#`
		Enters global configuration mode.

Step 2	**ip community-list standard** <*community-list-name*> **seq** <*seq-number*> **permit blackhole**
	Example:
	`switch (config)# ip community-list standard BH seq 10 permit blackhole`
	`switch(config)#`
		Configures a community list and permits routes that have the well-known "blackhole" community value.
		Beginning with Cisco NX-OS Release 10.3(2)F, the blackhole (well-known community) is added to the existing IP community list.

Step 3	**route-map** *map-name***[permit | deny]** <*seq-number*>
	Example:
	`switch(config)# route-map PREFER_BHC permit 10`
	`switch(config-route-map)#`
		Enters route-map configuration mode

Step 4	**match community** <*community-list*>
	Example:
	`switch(config-route-map)# match community BH`
	`switch(config-route-map)#`
		The BGP routes are matched using the community list.

Step 5	**set weight** *value*
	Example:
	`switch (config-route-map)# set weight 65535`
	`switch(config-route-map)#`
		Sets the weight for the incoming route with blackhole community. we recommend to set the **set weight** value to maximum value, to give the highest precedence to the null routes. The maximum value of **set weight** is 65535.

Step 6	**route-map** *map-name***permit** <*seq-number*>
	Example:
	`switch(config-route-map)# route-map PREFER_BHC permit 20`
	`switch(config-route-map)#`
		Configures a route-map with a fallback permit clause to allow other routes.

Step 7	**router bgp** *as-number*
	Example:
	`switch(config)# router bgp 100`
	`switch(config-router)#`
		Enables a routing process. The range of as-num is from 1 to 65535.

Step 8	**route-map** *route-map* {**in** | **out**} **Example:** `switch(config-router-neighbor-af)# route-map PREFER_BHC in`	Applies the route map to the neighbor in the configured direction.

Configuration Example for Null Route

The following example shows how to set the local/remote configuration on prefix-null and MAC/MAC-IP drop routes:

Configuration – Prefix Null

On local VTEP (Border leaf switch) where the Type-5 null route is to be advertised, perform the following steps:

1. Configure static IPv4/IPv6 address with Null0 adjacency

```
vrf context tenant-0001
  vni 3100001
  ip route 50.1.0.0/24 Null0 tag 6666
  ipv6 route 50::1:0/120 Null0 tag 6666
```

2. Configure route-map to set null route community on static route and redistribute into BGP

```
route-map SET_BHC permit 10
  match tag 6666
  set community blackhole
router bgp 100
  router-id 10.1.0.21
  vrf tenant-0001
    address-family ipv4 unicast
      redistribute static route-map SET_BHC
    address-family ipv6 unicast
      redistribute static route-map SET_BHC
```

On all other remote VTEPs, perform the following steps:

1. Configure route-map to match the null route community and set weight to highest value to ensure null route is always preferred.

```
ip community-list standard BH seq 10 permit blackhole
route-map PREFER_BHC permit 10
  match community BH
  set weight 65535
route-map PREFER_BHC permit 20
router bgp 100
 router-id 10.1.0.13
 address-family l2vpn evpn
 template peer LEAF_to_FABRIC_IBGP_OVERLAY
    remote-as 100
    address-family l2vpn evpn
      send-community
      send-community extended
      route-map PREFER_BHC in
```

Configuration – MAC/MAC-IP Drop

On local VTEP where Type-2 null route is to be advertised, perform the following steps:

1. Configure static MAC address with drop adjacency

```
mac address-table static 0013.e001.0001 vlan 2 drop
```

2. Configure static ARP/ND neighbor for same address

```
interface Vlan2
  no shutdown
  vrf member tenant-0001
  ip address 5.0.63.254/18
  ipv6 address 5::3f7f/114
  ipv6 neighbor 5::17fe 0013.e001.0001
  no ipv6 redirects
  ip arp 5.0.23.254 0013.e001.0001
  fabric forwarding mode anycast-gateway
```

On all other remote VTEPs, perform the following step:

1. Configure route-map to match the blackhole community and set weight to highest value to ensure null route is always preferred.

```
ip community-list standard BH seq 10 permit blackhole
route-map PREFER_BHC permit 10
  match community BH
  set weight 65535
route-map PREFER_BHC permit 20
router bgp 100
 router-id 10.1.0.13
 address-family l2vpn evpn
 template peer LEAF_to_FABRIC_IBGP_OVERLAY
    remote-as 100
    address-family l2vpn evpn
      send-community
      send-community extended
      route-map PREFER_BHC in
  neighbor 10.1.0.31
    inherit peer LEAF_to_FABRIC_IBGP_OVERLAY
```

Verifying EVPN Null Route Configuration

To display the EVPN null route configuration information, enter one of the following commands:

Command	Purpose
show bgp l2vpn evpn	Displays routing table information.
show ip arp static vlan <*vlan-id*> **vrf** <*vrf-name*>	Displays local ARP information.
show ip arp static remote vlan <*vlan-id*> **vrf** <*vrf-name*>	Displays remote ARP information.
show ip adjacency vlan <*vlan-id*> **detail vrf** <*vrf-name*>	Displays local adjacency information.
show ipv6 icmp neighbour static remote [*vlan* <*id*>] [*vrf* <*name*>]	Displays remote static neighbor information.
show mac address-table static vlan <*vlan-id*>	Displays local/remote MAC information.
show ip community-list *name*	Displays information about a IP community list.
show route-map *name*	Displays information about a route map.

The following example shows Type-2 EVPN Route sample output for the **show bgp l2vpn evpn** command:

```
switch# show bgp l2vpn evpn 1111.1111.1111
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 53.53.53.53:32769    (L2VNI 1000002)
BGP routing table entry for [2]:[0]:[0]:[48]:[1111.1111.1111]:[32]:[100.100.100.51]/272,
version 23
Paths: (1 available, best #1)
Flags: (0x000102) (high32 00000000) on xmit-list, is not in l2rib/evpn
Multipath: eBGP iBGP
  Advertised path-id 1
  Path type: local, path is valid, is best path, no labeled nexthop, has esi_gw
  AS-Path: NONE, path locally originated
  53.53.53.53 (metric 0) from 0.0.0.0 (53.53.53.53)
  Origin IGP, MED not set, localpref 100, weight 32768
  Received label 1000002 1000100
  Community: Blackhole
  Extcommunity: RT:23456:1000002 RT:23456:1000100 ENCAP:8
  Router MAC:0476.b0f0.8157
  Path-id 1 advertised to peers:
  111.111.54.1
```

The following example shows Type-5 EVPN Route (sent) sample output for the **show bgp l2vpn evpn** command:

```
switch# sh bgp ipv4 uni 44.44.44.0 vrf 100
BGP routing table information for VRF 100, address family IPv4 Unicast
BGP routing table entry for 44.44.44.0/24, version 6
Paths: (1 available, best #1)
Flags: (0x80c0002) (high32 0x000020) on xmit-list, is not in urib, exported, has label
vpn: version 5, (0x00000000100002) on xmit-list
local label: 492287

Advertised path-id 1, VPN AF advertised path-id 1
Path type: redist, path is valid, is best path, no labeled nexthop, is extd
AS-Path: NONE, path locally originated
0.0.0.0 (metric 0) from 0.0.0.0 (44.44.44.44)
Origin incomplete, MED 0, localpref 100, weight 32768
Community: blackhole
Extcommunity: RT:23456:1000100

VRF advertise information:
Path-id 1 not advertised to any peer

VPN AF advertise information:
Path-id 1 not advertised to any peer

switch# sh bgp l2 e 44.44.44.0
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 53.53.53.53:4 (L3VNI 1000100)
BGP routing table entry for [5]:[0]:[0]:[24]:[44.44.44.0]/224, version 5
Paths: (1 available, best #1)
Flags: (0x000002) (high32 00000000) on xmit-list, is not in l2rib/evpn
Multipath: eBGP iBGP

Advertised path-id 1
Path type: local, path is valid, is best path, no labeled nexthop, has esi_gw
Gateway IP: 0.0.0.0
AS-Path: NONE, path locally originated
53.53.53.53 (metric 0) from 0.0.0.0 (53.53.53.53)
Origin incomplete, MED 0, localpref 100, weight 32768
Received label 1000100
Community: blackhole
Extcommunity: RT:23456:1000100 ENCAP:8 Router MAC:0476.b0f0.8157
```

```
Path-id 1 advertised to peers:
111.111.54.1
```

The following example shows Type-5 EVPN Route (received) sample output for the **show bgp l2vpn evpn** command:

```
switch# sh bgp l2 e 44.44.44.0
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 53.53.53.53:4
BGP routing table entry for [5]:[0]:[0]:[24]:[44.44.44.0]/224, version 2
Paths: (1 available, best #1)
Flags: (0x000002) (high32 00000000) on xmit-list, is not in l2rib/evpn, is not in HW
Multipath: eBGP iBGP

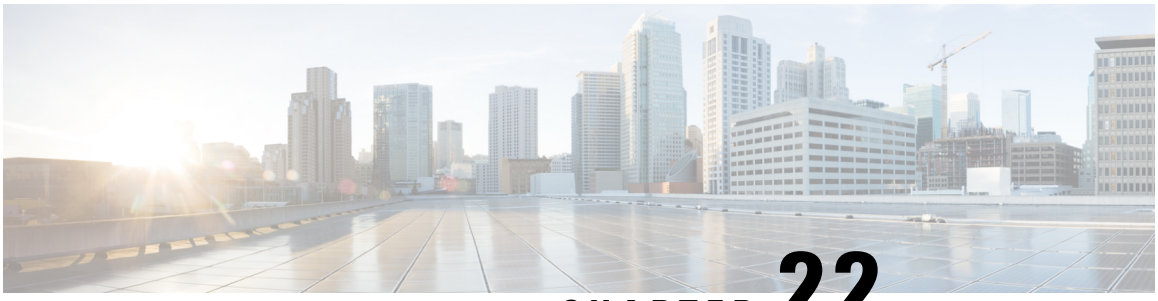
Advertised path-id 1
Path type: external, path is valid, is best path, no labeled nexthop, has esi_gw
Imported to 2 destination(s)
Imported paths list: 100 L3-1000100
Gateway IP: 0.0.0.0
AS-Path: 4241653625 , path sourced external to AS
53.53.53.53 (metric 2) from 111.111.53.1 (53.53.53.53)
Origin incomplete, MED 0, localpref 100, weight 0
Received label 1000100
Community: blackhole
Extcommunity: RT:11000:1000100 Route-Import:53.53.53.53:100
Source AS:4241653625:0 SOO:50529024:00000000 ENCAP:8
Router MAC:0476.b0f0.8157
Path-id 1 not advertised to any peer

switch# show bgp ipv4 uni 44.44.44.0 vrf 100
BGP routing table information for VRF 100, address family IPv4 Unicast
BGP routing table entry for 44.44.44.0/24, version 3
Paths: (1 available, best #1)
Flags: (0x8008001a) (high32 00000000) on xmit-list, is in urib, is best urib route, is in
HW
vpn: version 3, (0x00000000100002) on xmit-list

Advertised path-id 1, VPN AF advertised path-id 1
Path type: external, path is valid, is best path, no labeled nexthop, in rib, has esi_gw

Imported from 53.53.53.53:4:[5]:[0]:[0]:[24]:[44.44.44.0]/224
AS-Path: 4241653625 , path sourced external to AS
53.53.53.53 (metric 2) from 111.111.53.1 (53.53.53.53)
Origin incomplete, MED 0, localpref 100, weight 0
Received label 1000100
Community: blackhole
Extcommunity: RT:11000:1000100 Route-Import:53.53.53.53:100
Source AS:4241653625:0 SOO:50529024:00000000 ENCAP:8
Router MAC:0476.b0f0.8157

VRF advertise information:
Path-id 1 not advertised to any peer
```



CHAPTER 22

Configuring Port VLAN Mapping

This chapter contains these sections:

- About Translating Incoming VLANs, on page 537
- Guidelines and Limitations for Port VLAN Mapping, on page 538
- Configuring Port VLAN Mapping on a Trunk Port, on page 541
- Configuring Inner VLAN and Outer VLAN Mapping on a Trunk Port, on page 543
- About Port Multi-VLAN Mapping, on page 545
- Guidelines and Limitations for Port Multi-VLAN Mapping, on page 546
- Configuring Port Multi-VLAN Mapping , on page 547

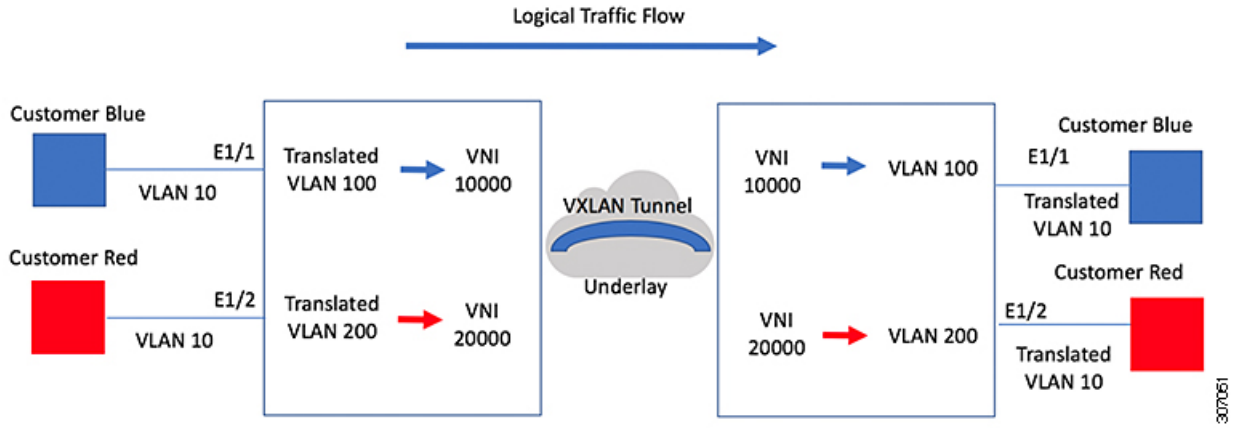
About Translating Incoming VLANs

Sometimes a VLAN translation is required or desired. One such use case is when a service provider has multiple customers connecting to the same physical switch using the same VLAN encapsulation, but they are not and should not be on the same Layer 2 segment. In such cases translating the incoming VLAN to a unique VLAN that is then mapped to a VNI is the right way to extending the segment. In the figure below two customers, Blue and Red are both connecting to the leaf using VLAN 10 as their encapsulation.

Customers Blue and Red should not be on the same VNI. In this example VLAN 10 for Customer Blue (on interface E1/1) is mapped/translated to VLAN 100, and VLAN 10 for customer Red (on interface E1/2) is mapped to VLAN 200. In turn, VLAN 100 is mapped to VNI 10000 and VLAN 200 is mapped to VNI 20000.

On the other leaf, this mapping is applied in reverse. Incoming VXLAN encapsulated traffic on VNI 10000 is mapped to VLAN 100 which in turn is mapped to VLAN 10 on Interface E1/1. VXLAN encapsulated traffic on VNI 20000 is mapped to VLAN 200 which in turn is mapped to VLAN 10 on Interface E1/2.

Figure 48: Logical Traffic Flow



You can configure VLAN translation between the ingress (incoming) VLAN and a local (translated) VLAN on a port. For the traffic arriving on the interface where VLAN translation is enabled, the incoming VLAN is mapped to a translated VLAN that is VXLAN enabled.

On the underlay, this is mapped to a VNI, the inner dot1q is deleted, and switched over to the VXLAN network. On the egress switch, the VNI is mapped to a translated VLAN. On the outgoing interface, where VLAN translation is configured, the traffic is converted to the original VLAN and egressed out. Refer to the VLAN counters on the translated VLAN for the traffic counters and not on the ingress VLAN. Port VLAN (PV) mapping is an access side feature and is supported with both multicast and ingress replication for flood and learn and MP-BGP EVPN mode for VXLAN.

Guidelines and Limitations for Port VLAN Mapping

The following are the guidelines and Limitations for Port VLAN Mapping:

- Support is added for vPC Fabric Peering.
- Beginning with Cisco NX-OS Release 10.3(3)F, VLAN translation is supported on both VXLAN and non-VXLAN VLANs.
- The ingress (incoming) VLAN does not need to be configured on the switch as a VLAN. The translated VLAN needs to be configured and a vn-segment mapping given to it. An NVE interface with VNI mapping is essential for the same.
- All Layer 2 source address learning and Layer 2 MAC destination lookup occurs on the translated VLAN. Refer to the VLAN counters on the translated VLAN and not on the ingress (incoming) VLAN.
- Cisco Nexus 9300 and 9500 switches support switching and routing on overlapped VLAN interfaces. Only VLAN-mapping switching is applicable for Cisco Nexus 9300-FX/FX2/FX3 platform switches and Cisco Nexus 9500 with X97160YC-EX, 9700-FX/FX3 line cards.
- Port VLAN routing is supported on the following platforms:
  - Beginning with Cisco NX-OS Release 7.x, this feature is supported on Cisco Nexus 9300-FX/FX2 platform switches.

  - Beginning with Cisco NX-OS Release 9.2(x), this feature is supported on Cisco Nexus 9300-GX platform switches.
  - Beginning with Cisco NX-OS Release 9.3(x), this feature is supported on Cisco Nexus 9300-FX3 platform switches.
  - Beginning with Cisco NX-OS Release 10.2(3)F, this feature is supported on the Cisco Nexus 9300-GX2 platform switches.
  - Beginning with Cisco NX-OS Release 10.4(1)F, this feature is supported on the Cisco Nexus 9332D-H2R switches.
  - Beginning with Cisco NX-OS Release 10.4(2)F, this feature is supported on the Cisco Nexus 93400LD-H1 switches.
  - Beginning with Cisco NX-OS Release 10.4(3)F, this feature is supported on the Cisco Nexus 9364C-H1 switches.
  - Beginning with Cisco NX-OS Release 10.5(2)F, this feature is supported on the Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.

- Beginning with Cisco NX-OS Release 9.3(3), PV Translation is supported for Cisco Nexus 9300-GX platform switches.
- Beginning with Cisco NX-OS Release 10.2(3)F, PV Translation is supported on the Cisco Nexus 9300-GX2 platform switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, PV Translation is supported on the Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, PV Translation is supported on the Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, PV Translation is supported on the Cisco Nexus 9364C-H1 switches.
- On Cisco Nexus 9300 Series switches with NFE ASIC, PV routing is not supported on 40 G ALE ports.
- PV routing supports configuring an SVI on the translated VLAN for flood and learn and BGP EVPN mode for VXLAN.
- When changing a property on a translated VLAN, the port that has a mapping configuration with that VLAN as the translated VLAN, must be flapped to ensure correct behavior. This is applicable only to the following platforms:
  - N9K-C9504 modules
  - N9K-C9508 modules
  - N9K-C9516 modules
  - Nexus 9500 line cards
  - Nexus 9600 line cards
  - Nexus 9700-X Cloud Scale line cards
  - Nexus 9600-R and R2 line cards

```
Int eth 1/1
switchport vlan mapping 101 10
.
.
.

/***Deleting vn-segment from vlan 10.***/
/***Adding vn-segment back.***/
/***Flap Eth 1/1 to ensure correct behavior.***/
```

- The following example shows incoming VLAN 10 being mapped to local VLAN 100. Local VLAN 100 will be the one mapped to a VXLAN VNI.

```
interface ethernet1/1
switchport vlan mapping 10 100
```

- The following is an example of overlapping VLAN for PV translation. In the first statement, VLAN-102 is a translated VLAN with VNI mapping. In the second statement, VLAN-102 the VLAN where it is translated to VLAN-103 with VNI mapping.

```
interface ethernet1/1
switchport vlan mapping 101 102
switchport vlan mapping 102 103/
```

- When adding a member to an existing port channel using the force command, the "mapping enable" configuration must be consistent. For example:

```
Int po 101
switchport vlan mapping enable
switchport vlan mapping 101 10
switchport trunk allowed vlan 10

int eth 1/8
/***No configuration***/
```



Note The **switchport vlan mapping enable** command is supported only when the port mode is trunk.

- VLAN mapping helps with VLAN localization to a port, scoping the VLANs per port. A typical use case is in the service provider environment where the service provider leaf switch has different customers with overlapping VLANs that come in on different ports. For example, customer A has VLAN 10 coming in on Eth 1/1 and customer B has VLAN 10 coming in on Eth 2/2.

  In this scenario, you can map the customer VLAN to a provider VLAN and map that to a Layer 2 VNI. There is an operational benefit in terminating different customer VLANs and mapping them to the fabric-managed VLANs, L2 VNIs.

- An NVE interface with VNI mapping must be configured for Port VLAN translation to work.

- You should not enable super bridging VLAN in the provider VLAN list of the **system dot1q-tunnel transit vlan** *id* command. If enabled it will end up in unrecoverable functional and forwarding impacts.

- Port VLAN mapping is not supported on FEX ports.

- Beginning with Cisco NX-OS Release 10.3(3)F, IPv6 underlay is supported on Port VLAN Mapping for VXLAN EVPN on Cisco Nexus 9300-FX/FX2/FX3/GX/GX2 switches and Cisco Nexus 9500 switches with X97160YC-EX and 9700-FX/GX line cards.

- Beginning with Cisco NX-OS Release 10.4(1)F, IPv6 underlay is supported on Port VLAN Mapping for VXLAN EVPN on Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, IPv6 underlay is supported on Port VLAN Mapping for VXLAN EVPN on Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, IPv6 underlay is supported on Port VLAN Mapping for VXLAN EVPN on Cisco Nexus 9364C-H1 switches.
- Beginning with Cisco NX-OS Release 10.5(2)F, IPv6 underlay is supported on Port VLAN Mapping for VXLAN EVPN is supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.

Configuring Port VLAN Mapping on a Trunk Port

Before you begin

- Ensure that the physical or port channel on which you want to implement VLAN translation is configured as a Layer 2 trunk port.
- Ensure that the translated VLANs are created on the switch and are also added to the Layer 2 trunk ports trunk-allowed VLAN vlan-list.

Note As a best practice, do not add the ingress VLAN ID to the switchport allowed vlan-list under the interface.

- Ensure that all translated VLANs are VXLAN enabled.

SUMMARY STEPS

1. **configure terminal**
2. **interface** *type/port*
3. **[no] switchport vlan mapping enable**
4. **[no] switchport vlan mapping** *vlan-id translated-vlan-id*
5. **[no] switchport vlan mapping all**
6. **copy running-config startup-config**
7. **show interface** [*if-identifier*] **vlan mapping**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.

	Command or Action	Purpose
Step 2	**interface** *type/port* **Example:** `switch(config)# interface Ethernet1/1`	Specifies the interface that you are configuring.
Step 3	**[no] switchport vlan mapping enable** **Example:** `switch(config-if)# [no] switchport vlan mapping enable`	Enables VLAN translation on the switch port. VLAN translation is disabled by default. **Note** Use the **no** form of this command to disable VLAN translation.
Step 4	**[no] switchport vlan mapping** *vlan-id translated-vlan-id* **Example:** `switch(config-if)# switchport vlan mapping 10 100`	Translates a VLAN to another VLAN. • The range for *vlan-id* is from 1 to 4094. For *translated-vlan-id*, only non-reserved VLAN IDs are allowed. • You can configure VLAN translation between the ingress (incoming) VLAN and a local (translated) VLAN on a port. For the traffic arriving on the interface where VLAN translation is enabled, the incoming VLAN is mapped to a translated VLAN that is VXLAN enabled. On the underlay, this is mapped to a VNI, the inner dot1q is deleted, and switched over to the VXLAN network. On the egress switch, the VNI is mapped to a local translated VLAN. On the outgoing interface, where VLAN translation is configured, the traffic is converted to the original VLAN and egresses out. **Note** Use the **no** form of this command to clear the mappings between a pair of VLANs.
Step 5	**[no] switchport vlan mapping all** **Example:** `switch(config-if)# switchport vlan mapping all`	Removes all VLAN mappings configured on the interface.
Step 6	**copy running-config startup-config** **Example:** `switch(config-if)# copy running-config startup-config`	Copies the running configuration to the startup configuration. **Note** The VLAN translation configuration does not become effective until the switch port becomes an operational trunk port.
Step 7	**show interface** [*if-identifier*] **vlan mapping** **Example:** `switch# show interface ethernet1/1 vlan mapping`	Displays VLAN mapping information for a range of interfaces or for a specific interface.

Example

This example shows how to configure VLAN translation between (the ingress) VLAN 10 and (the local) VLAN 100. The show vlan counters command output shows the statistic counters as translated VLAN instead of customer VLAN.

```
switch#  configure terminal
switch(config)#  interface ethernet1/1
switch(config-if)#  switchport vlan mapping enable
switch(config-if)#  switchport vlan mapping 10 100
switch(config-if)#  switchport trunk allowed vlan 100
switch(config-if)#  show interface ethernet1/1 vlan mapping
Interface eth1/1:
Original VLAN           Translated VLAN
------------------      ---------------
10                           100

switch(config-if)#  show vlan counters
Vlan Id                              :100
Unicast Octets In                    :292442462
Unicast Packets In                   :1950525
Multicast Octets In                  :14619624
Multicast Packets In                 :91088
Broadcast Octets In                  :14619624
Broadcast Packets In                 :91088
Unicast Octets Out                   :304012656
Unicast Packets Out                  :2061976
L3 Unicast Octets In                 :0
L3 Unicast Packets In                :0
```

Configuring Inner VLAN and Outer VLAN Mapping on a Trunk Port

Configuring Inner VLAN and Outer VLAN Mapping on a Trunk Port is applicable only for Cisco Nexus 9300 platforms and not supported on Cisco Nexus 9300-FX, 9300-FX2, 9300-FX3, 9300-GX, 9300-GX2, 9332D-H2R, 93400LD-H1, 9364C-H1 platforms.

You can configure VLAN translation from an inner VLAN and an outer VLAN to a local (translated) VLAN on a port. For the double tag VLAN traffic arriving on the interfaces where VLAN translation is enabled, the inner VLAN and outer VLAN are mapped to a translated VLAN that is VXLAN enabled.

Notes for configuring inner VLAN and outer VLAN mapping:

- Inner and outer VLAN cannot be on the trunk allowed list on a port where inner VLAN and outer VLAN is configured.

  For example:

  ```
  switchport vlan mapping 11 inner 12 111
  switchport trunk allowed vlan 11-12,111 /***Not valid because 11 is outer VLAN and 12
  is inner VLAN.***/
  ```

- On the same port, no two mapping (translation) configurations can have the same outer (or original) or translated VLAN. Multiple inner VLAN and outer VLAN mapping configurations can have the same inner VLAN.

  For example:

```
switchport vlan mapping 101 inner 102 1001
switchport vlan mapping 101 inner 103 1002  /***Not valid because 101 is already used
as an original VLAN.***/
switchport vlan mapping 111 inner 104 1001  /***Not valid because 1001 is already used
 as a translated VLAN.***/
switchport vlan mapping 106 inner 102 1003  /***Valid because inner vlan can be the
same.***/
```

- When a packet comes double-tagged on a port which is enabled with the inner option, only bridging is supported.
- VXLAN PV routing is not supported for double-tagged frames.

SUMMARY STEPS

1. **configure terminal**
2. **interface** *type port*
3. [**no**] **switchport mode trunk**
4. **switchport vlan mapping enable**
5. **switchport vlan mapping** *outer-vlan-id* **inner** *inner-vlan-id translated-vlan-id*
6. (Optional) **copy running-config startup-config**
7. (Optional) **show interface** [*if-identifier*] **vlan mapping**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**interface** *type port*	Enters interface configuration mode.
Step 3	[**no**] **switchport mode trunk**	Enters trunk configuration mode.
Step 4	**switchport vlan mapping enable**	Enables VLAN translation on the switch port. VLAN translation is disabled by default. **Note** Use the**no** form of this command to disable VLAN translation.
Step 5	**switchport vlan mapping** *outer-vlan-id* **inner** *inner-vlan-id translated-vlan-id*	Translates inner VLAN and outer VLAN to another VLAN.
Step 6	(Optional) **copy running-config startup-config**	Copies the running configuration to the startup configuration.

	Command or Action	Purpose
		Note The VLAN translation configuration does not become effective until the switch port becomes an operational trunk port
Step 7	(Optional) **show interface** [*if-identifier*] **vlan mapping**	Displays VLAN mapping information for a range of interfaces or for a specific interface.

Example

This example shows how to configure translation of double tag VLAN traffic (inner VLAN 12; outer VLAN 11) to VLAN 111.

```
switch# configure terminal
switch(config)# interface ethernet1/1
switch(config-if)# switchport mode trunk
switch(config-if)# switchport vlan mapping enable
switch(config-if)# switchport vlan mapping 11 inner 12 111
switch(config-if)# switchport trunk allowed vlan 101-170
switch(config-if)# no shutdown

switch(config-if)# show mac address-table dynamic vlan 111

Legend:
        * - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC
        age - seconds since last seen,+ - primary entry using vPC Peer-Link,
        (T) - True, (F) - False
   VLAN     MAC Address      Type      age     Secure NTFY Ports
---------+-----------------+--------+---------+------+----+------------------
*  111     0000.0092.0001   dynamic  0         F      F    nve1(100.100.100.254)
*  111     0000.0940.0001   dynamic  0         F      F    Eth1/1
```

About Port Multi-VLAN Mapping

With Port Multi-VLAN Mapping feature multiple VLANs are mapped on a trunk interface to a single global VLAN/VNI. Layer 2 (L2) sub-interface has to be created for the mapping and a qTag has to be provided for each L2 sub-interface.

Different Port-VLANs can serve different services on the same physical interface.

For the Port Multi-VLAN mappings per trunk port, ACLs are installed per each of the mapping using L2 sub-interface. Some ACLs are installed automatically by default and some are installed with static MAC address configuration. L2 sub-interface has a qtag, flood-domain or provider-VLAN. The provider-VLAN is configured on the switch and is used for traffic forwarding. There can be only one provider-VLAN on the switch.

This static MAC configuration is done using the **switchport mac-address static-only** command configured on L2 sub-interface parent port. This command disables the MAC learning on the parent port and enables MAC-ACL per each static MAC configured on the L2 sub-interfaces.

Guidelines and Limitations for Port Multi-VLAN Mapping

The following are the guidelines and limitations for Port Multi-VLAN Mapping:

- Beginning with Cisco NX-OS Release 10.2(3)F, the Port Multi-VLAN feature is supported on N9K-C9316D-GX, N9K-C93600CD-GX, N9K-C9364C-GX, and Cisco Nexus 9300-GX2 switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, the Port Multi-VLAN feature is supported on Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, the Port Multi-VLAN feature is supported on the Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, the Port Multi-VLAN feature is supported on the Cisco Nexus 9364C-H1 switches.
- Beginning with Cisco NX-OS Release 10.1(2), Port Multi-VLAN Mapping is supported on Cisco Nexus 9300-FX, and FX2 platform switches.
- Beginning with Cisco NX-OS Release 10.2(3)F, Port Multi-VLAN Mapping is supported on the Cisco Nexus 9300-FX3 platform switches.
- Port Multi-VLAN Mapping is an access side feature and is supported with both multicast and ingress replication for VXLAN flood and learn mode. This feature is not supported for VXLAN MP-BGP EVPN mode in Cisco NX-OS Release 10.1(2).
- For a device that is running on Cisco Nexus Release 10.1(2) or Cisco Nexus Release 10.2(1)F ND-ISSU is not supported if L2 sub-interfaces are configured.
- This feature is not supported with vPC fabric peering configuration.
- In order to protect against broadcast or multicast flood, all flooding traffic is dropped except ARP and NS/ND.
- Layer 2 is supported.
- STP is not supported.
- Static default route or specific route to remote VTEP is recommended to be configured on ToRs.
- Interaction with other access features like QinQ/QinVNI, Port VLAN mapping, PVLAN and Xconnect are not supported.

The following are the guidelines and limitations related to the parent interface:

- TCAM entries are only installed on the slice where the parent port exists. To check TCAM utilization, use the **show system internal access-list resource utilization** command.
- To check the port slice, use the **show interface hardware-mappings** command.
- For hosts using static ARP, add on ToR static MAC entry for remote host on interface nve 1. Example:

```
mac address-table static 0034.0100.0001 vni 10013001 interface nve 1 peer-ip 192.168.75.2
```

- Port-security/dot1x is not supported on the parent interface.
- vPC mode is not supported for parent interface or L2 sub interface.

The following are the guidelines and limitations related to the sub interface:

- Maximum of 510 sub-interfaces are supported per switch.
- ACL and storm-control per sub-interface cannot be configured under the switch port mapping.
- TCAM region must be re-configured in order to support Max 510 L2 sub interfaces. For each L2 sub interface nine TCAM ing-pacl-sb entries are allocated.
- Static MAC is configured on L2 sub interface using the **switchport mac-address static-only** command on the parent interface.
- L2 sub interfaces are not supported without VXLAN deployment. The provider VLAN must be a VXLAN VLAN.
- Dynamic MAC learning is disabled on L2 sub interface.
- Storm control is not supported for L2 sub interface.
- The **hardware profile svi-and-si flex-stats-enable** command supports only ingress L2 sub interface counters. This profile statistics command does not support egress L2 sub interface counters and VxLAN statistics.
- IGMP snooping is not supported on the provider VLAN where L2 sub interface is configured.

Configuring Port Multi-VLAN Mapping

A sample configuration of Port Multi-VLAN Mapping is provided below:

```
feature ospf
feature pim
feature bfd
feature interface-vlan
feature vn-segment-vlan-based
feature private-vlan
feature lacp
feature nv overlay

hardware access-list tcam region ing-pacl-sb 2560
hardware profile svi-and-si flex-stats-enable

ip pim rp-address 2.0.0.254 group-list 224.0.0.0/4

vlan 3001
  vn-segment 10013001

interface Ethernet1/22
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 3001
  mtu 9216
  storm-control broadcast level 0.01
  storm-control action trap
  switchport isolated
  switchport mac-address static-only
  no shutdown

interface Ethernet1/22.1
  encapsulation dot1q 301 provider-vlan 3001
  no shutdown
```

```
interface Ethernet1/22.2
  encapsulation dot1q 302 provider-vlan 3001
  no shutdown

interface Ethernet1/22.3
  encapsulation dot1q 303 provider-vlan 3001
  no shutdown

interface Ethernet1/22.4
  encapsulation dot1q 304 provider-vlan 3001
  no shutdown

interface Ethernet1/22.5
  encapsulation dot1q 305 provider-vlan 3001
  no shutdown

interface port-channel1
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 3001
  mtu 9216
  storm-control broadcast level 0.01
  storm-control multicast level 0.01
  storm-control unicast level 0.01
  storm-control action trap
  switchport isolated
  switchport mac-address static-only

interface port-channel1.1
  encapsulation dot1q 301 provider-vlan 3001
  no shutdown

interface port-channel1.2
  encapsulation dot1q 302 provider-vlan 3001
  no shutdown

interface port-channel1.3
  encapsulation dot1q 303 provider-vlan 3001
  no shutdown

interface port-channel1.4
  encapsulation dot1q 304 provider-vlan 3001
  no shutdown

interface port-channel1.5
  encapsulation dot1q 305 provider-vlan 3001
  no shutdown

interface Ethernet1/24
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 3001
  mtu 9216
  storm-control broadcast level 0.01
  storm-control multicast level 0.01
  storm-control unicast level 0.01
  storm-control action trap
  switchport isolated
  switchport mac-address static-only
  channel-group 1 mode active
  no shutdown

interface Ethernet1/25
```

```
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 3001
  mtu 9216
  storm-control broadcast level 0.01
  storm-control multicast level 0.01
  storm-control unicast level 0.01
  storm-control action trap
  switchport isolated
  switchport mac-address static-only
  channel-group 1 mode active
  no shutdown

mac address-table static 0035.0100.0001 vlan 3001 interface Ethernet1/22.1
mac address-table static 0035.0100.0002 vlan 3001 interface Ethernet1/22.2
mac address-table static 0035.0100.0003 vlan 3001 interface Ethernet1/22.3
mac address-table static 0035.0100.0004 vlan 3001 interface Ethernet1/22.4
mac address-table static 0035.0100.0005 vlan 3001 interface Ethernet1/22.5

mac address-table static 003b.0100.0001 vlan 3001 interface port-channel1.1
mac address-table static 003b.0100.0002 vlan 3001 interface port-channel1.2
mac address-table static 003b.0100.0003 vlan 3001 interface port-channel1.3
mac address-table static 003b.0100.0004 vlan 3001 interface port-channel1.4
mac address-table static 003b.0100.0005 vlan 3001 interface port-channel1.5


router ospf p1
  bfd
  router-id 192.168.210.1

interface loopback0
  ip address 192.168.210.1/32
  ip router ospf p1 area 0.0.0.0
  ip pim sparse-mode

interface loopback1
  description NVE_IP
  ip address 192.168.210.2/32
  ip router ospf p1 area 0.0.0.0
  ip pim sparse-mode

interface Ethernet1/49
  mtu 9216
  no ip redirects
  ip address 10.0.1.16/31
  ip router ospf p1 area 0.0.0.0
  ip pim sparse-mode
  no shutdown

interface Ethernet1/54
  mtu 9216
  no ip redirects
  ip address 10.0.1.18/31
  ip router ospf p1 area 0.0.0.0
  ip pim sparse-mode
  no shutdown

interface nve1
  no shutdown
  source-interface loopback1
 member vni 10013001
    mcast-group 227.1.1.1
```

The following examples provide show command outputs related to Port Multi-VLAN Mapping:

```
switch# show hardware access-list resource utilization | grep Super

Ingress PACL Super Bridge               2445    115     95.50
 Ingress PACL Super Bridge IPv4         0               0.00
 Ingress PACL Super Bridge IPv6         0               0.00
 Ingress PACL Super Bridge MAC          0               0.00
 Ingress PACL Super Bridge ALL          1956            76.40
 Ingress PACL Super Bridge OTHER        489             19.10

switch # show hardware access-list resource entries | in Super

Ingress PACL Super Bridge                : 2445 valid entries   115 free entries

switch# show interface ethernet 1/22.1-5 brief
--------------------------------------------------------------------------------
Ethernet        VLAN    Type Mode   Status  Reason                 Speed     Port
Interface                                                                    Ch #
--------------------------------------------------------------------------------
Eth1/22.1       301     eth  trunk  up      none                    10G(D) --
Eth1/22.2       302     eth  trunk  up      none                    10G(D) --
Eth1/22.3       303     eth  trunk  up      none                    10G(D) --
Eth1/22.4       304     eth  trunk  up      none                    10G(D) --
Eth1/22.5       305     eth  trunk  up      none                    10G(D) --


switch# show interface port-channel 1.1-5 brief
--------------------------------------------------------------------------------------------------
Port-channel VLAN    Type Mode   Status  Reason     Speed   Protocol
Interface
--------------------------------------------------------------------------------------------------
Po1.1        301     eth  trunk  up      none       a-10G(D)    --
Po1.2        302     eth  trunk  up      none       a-10G(D)    --
Po1.3        303     eth  trunk  up      none       a-10G(D)    --
Po1.4        304     eth  trunk  up      none       a-10G(D)    --
Po1.5        305     eth  trunk  up      none       a-10G(D)    --

switch# show interface ethernet 1/22.1 counters

--------------------------------------------------------------------------------
Port                                    InOctets                    InUcastPkts
--------------------------------------------------------------------------------
Eth1/22.1                          1145503766466                     125246421

--------------------------------------------------------------------------------
Port                                  InMcastPkts                    InBcastPkts
--------------------------------------------------------------------------------
Eth1/22.1                                    0                              0

--------------------------------------------------------------------------------
Port                                   OutOctets                   OutUcastPkts
--------------------------------------------------------------------------------
Eth1/22.1                                    0                              0

--------------------------------------------------------------------------------
Port                                 OutMcastPkts                   OutBcastPkts
--------------------------------------------------------------------------------
Eth1/22.1                                    0                              0

switch# show consistency-checker l2 sub-interface port-channel 1.1

Getting details for port-channel1.1 (0x16001000)
========================================
Running CC for port-channel1.1
========================================
```

```
CC for Permit Static: PASSED
CC for Deny ACL: PASSED
CC for Permit ARP ACL: PASSED
CC for Permit Multi-Dest ACL: PASSED
CC for info_src_idx: PASSED
CC for info_bd_xlate_idx: PASSED
CC for info_vlan_mbr_chk_bypasss: PASSED
CC for info_set_dont_learn: PASSED
CC for VlanXlate Table: PASSED
CC for BD State Table: PASSED
CC for QSMT BD State Table: PASSED
CC for Local Multipath Table: PASSED
CC for Rw VifTable: PASSED
CC for Rwx VlanXlate Table: PASSED


switch# show system internal access-list interface eth 1/22.1

slot 1
=======

Policies in ingress direction:
Policy type Policy Id Policy name
------------------------------------------------------------
PACL Super Bridge 341 l2fm-acl-mac-Eth1/22.1
PACL Super Bridge 342 l2fm-acl-ipv6-Eth1/22.1

No Netflow profiles in ingress direction

INSTANCE 0x0
---------------

Tcam 20 resource usage:
----------------------
LBL AB = 0x11
Bank 0
------
IPv6 Class
Policies: PACL Super Bridge(l2fm-acl-ipv6-Eth1/22.1)
Netflow profile: 0
Netflow deny profile: 0
2 tcam entries
MAC Class
Policies: PACL Super Bridge(l2fm-acl-mac-Eth1/22.1)
Netflow profile: 0
Netflow deny profile: 0
3 tcam entries

0 l4 protocol cam entries
0 mac etype/proto cam entries
0 lous
0 tcp flags table entries
0 adjacency entries

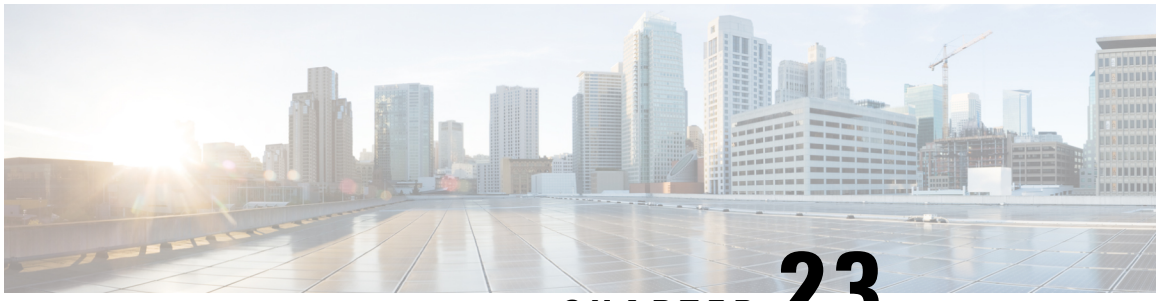
No egress policies
No Netflow profiles in egress direction


switch# show system internal access-list interface eth 1/22.1 input statistics

slot 1
=======
INSTANCE 0x0
```

```
---------------

Tcam 20 resource usage:
----------------------
LBL AB = 0xb
Bank 0
------
IPv6 Class
Policies: PACL Super Bridge(l2fm-acl-ipv6-Eth1/22.1)
Netflow profile: 0
Netflow deny profile: 0
Entries:
[Index] Entry [Stats]
---------------------
[0x0038:0x0038:0x0038] permit lbl(0x0) 0000.0000.0000 ffff.ffff.ffff 0000.0000.0000
ffff.ffff.ffff vlan 502 [9]
[0x003a:0x003a:0x003a] permit lbl(0x0) 0000.0000.0000 ffff.ffff.ffff 0000.0000.0000
ffff.ffff.ffff vlan 502 [0]
MAC Class
Policies: PACL Super Bridge(l2fm-acl-mac-Eth1/22.1)
Netflow profile: 0
Netflow deny profile: 0
Entries:
[Index] Entry [Stats]
---------------------
[0x003c:0x003c:0x003c] permit lbl(0x0) arp [7]
[0x003d:0x08de:0x08de] permit lbl(0x0) 0035.0100.0001 ffff.ffff.ffff 0000.0000.0000
ffff.ffff.ffff vlan 502 [6279856]
[0x08dd:0x08e0:0x08e0] deny lbl(0x0) 0000.0000.0000 ffff.ffff.ffff 0000.0000.0000
ffff.ffff.ffff vlan 502 [279]
```



CHAPTER 23

Micro-segmentation for VXLAN Fabrics Using Group Policy Option (GPO)

- Overview, on page 553
- GPO, on page 553
- Terminology, on page 554
- Guidelines and Limitations, on page 555
- Configuring Micro-segmentation using GPO, on page 558
- Configuration Examples for GPO, on page 568
- Verifying GPO, on page 569
- VXLAN Multi-Site and GPO Interoperability, on page 572

Overview

Network administrators can use micro-segmentation to logically group network resources based on specific criteria such as application attributes. You can use micro-segmentation with Security Groups (SGs) and Security Group ACLs (SGACLs) to create and enforce tailored application centric security policies between security groups regardless of network topology.

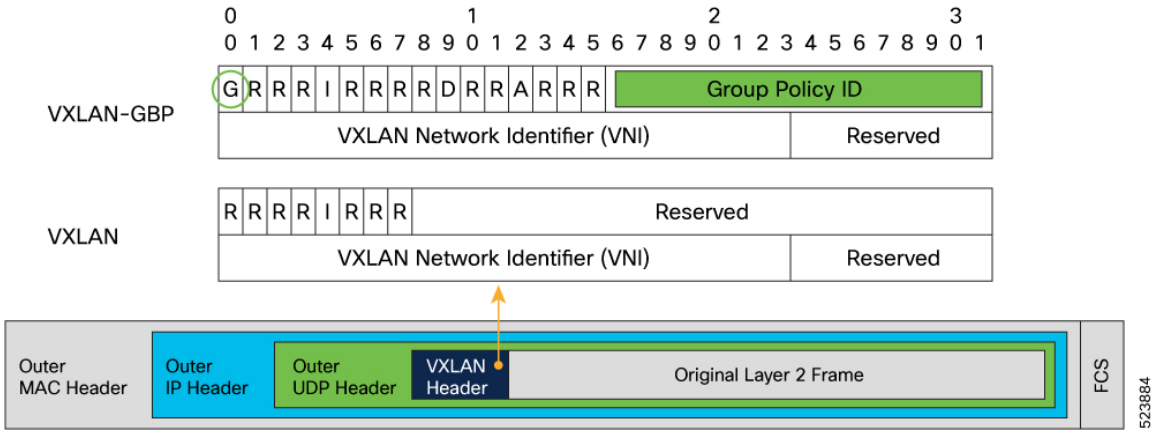
In traditional data center environments, the application or workload security is often implemented at the perimeter or the north-south boundary where users from outside the data center fabric enter. This is often implemented using perimeter firewalls and other security inspection devices. However, this approach is not effective against the advanced nature of the latest attacks. The attack surface spans the entire data center including the east-west and north-south flows.

Using micro-segmentation with security groups and security group ACLs, this feature can provide an effective security solution to the users of NX-OS platforms. Micro-segmentation provides more flexibility and lower complexity than traditional general purpose Access Control Lists (ACLs). With micro-segmentation, organizations can provide specific policies that dictate how the application workloads communicate regardless of where these applications reside within the network.

GPO

Group Policy Option (GPO) is a backward-compatible extension to VXLAN that adds a Security Group Tag (SGT) to the VXLAN header for security policy enforcement purposes.

Figure 49: Security Group Tags on VXLAN Header



In a GPO enabled VXLAN network, you can create Security Groups in the VXLAN EVPN fabrics to define segmentation. By defining smaller, isolated application segments, you can deploy micro-segmentation policies that allow for better control over the flow of network traffic among the application tiers and across applications. Micro-segmentation ensures that security policies are applied only where they are needed, improving application and workload security, thereby improving the security posture.

You can classify network resources to a Security Group tag based on multiple attributes. Traffic between Security Groups can be controlled by Security Group Access Control Lists (SGACLs) also known as Security Contracts, which match source and destination Security Groups using Security Group tags.

Terminology

Security Group (SG)

A Security Group is a logical entity that contains a collection of physical or virtual network endpoints that are classified based on attributes or selectors.

Source Security Group Tags (S-SGT)

Tags derived from source attributes are called Source Security Group Tags.

Destination Security Group Tags (D-SGT)

Tags derived from destination attributes are called Destination Security Group Tags.

Security Group Access Control List (SGACL)

An SGACL uses Security Tags for enforcing specific security rules (L4 filters) between different Security Groups. The Tags are derived from IP, VLAN, and VM Attributes. SGACL allows to enforce security policies between SGs. An SGACL is also known as a Contract. In some parts of the document, SGACL is referred to as contract.

VRF Level Enforcement

The security group selectors define which endpoints and external IPs belong to the Security Group. Security Groups can contain endpoints, which are part of different VRFs. If endpoints part of different VRFs are associated to the same SG, communication between them would be possible only after applying the required VRF route-leaking configuration.

By default, a newly defined Tenant VRF has policy enforcement set to Unenforced. This means that even if classification criteria and SGACLs between secure groups were to be provisioned, no policy enforcement would be possible. To enable SGACL enforcement in the VRF, the VRF needs to be explicitly configured in **Enforced** mode.

When you configure the VRF in enforced mode, you can define the default behavior to be either of the following:

- **Deny:** All unicast traffic flows are dropped unless permitted by an Allowlist.
- **Permit:** All unicast traffic flows are allowed unless denied by a Denylist.

Hosts within a SG can communicate freely without explicit SGACLs. SGACLs create security rules, only.

Guidelines and Limitations

GPO has the following guidelines and limitations:

- All supported N9000 platforms require a minimum system memory of 24GB to support this feature.
- Beginning with Cisco NX-OS Release 10.4(3)F, GPO is supported on the following platforms:
  - 9300-FX3
  - 9300-GX
  - 9300-GX2
- Beginning with Cisco NX-OS Release 10.5(1)F, GPO is supported on Cisco Nexus 9408 switch.
- Beginning with Cisco NX-OS Release 10.5(1)F, GPO supports VLAN range match.
- Beginning with Cisco NX-OS Release 10.5(1)F, analytics support for GPO packets is added.
- Beginning with Cisco NX-OS Release 10.5(1)F, service-redirection with GPO is supported.
- Beginning with Cisco NX-OS Release 10.5(2)F, GPO is supported on the following platforms:
  - 9300-FX
  - 9300-FX2
  - 9300-HX
- SGACLs are supported only in the context of a VXLAN EVPN deployment. SGACL cannot be deployed on non-VXLAN enabled VRFs.
- SGACLs are not applicable to BUM and multicast traffic. System generated default permit policies exist for BUM and multicast traffic.

- You cannot configure VLAN-based Security Group selectors with a VLAN part of **system reserved-vlan-range** values.
- Coexistence of PVLAN and SGACL configurations is not supported.
- If VLAN-based Security Group selectors are already configured, system-reserved-vlan-range cannot be modified to include VLAN values used in the SG selectors.
- GPO is not supported on a site that has policy-aware and policy-unaware nodes. GPO is supported on policy-aware sites or a mix of policy-aware and policy-unaware sites.
- Unicast Reverse Path Forwarding (URPF) is not supported with routing profile template-security-groups (**system routing template-security-groups**).
- For the adjacencies learned on Layer 3 connected interfaces, if the matching ESG selector (connected endpoints) is configured on the switch, the host routes are installed with admin distance of 205.
- Beginning with Cisco NX-OS Release 10.5(3)F, users can enforce SGACL on an EVPN-VXLAN fabric for traffic flows to and from external endpoints reachable through an SR-MPLS based Layer-3 VPN.
  - ACL Permit logging action for the SGACL Policy shall not work for MPLS terminated flows.
  - On the BGW or BL, the SGACL based service redirection is not supported for flows destined towards MPLS Core
  - This feature is only supported on Virtual MCT BL and is not supported on Physical MCT BGW/BL.
  - SR-MPLS L3VPN interworking is supported on BLs and Anycast BGWs.
  - When this feature is enabled on Cisco Nexus 9300 FX2 platform switches, MPLS-SR LSR functionality is not supported.
  - Non-disruptive ISSU for MPLS features is only supported with BGP labelled unicast in the MPLS core. This limitation will be applicable for this feature also.
  - Beginning with Cisco NX-OS Release 10.5(3)F, users can select a combination of external-subnet IP prefix and next-hop IP address to create groups and enforce SGACL policy. Users can also use next-hop IP address with encapsulation type option to specify the next-hop address used in the VPN address family. In this release, this feature is supported only on BGP and static routes.
    - Beginning with Cisco NX-OS Release 10.6(1)F, support for this feature is extended on OSPF.
    - The next-hop qualifier in the selector can be a recursive next-hop or a connected next-hop for the protocol or client.
    - If the user specifies IP along with next-hop selector match for only a subset of expected next-hops of a route but not all, then
      - The ESG lookup with IP and next-hop for the other next-hops may yield a less specific tag (say based on IP). That tag is installed for such next-hops.
      - Details on the SGT returned to the protocol learned prefixes are as follows:

        Unicast Longest prefix match (LPM) first, then the next-hop match. If LPM matches, but the next-hop List doesn't match, then the protocol will look for the next better LPM (lower mask length).

        5

Unordered List bullet 5 If an IP selector does not have any next-hop specified, then it would be treated as a match against any next-hop. Following is an example.

Unordered List bullet 6 For the below config:

```
SGT:100 -> 10.1.1.0/24+NH:1.1.1.1
SGT:300 -> 10.1.1.0/24+NH:1.1.1.2

  SGT:500 -> 10.1.0.0/16
```

Unordered List bullet 6 When the protocol learns 10.1.1.0/25 with NH:1.1.1.1, then the traffic against this prefix will get SGT 100.

Unordered List bullet 6 When the protocol learns 10.1.1.0/20 with NH:1.1.1.1, then the traffic against this prefix will get SGT 500.

- It would be a misconfiguration to have mismatch of SGTs among next-hop of a route when the paths may become ECMP.
  - In such a case, traffic could get dropped if the tag selected by RIB/FIB for installation if from a NH with tag 0, or not forwarded/processed correctly.
- The next-hop is expected to be of the same address family as the route prefix.
  - For the VPN case, the next-hop could be of IPv4 address family, even if the route's prefix is of IPv6 address family.
- The VRF of next-hop in the IP with next-hop selector should be in the VRF configured for the IP.
- For IPv6, next-hops can be link-local (since they will be unique across different neighbor devices).
- Unlike IP selector, route-Injection for IP with next-hop selector is not supported from Cisco NX-OS Release 10.5(3)F.

Note Route-Injection for IP selector was introduced from 10.5(2)F

- Beginning with Cisco NX-OS Release 10.5(3)F, users can select VLAN along with MAC address to create a security group. VLAN and VLAN along with MAC address selector both can coexist in the same security group.
- Beginning with Cisco NX-OS Release 10.5(3)F, support for ISSU is added for all the SGACL features along with service redirection for multi-node and multisite fabrics.

- Beginning with Cisco NX-OS Release 10.6(1)F, mac based micro-segmentation is supported with L2 VLANs with SVI and without SVI.

- Mac based micro-segmentation is a fabric wide feature. All nodes should have this feature enabled or disabled.
- When mac based micro-segmentationis enabled, IP based micro-segmentation is disables by default.
- For mac based micro-segmentation, associate-vrf or SVI interface is needed, without which contracts will not be applied.

- Beginning with Cisco NX-OS Release 10.6(1)F, users can select combination of VLAN and an interface to create a security group.
  - Port and VLAN mappings are supported only on trunk ports (physical and PO).
  - This feature is not supported on ports connected to service nodes.
  - All VLAN selectors type can exist under a security-group.
  - For a given VLAN, if different VLAN selector types are configured against different security-groups, then the preference order is:
    1. Combination of MAC and VLAN then
    2. Combination of Interface and VLAN then
    3. VLAN selector

    Further, if there is a match against both IP and VLAN selector, then the security group against the IP selector is preferred.

Configuring Micro-segmentation using GPO

Enabling GPO

Perform the following steps to enable the micro-segmentation feature. The first time you enable the feature, the routing template should be configured to **system routing template-security-groups**.

Warning

- This routing template is required for **feature security-group**. Ensure the feature is enabled after applying the template mode.
- This routing template requires extended SSD re-partitioning. This can be achieved by executing the **copy running-config startup-config** and **system flash sda resize extended** commands.

Note We recommend backing up contents within bootflash, logflash, and running configuration prior to proceeding. For more information see Cisco Nexus 9000 Series NX-OS Fundamentals Configuration Guide, Release 10.4(x).

Subsequent disabling and re-enabling of the **feature security-group** can be done without requiring a switchreload.

SUMMARY STEPS

1. **configure terminal**
2. **system routing template-security-groups**
3. **copy running-config startup-config**
4. **system flash sda resize extended**
5. **[no] feature security-group**
6. **show nve peers detail**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**system routing template-security-groups** **Example:** `switch(config-if)# system routing template-security-groups`	Changes the switch routing profile. **Note** The routing template should be configured to **system routing template-security-groups**. Routing template requires extended SSD partitioning executed through **system flash sda resize** which will initiate a reload.
Step 3	**copy running-config startup-config** **Example:** `switch(config-if)# copy running-config startup-config`	Copies the running configuration to the startup configuration.

Step 4	**system flash sda resize extended**
	Example:
	`switch(config-if)# system flash sda resize extended`
	`!!!! WARNING !!!!`
	
	`Attempts will be made to preserve drive contents during`
	`the resize operation, but risk of data loss does exist.`
	`Backing up of bootflash, logflash, and running configuration`
	`is recommended prior to proceeding.`
	
	`!!!! WARNING !!!!`
		Increases storage space.

	Command or Action	Purpose
	current scheme is sda 8:0 0 119.2G 0 disk |-sda1 8:1 0 512M 0 part |-sda2 8:2 0 32M 0 part /mnt/plog |-sda3 8:3 0 128M 0 part /mnt/pss |-sda4 8:4 0 110.5G 0 part /bootflash |-sda5 8:5 0 64M 0 part /mnt/cfg/0 |-sda6 8:6 0 64M 0 part /mnt/cfg/1 `-sda7 8:7 0 8G 0 part /logflash target scheme is sda 8:0 0 120GB|250GB 0 disk |-sda1 8:1 0 512M 0 part |-sda2 8:2 0 32M 0 part /mnt/plog |-sda3 8:3 0 128M 0 part /mnt/pss |-sda4 8:4 0 rem 0 part /bootflash |-sda5 8:5 0 1.0G 0 part /mnt/cfg/0 |-sda6 8:6 0 1.0G 0 part /mnt/cfg/1 |_sda7 8:7 0 39G 0 part /logflash Continue? (y/n) [n] y A module reload is required for the resize operation to proceed Please, do not power off the module during this process.	
Step 5	**[no] feature security-group** **Example:** switch(config-if)# **feature security-group**	Enables the group policy option (GPO) feature. Use the 'no' prefix to disable the feature. The GPO feature can be disabled or enabled in runtime.
Step 6	**show nve peers detail** **Example:** switch(config-if)# **show nve peers detail** Details of nve Peers: ---------- Peer-Ip: 1.1.1.1 NVE Interface : nve1 Peer State : Up Peer Uptime : 1d12h Router-Mac : 5292.ca60.1b08 Peer First VNI : 101 Time since Create : 1d12h Configured VNIs : 100-101,200-201 Provision State : peer-add-complete Learnt CP VNIs : 100-101,200-201 vni assignment mode : SYMMETRIC Peer Location : FABRIC Group policy option : yes ----------	Verifies that the group policy option is enabled for peer device.

Creating a Security Group

Perform the following steps to create or update a Security Group and to configure member selection criteria. To select group members, you can specify any combination of the following attributes:

- IPv4 address or subnet for connected-endpoints and external-subnets.
- IPv6 address or subnet for connected-endpoints and external-subnets.
- Match VLAN at the switch level.
- External-subnets and next-hop IP addresses. Along with this, users can also use encapsulation type for next-hop to specify the next-hop address used in the VPN address family.
- VLAN along with MAC address.

SUMMARY STEPS

1. **configure terminal**
2. **security-group***sg-id* **name***sg-name*
3. **[no] match {connected-endpoints | external-subnets} vrf***vrf-name***[ipv4|ipv6]** *ip-prefix*
4. **[no] match external-subnets vrf** *name* **{ipv4 | ipv6}** *prefix* **nexthop ipv4** *ip-address* **encap-type mpls-vpn**
5. **[no] match external-subnets vrf** *name* **{ipv4 | ipv6}** *prefix* **nexthop {ipv4 | ipv6}** *ip-address*
6. **[no] match vlan** *vlan-id* **mac** *mac-addr*
7. **[no] match vlan** *vlan-id* **interface [port-channel | ethernet]**
8. **[no] match vlan** *vlan-id* **[mac** *mac-addr***]**
9. **[no] match vlan** *vlan-id* **interface [port-channel | ethernet]**
10. **[no] match vlan** *<list of vlan-ranges>*
11. **type layer4-7**
12. **match interface** *interface*
13. **show security-group id** *sg-id*

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# configure terminal`
	`switch(config)#`
		Enters global configuration mode.

Step 2	**security-group***sg-id* **name***sg-name*
	Example:
	`switch(config)# security-group 100 name webservers`
	`switch(config-security-group)#`
		Creates (or selects an existing) security group whose unique ID is *sg-id* and whose name is *sg-name*.

	Command or Action	Purpose

Step 3	**[no] match {connected-endpoints | external-subnets} vrf***vrf-name***[ipv4|ipv6]** *ip-prefix*
	Example:
	`switch(config-security-group)# match connected-endpoints vrf vrf_blue ipv4 61.1.1.141/32`
	`switch(config-security-group)# match external-subnets vrf vrf_blue ipv4 10.0.0.0/8`
	`switch(config-security-group)# match connected-endpoints vrf vrf_blue ipv6 61:1:1:2:1::141/128`
	`switch(config-security-group)# match external-subnets vrf vrf_blue ipv6 10:11:12:13::/64`
		This command is an IPv4-VRF or IPv6-VRF selector for a host (connected-endpoints) or external (external-subnets) resource.
		Use the '**no**' prefix to disable the specific classification.

Step 4	**[no] match external-subnets vrf** *name* **{ipv4 | ipv6}** *prefix* **nexthop ipv4** *ip-address* **encap-type mpls-vpn**
	Example:
	`switch(config-security-group)# match external-subnets vrf vrf_blue ipv4 10.0.0.0/8 nexthop ipv4 100.0.1.1 encap-type mpls-vpn`
	`switch(config-security-group)# match external-subnets vrf tenant-5 ipv6 78:1:1:1::/64 nexthop ipv4 100.0.1.1 encap-type mpls-vpn`
		Configures VRF selector using external subnet, next-hop IP address, and encapsulation type. This step is applicable for route learned against SR-MPLS.

Step 5	**[no] match external-subnets vrf** *name* **{ipv4 | ipv6}** *prefix* **nexthop {ipv4 | ipv6}** *ip-address* **Example:** `switch(config-security-group)# match external-subnets vrf vrf_blue ipv4 10.0.0.0/8 nexthop ipv4 100.0.1.1`	Configures VRF selector using external subnet and next-hop IP address.

Step 6	**[no] match vlan** *vlan-id* **mac** *mac-addr*
	Example:
	`switch(config-security-group)# match vlan 10`
	`switch(config-security-group)# match vlan 10 mac 001b.6384.45e6`
		Configures VLAN selector at the switch level. Users can select VLAN along with MAC address to create a security group. VLAN and VLAN along with MAC address selector both can coexist in the same security group.

Step 7	**[no] match vlan** *vlan-id* **interface [port-channel | ethernet]**
	Example:
	`switch(config-security-group)# match vlan 400 interface port-channel1`
	`switch(config-security-group)# match vlan 400 interface Ethernet1/1`
		Configures VLAN with port selector. Users can select port channel or ethernet.

Step 8	**[no] match vlan** *vlan-id* **[mac** *mac-addr***]** **Example:**	Configures VLAN selector at the switch level. Users can select VLAN along with MAC address to create a security group. VLAN and VLAN along with MAC

	Command or Action	Purpose
	`switch(config-security-group)# match vlan 10` `switch(config-security-group)# match vlan 10 mac 001b.6384.45e6`	address selector both can coexist in the same security group.

Step 9	**[no] match vlan** *vlan-id* **interface [port-channel | ethernet]**
	Example:
	`switch(config-security-group)# match vlan 400 interface port-channel1`
	`switch(config-security-group)# match vlan 400 interface Ethernet1/1`
		Configures VLAN with port selector. Users can select port channel or ethernet channel.

Step 10	**[no] match vlan** *<list of vlan-ranges>* **Example:** `switch(config-security-group)# match vlan 10-20, 30, 35, 40-45`	Configures VLAN with range selector. **Note** VLAN based ESG selectors cannot be configured with a VLAN inside system-reserved-vlan-range values. If vlan-based ESG selectors are already configured, system-reserved-vlan-range cannot be changed to VLAN values used in the ESG selectors.
Step 11	**type layer4-7** **Example:** `switch(config-security-group)# type layer4-7`	Creates special ESG for service attachment. Any selector type can be configured under this security-group. Layer4-7 attribute applies only to interface selector, results in hardware programming to tag all incoming traffic on the interface. Configuration of layer4-7 type for an ESG need be consistent on all the leaf. Configuration of same layer4-7 security-group without layer4-7 type on a different leaf causes misconfiguration. **Note** This is applicable for service redirection. See Cisco Nexus 9000 Series NX-OS ePBR Configuration Guide, Release 10.5(x), chapter Configuring ePBR for Micro-segmentation.
Step 12	**match interface** *interface* **Example:** `switch(config-security-group)# match interface vlan10`	Configures interface. In Release 10.5(1)F, only SVI interface is supported as a service interface. Type layer4-7 attribute applies only to interface selector. In Release 10.5(1)F, match interface <> without type layer4-7 attribute configured under the security-group is not supported. **Note** This is applicable for service redirection. See *Cisco Nexus 9000 Series NX-OS ePBR Configuration Guide, Release 10.5(x)*, chapter *Configuring ePBR for Micro-segmentation*.
Step 13	**show security-group id** *sg-id* **Example:**	Verifies the group policy selectors.

	Command or Action	Purpose

	switch(config-if)# show security-group id all
	Security Group ID 100 , Name webservers, Type Layer4-7 Service
	  Selector Type : External IPv4 Subnets
	  VRF-Name
	IPv4-Address/mask-len
	  vrf_blue 10.0.0.0/8
	
	  Selector Type : Connected IPv4 Endpoints
	  VRF-Name
	IPv4-Address/mask-len
	  vrf_blue
	61.1.1.141/32
	  Selector Type : External IPv6 Subnets
	  VRF-Name
	IPv6-Address/mask-len
	  vrf_blue
	10:11:12:13::/64
	  Selector Type : Connected IPv6 Endpoints
	  VRF-Name
	IPv6-Address/mask-len
	  vrf_blue
	61:1:1:2:1::141/128
	  Selector Type : Vlan
	  VLAN-Id
	  Vlan10
	  Vlan11
	  Vlan12
	  Vlan13
	  Vlan14
	  Vlan15
	  Vlan16
	  Vlan17
	  Vlan18
	  Vlan19
	  Vlan20
	  Vlan30
	  Vlan35
	  Vlan40
	  Vlan41
	  Vlan42
	  Vlan43
	  Vlan44
	  Vlan45
	  Selector Type : Interface
	  Interface
	  Vlan10	

Creating Security Group using MAC

Execute the following system level CLI to create a security group using MAC.

SUMMARY STEPS

1. **[no] system security-group mac-segmentation**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**[no] system security-group mac-segmentation**	Enables mac-segmentation Use **no** form of this command to disable mac-segmentation.

Creating a Security Class-Map

A Class-Map classifies network traffic based on various match criteria configured within a class map. Perform the following steps to create a Security Class-Map to define the filters identifying specific traffic flows.

SUMMARY STEPS

1. **configure terminal**
2. **class-map type security match-any***web-class*
3. **match [default | ip | ipv4 | ipv6]**

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# configure terminal`
	`switch(config)#`
		Enters global configuration mode.

Step 2	**class-map type security match-any***web-class* **Example:** `switch(config)# class-map type security match-any web-class2`	Create a security class-map to identify specific traffic flows.

Step 3	**match [default | ip | ipv4 | ipv6]**
	Example:
	`switch(config-cmap-sec)# match ipv4 udp sport 399 to 402 dport 400 to 403`
	`switch(config-cmap-sec)# match ipv6 udp sport 399 to 402 dport 400 to 403`
		Configures the security class by matching based on traffic type.

Creating a Security Policy-Map

A policy map defines a policy stating what happens to traffic that is classified using class maps and ACLs. Perform the following steps to create a Security Policy-Map to define the action (permit, deny, log) traffic flows identified by the previously created security class-map.

SUMMARY STEPS

1. **configure terminal**
2. **policy-map type security***policy-map*
3. **class web-class**
4. **[no] [permit | deny | log]**
5. **[no] service-chain** <*svc-chain*>

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# configure terminal`
	`switch(config)#`
		Enters global configuration mode.

Step 2	**policy-map type security***policy-map*	Creates a security policy-map.
Step 3	**class web-class**	Specifies a security class-map to be associated with the policy-map to define the traffic the rule will be applied to.
Step 4	**[no] [permit | deny | log]**	Defines action to be taken on matching traffic. • **Deny:** Deny the matching traffic. • **Log:** Log the matching traffic. • **Permit:** Permit the matching traffic. **Permit** is the default action if none specified. **Log** action can be set with permit or deny. Matching traffic is logged under **show logging ip access-list cache [detail]**.
Step 5	**[no] service-chain** <*svc-chain*>	Configure service-chain for match class-maps to enable service redirection for the contract. SGACL with service-chain can only be enabled with permit action. This is applicable for service redirection. For more details about service redirection, see Cisco Nexus 9000 Series NX-OS ePBR Configuration Guide.

Configuring Security contracts between Security Groups

This procedure creates an SGACL (contract) to enforce a security policy between Security Groups.

Before you begin

- Creating a Security Group
- Creating a Security Class-Map
- Creating a Security Policy-Map

SUMMARY STEPS

1. **configure terminal**
2. **vrf context***vrf-name*
3. **security enforce tag***sg-id* **default [permit | deny]**
4. **security contract source [***sg-id* | *any***] destination [***sg-id* | *any***] policy policy-map-name** [*bidir* | *unidir*]

DETAILED STEPS

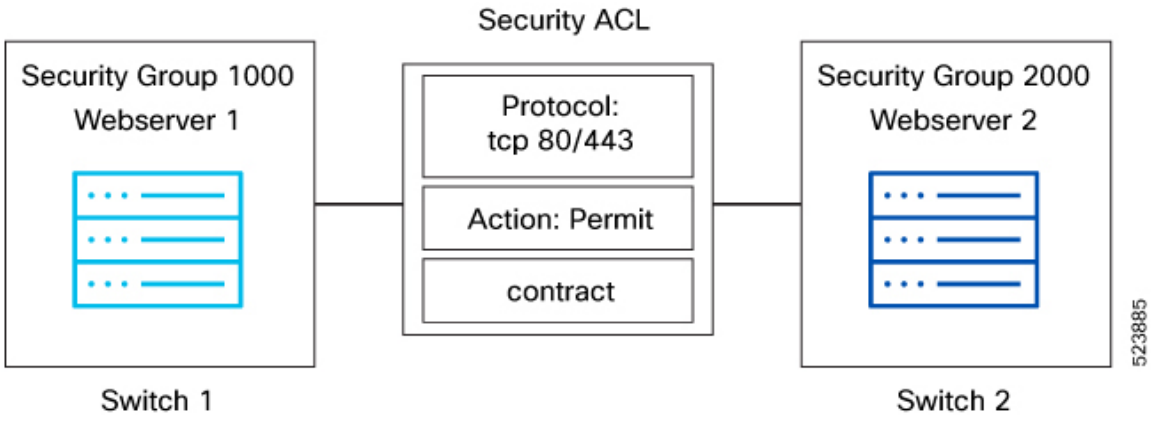
Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**vrf context***vrf-name* **Example:** `switch(config)# vrf context blue`	Enters configuration mode for the specified VRF.
Step 3	**security enforce tag***sg-id* **default [permit | deny]** **Example:** `switch(config-vrf)# security enforce tag 100 default deny`	Moves the VRF to enforced mode. • **sg-id:** defines the security group tag for the tenant VRF on which the default options are added. • **default deny:** Denies all traffic within the VRF without explicit security contracts. • **default permit:** Allows all traffic within the VRF without explicit security contracts.
Step 4	**security contract source [***sg-id* | *any***] destination [***sg-id* | *any***] policy policy-map-name** [*bidir* | *unidir*] **Example:** `switch(config-vrf)# security contract source 100 destination 200 policy policyMap1 bidir`	Applies the previously defined security policy map, with the corresponding action, between the specified security groups. The default option is **bidir** if no direction is specified. Default option **bidir** applies the SGACL to traffic in both direction (in example, 100 to 200 and 200 to 100).

	Command or Action	Purpose
		For example, if you create a security rule between SG 100 and SG 200 with a filter that specifies destination port 80, the use of **bidir** ensures that a rule is also applied for communication between SG 200 and SG 100 with source port 80 so that the two-ways communication can be successfully established.

Configuration Examples for GPO

Figure 50: Creating Security Group



Step 1: Enabling GPO.

```
Switch1# configure terminal
Switch1(config)# system routing template-security-groups
Switch1(config)#feature security-group

Switch2# configure terminal
Switch2(config)# system routing template-security-groups
Switch2(config)#feature security-group
```

Step 2: Creating a security class-map to identify specific traffic flows.

```
Switch1(config)#class-map type security match-any web-class
match ipv4 tcp dport 443
match ipv4 tcp dport 80

Switch2(config)#class-map type security match-any web-class
match ipv4 tcp dport 443
match ipv4 tcp dport 80
```

Step 3: Creating a security policy map

```
Switch1(config)#policy-map type security policyMap1
class web-class
permit
Switch2(config)#policy-map type security policyMap1
class web-class
permit
```

Step 4: Creating security group.

```
switch1(config)security-group 1000 name webserver1
switch1(config-security-group)# match connected-endpoints vrf vrf_blue ipv4 61.1.1.141/32
switch1(config-security-group)# match external-subnets vrf vrf_blue ipv4 10.0.0.0/8
switch1(config-security-group)# match connected-endpoints vrf vrf_blue ipv6
61:1:1:2:1::141/128
switch1(config-security-group)# match external-subnets vrf vrf_blue ipv6 10:11:12:13::/64
switch1(config-security-group)# match connected-endpoints vrf vrf_red ipv4 100.5.150.125/32

switch1(config-security-group)# match connected-endpoints vrf vrf_red ipv6
100:1:1:495::125/128
switch1(config-security-group)# match external-subnets vrf vrf_red ipv4 11.0.0.0/8
switch1(config-security-group)# match vlan 10

switch2(config)security-group 2000 name webserver2
switch2(config-security-group)# match connected-endpoints vrf vrf_blue ipv4 61.1.1.142/32
switch2(config-security-group)# match external-subnets vrf vrf_blue ipv4 20.0.0.0/8
switch2(config-security-group)# match connected-endpoints vrf vrf_blue ipv6
61:1:1:2:1::142/128
switch2(config-security-group)# match external-subnets vrf vrf_blue ipv6 20:11:12:14::/64
switch2(config-security-group)# match connected-endpoints vrf vrf_red ipv4 100.5.150.126/32

switch2(config-security-group)# match connected-endpoints vrf vrf_red ipv6
100:1:1:495::126/128
switch2(config-security-group)# match external-subnets vrf vrf_red ipv4 21.0.0.0/8
switch2(config-security-group)# match vlan 10
```

Step 5: Moving VRF to enforce mode

```
switch1(config)# vrf context vrf_blue
switch1(config-vrf)# security enforce tag 100 default deny
switch2(config)# vrf context vrf_red
switch2(config-vrf)# security enforce tag 101 default deny
```

Step 6: Apply contract

```
switch1(config-vrf-blue)# security contract source 1000 destination 2000 policy policyMap1

switch1(config-vrf-red)# security contract source 1000 destination 2000 policy policyMap1


switch2(config-vrf-blue)# security contract source 1000 destination 2000 policy policyMap1

switch2(config-vrf-red)# security contract source 1000 destination 2000 policy policyMap1
```

Verifying GPO

Following are the show commands associated with GPO configuration:

show contracts

Displays all the contracts applied in the switch for all the vrfs.

```
switch(config)# show contracts

VRF             SGT   DGT   Policy           Dir    Stats        Class            Action
      OperSt
-------------------------------------------------------------------------------------------------
vrf_blue            1000   2000 policyMap1        bidir  350370       web-class
permit,log   enabled
vrf_red             1000   2000 policyMap1        bidir  373270       web-class
permit,log   enabled
```

show run security-group

Displays all the security-group related configurations in the switch.

```
switch1(config)# show run security-group
!Command: show running-config security-group
!Running configuration last done at: Fri Dec  8 12:23:52 2023
!Time: Fri Dec  8 12:27:09 2023
version 10.4(2) Bios:version 05.50
feature security-group
security-group 1000 name webserver1
match connected-endpoints vrf vrf_blue ipv4 61.1.1.141/32
match external-subnets vrf vrf_blue ipv4 10.0.0.0/8
match connected-endpoints vrf vrf_blue ipv6 61:1:1:2:1::141/128
match external-subnets vrf vrf_blue ipv6 10:11:12:13::/64
match connected-endpoints vrf vrf_red ipv4 100.5.150.125/32
match connected-endpoints vrf vrf_red ipv6 100:1:1:495::125/128
match external-subnets vrf vrf_red ipv4 11.0.0.0/8
match vlan 10

class-map type security match-any web-class
  match ip udp
  match ip tcp

policy-map type security policyMap1
  class web-class

vrf context vrf_blue
  security contract source 1000 destination 2000 policy policyMap1
  security enforce tag 100 default deny

vrf context vrf_red
  security contract source 1000 destination 2000 policy policyMap1
  security enforce tag 101 default deny
```

show contracts detail

Displays all the contracts details applied in the switch includes class-map and policy-map details.

```
switch1(config)# show contracts detail

VRF: vrf_blue
  Contract source group any dest group 2000
    Policy: policyMap1 Direction: bidir
      Stats: 350370
      Class: web-class
        match ip udp
        match ip tcp
      Action: permit,log
      OperSt: enabled

VRF: vrf_red
  Contract source group any dest group 2000
    Policy: policyMap1 Direction: bidir
      Stats: 373270
      Class: web-class
        match ip udp
        match ip tcp
      Action: permit,log
      OperSt: enabled
```

show contracts policy policyMap1

Displays contracts based on policy name.

```
Switch1(config)show contracts policy policyMap1
VRF             SGT   DGT   Policy           Dir    Stats        Class            Action
      OperSt

-------------------------------------------------------------------------------------------------


vrf_blue            1000  2000   policyMap1       bidir  0            web-class
permit       enabled
vrf_red             1000  2000   policyMap1       bidir  0            web-class
permit       enabled
```

show contracts vrf vrf_blue

Displays contracts based on vrf.

```
switch1(config)show contracts vrf vrf_blue
VRF             SGT   DGT   Policy           Dir    Stats        Class            Action
      OperSt
-------------------------------------------------------------------------------------------------
vrf_blue         1000  2000   policyMap1       bidir  0            web-class       permit
      enabled
```

show contracts sgt 1000

Displays contract based on a given SGT.

```
switch1(config)# show contracts sgt 1000

VRF             SGT   DGT   Policy           Dir    Stats        Class            Action
      OperSt
-------------------------------------------------------------------------------------------------
vrf_blue     1000  2000  policyMap1       bidir  0           web-class          permit,log
   enabled
vrf_red      1000  2000  policyMap1       bidir  0           web-class          permit,log
   enabled
```

show contracts dgt 2000

Displays contract based on a given DGT.

```
switch1(config)# show contracts dgt 2000

VRF             SGT   DGT   Policy           Dir    Stats        Class            Action
      OperSt
-------------------------------------------------------------------------------------------------
vrf_blue     1000  2000  policyMap1       bidir  0           web-class          permit,log
   enabled
vrf_red      1000  2000  policyMap1       bidir  0           web-class          permit,log
   enabled
```

show contracts sgt 1000 dgt 2000

Displays contract based on a given SGT and DGT.

```
switch1(config)# show contracts sgt 1000 dgt 2000

VRF             SGT   DGT   Policy           Dir    Stats        Class            Action
      OperSt
-------------------------------------------------------------------------------------------------
vrf_blue     1000  2000  policyMap1       bidir  0           web-class          permit,log
   enabled
vrf_red      1000  2000  policyMap1       bidir  0           web-class          permit,log
   enabled
```

Use the Following show commands to see the contracts with service-redirection.

```
switch1(config)# show contracts policy ipv4tcp
VRF     SGT   DGT     Policy   Dir   Stats   Class    Action   OperSt
-------------------------------------------------------------------------------------------------
ixia    2004  3004   ipv4tcp   bidir   0    ipv4tcp   redir    enabled

switch1(config)# show contracts policy ipv4tcp detail
VRF: ixia
  Contract source group 2004 dest group 3004
   Policy: ipv4tcp Direction: bidir
     Stats: 0
     Class: ipv4tcp
       match ipv4 tcp
     Action: redir-1node2arm
     OperSt: enabled

switch1(config)# show contracts sgt 2004 dgt 3004
VRF    SGT    DGT    Policy    Dir    Stats   Class     Action   OperSt
-------------------------------------------------------------------------------------------------
ixia   2004   3004   ipv4tcp   bidir   0      ipv4tcp   redir    enabled
```

Checking Consistency of GPO

Beginning with Cisco NX-OS Release 10.5(1)F, use the following show command to check the consistency of contracts and polices.

Command or Action	Purpose

show consistency-checker contract *vrf < all | vrf-name > | rule < rule-id >*
 	
`switch1(config)# show consistency-checker contract vrf all`
`Contract CC: Starting for vrf= all`
`Contract CC: Checking if SGT feature is enabled`
`Contract CC: Checking DME & Policy-Mgr programming inconsistencies`
`Contract CC: Processing configuration data...`
`Contract CC: Starting validation for 15 rules`
`Contract CC: Fetching PT Tiles, TCAM entries from the Hardware`
`Contract CC: Retrieved Tile Entries 0 Tcam Entries 130`
`Contract CC: Examining SG-Agent and HW policy programming inconsistencies`
 	
`Contract CC: Final Status = SUCCESS`
	Displays the software and hardware programming status of contracts and polices.
	
	<all> - allows users to execute a consistency check for all contracts under all VRFs.
	
	<vrf-name> - allows users to execute a consistency check for all contracts under a specific VRF.
	
	<rule> - allows users to execute a consistency check for a specific rule by its identifier.

VXLAN Multi-Site and GPO Interoperability

Beginning with Cisco NX-OS Release 10.4(3)F, SG and SGACLs are supported on VXLAN EVPN fabrics part of the same Multi-Site domain. Policy aware and policy-unaware fabrics can be deployed as part of the same Multi-Site domain.

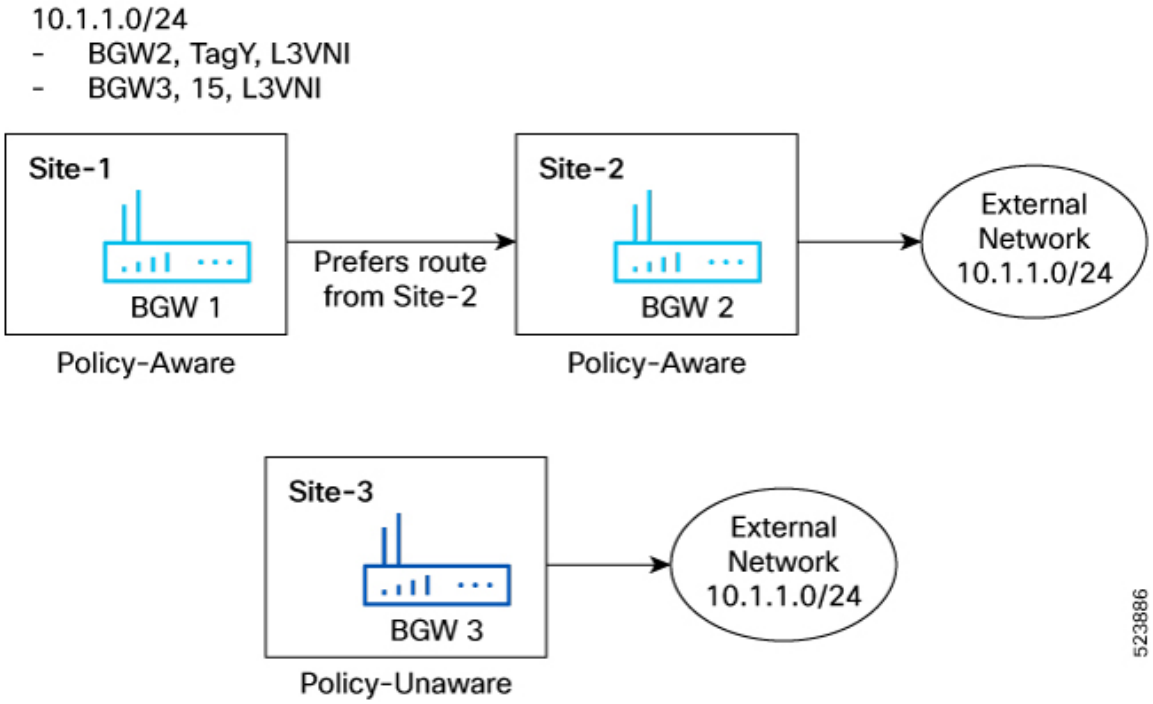
Anycast Border Gateways (BGWs) and vPC BGWs are supported with SGACL feature. Border-Gateway nodes in a policy-aware site can establish Multi-Site EVPN connectivity with policy-aware sites as well as

policy-unaware sites. SGACLs can be applied between SGs locally defined on separate fabrics. Additionally, it is possible to stretch a SG across multiple fabrics.

All the EVPN routes from policy-unaware sites are distributed into policy-aware sites and installed with a reserved security group tag value 15. To allow workload communication between policy unaware to policy aware sites, the user must create explicit contracts with tag 15 and the intended destination tag.

When a remote prefix is learned with multiple next-hops belonging to a mix of policy-unaware and policy-aware fabrics, the next-hop(s) of policy-aware fabrics are preferred. If multiple next-hops all belong to policy-unaware fabrics, then the received prefix route is installed and re-originated inside the policy-aware fabric with policy-unaware tag.

Figure 51: Policy-Aware and Policy-Unaware Tag Next-Hop Preference



In the figure, BGW 1 and BGW 2 are part of policy aware sites and BGW 3 belongs to a policy unaware site. External network 10.1.1.0/24 is advertised from Site 3 and Site 2 to Site 1. Since Site-2 is policy-aware, Route 10.1.1.0/24 is advertised with “TagY”, which is configured on Site2. Site-3 is policy-unaware, so the route advertised from BGW3 would not carry any policy information and would be locally assigned to have the default Policy-unaware tag which is ‘15’ when received by BGW1.

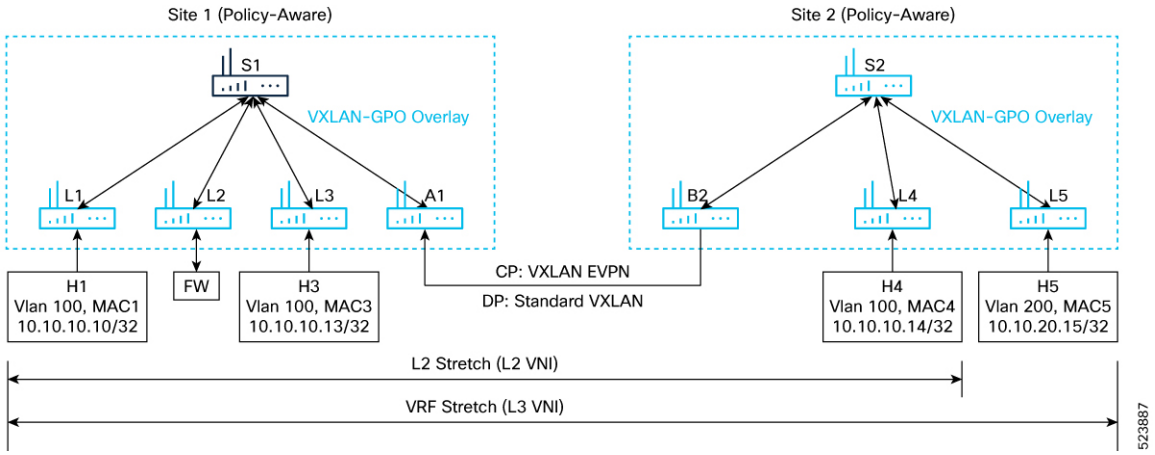
On BGW1, the Route would have overlay ECMP of BGW2 and BGW3 with tags of TagY and, ‘15’ respectively. However, since TagY is a valid tag from the policy-aware site, 10.1.1.0/24 is programmed with Tag ‘TagY’. Similarly, when BGW1 re-originates the route to a leaf in Site-1, it adds the tag ‘TagY’ to the route.

In a typical Anycast BGW setup, there are no SVI configurations for L2VNIs. However, to support GPO for L2 Bridged IP traffic, you need to configure an SVI for the L2VNIs on the Anycast BGWs. The SVI configuration does not need an IP Address or the “ip forward” command. The only requirement is to be configured under the Tenant VRF. This is needed to derive the Tenant VRF information for Host-IP lookup of the endpoints connected to the L2VNI segment and provide the SGT & DGT tags necessary to enforce the security policy.

Policy Aware Fabrics in a Multi-Site Domain

This section describes how the route advertisement and packet movement happens between 2 hosts in a policy aware Multi-Site domain.

Figure 52: Policy-Aware Multi-site



In the above topology, site 1 and site 2 are policy aware.

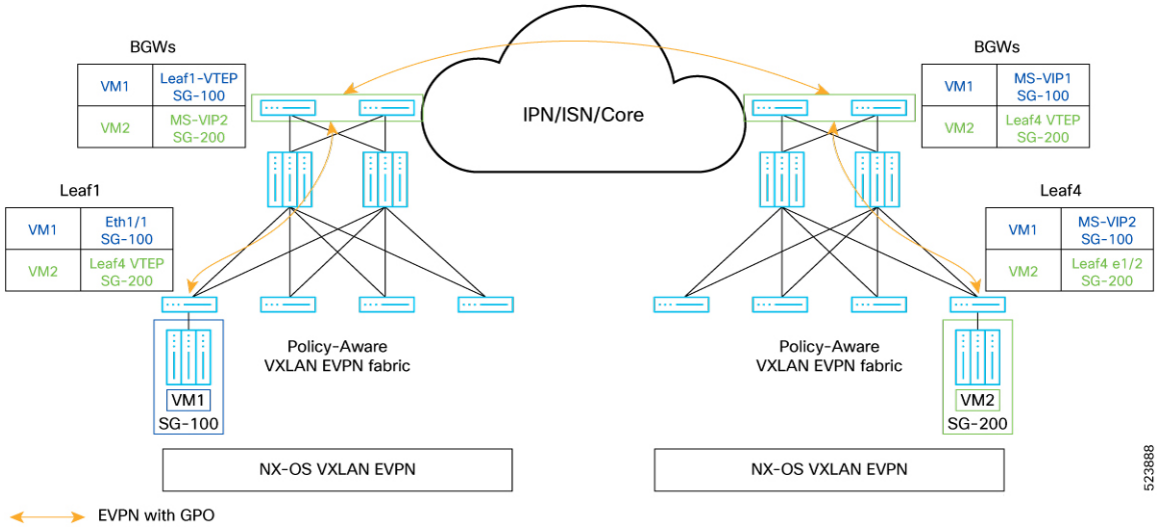
Route-advertisement of Host H1 from Leaf L1 to Leaf L5

When a host route H1 is advertised from L1, site 1 to L5 site 2, the route advertisement flow is as follows:

1. Next-hop Router L1 advertises the route with L1 next-hop and source group tag (TagX) as configured by the policy on router L1.
2. Router A1, which is the BGW in the Site-1, re-originates the route with the Multi-Site VIP of the BGW as next-hop, however, retains the SGT tag (TagX) received from L1.
3. Similarly, router B2 (BGW in Site-2) re-originates the route H1 inside the local fabric with the Multi-Site VIP of B2 BGW as next-hop and the same SGT tag (TagX).
4. L5 receives the route from router B2 and installs it in the forwarding table with the associated security tag “TagX”. This way, the tag from the originator leaf is retained across the entire multi-site domain.

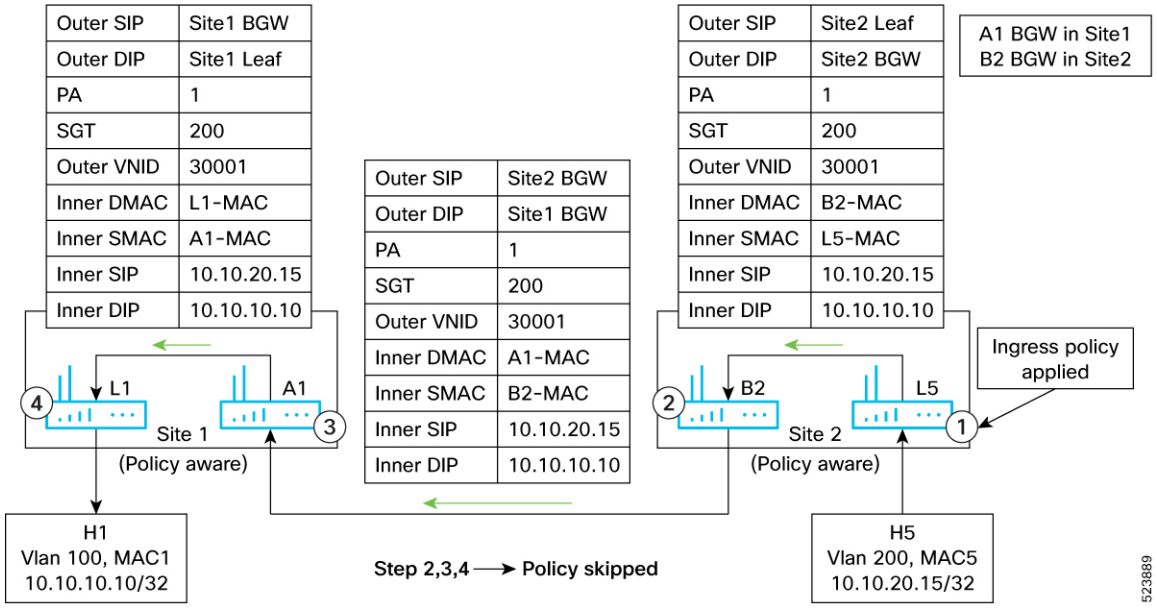
Refer to the following figure to know about the learned endpoint information on the various nodes.

Figure 53: Route Advertisement in a Policy Aware Multi-Site



Packet-Flow from Host H5 to Host H1

Figure 54: Packet-Flow from Host H5 to Host H1



The packet flow between the hosts H5 and H1 would be as follows.

1. Router L5 receives the traffic from the host H5. Since both SRC and DST tags are available, L5 would locally apply the security policy. If the policy-action is permitted, it would set the policy-applied (PA) bit in the VXLAN-GPO header and send the traffic to BGW router B2, representing the next-hop to reach the destination host H1.

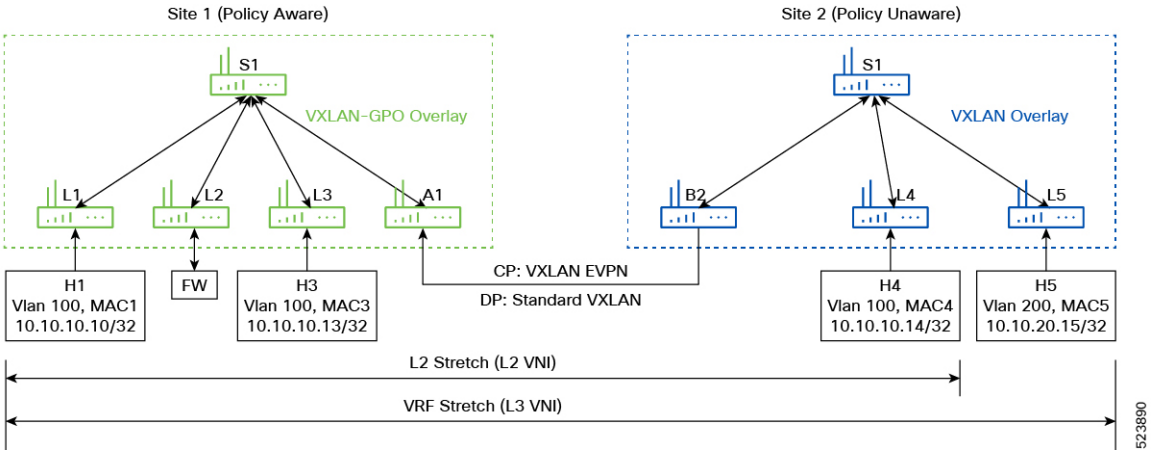
2. Along with this, the PA bit setting is retained in the VXLAN GPO header B2, on seeing the PA bit set in the VXLAN header, would not reapply the policy, and just decapsulate and re-encapsulate the traffic

sending it to the BGW A1, site 1, representing the next-hop to reach the destination host H1. Along with this, the PA bit setting is retained in the VXLAN GPO header.

3. BGW A1 would take a similar action as BGW B2 and forward the traffic to router L1.
4. Router L1 would not apply Policy as well because of the PA bit set and would forward the traffic to the destination host H1.

Policy Aware and Policy Unaware Fabrics in a Multi-Site Domain

Figure 55: Policy-Aware Policy Unaware Multi-Site



In the above picture, Site-1 is a policy aware site and Site-2 is a policy unaware site.

Route advertisement of Host H1 from Router L1 to Router L5

The route advertisement of Host H1 from router L1 to router L5 would be as follows.

1. Router L1 advertises the route H1 with next-hop as L1 and SGT Tag (TagY) based on the configuration on L1.
2. Router A1 locally installs the route with next-hop L1 (and associated tag TagY) and re-originates the route to the remote BGW B2 with next-hop as A1 Multi-Site VIP and the same SGT Tag (TagY).
3. The BGW B2, upon receiving the route with SGT Tag (TagY), locally installs the route with A1 as next-hop and ignores the SGT Tag as it is policy unaware. It then re-originates the route to router L5 with next-hop as B2 Multi-Site VIP.
4. Router L5 installs the route H1.

Route advertisement of Host H5 from Router L5 to Router L1

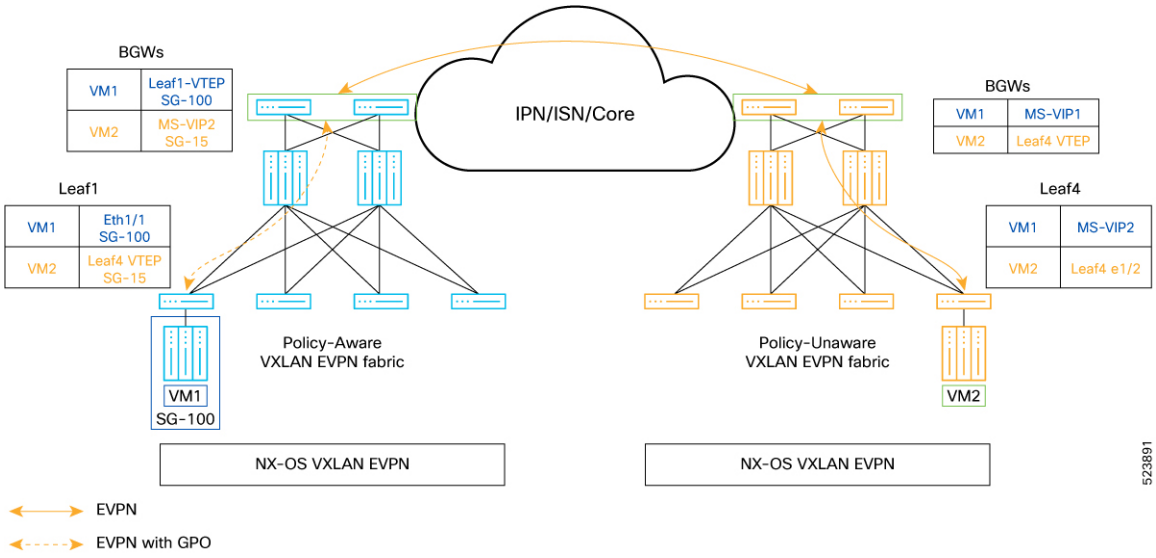
The route advertisement of Host H5 from router L5 to router L1 would be as follows.

1. Router L5 advertises the route for host H5 with next-hop as L5.
2. BGW B2 locally installs the route with L5 as next-hop and re-originates the prefix with next-hop as B2 Multi-Site VIP.
3. Since A1 is in a policy aware site, and received the route from a policy unaware fabric, A1 installs the route with the Default Tag for Policy unaware sites, Tag ‘15’.

4. BGW router A1 re-originates the route to L1 with next-hop as A1 Multi-Site VIP and Tag 15.
5. Router L1 installs the route with SGT TAG 15 and next-hop as A1 Multi-Site VIP.

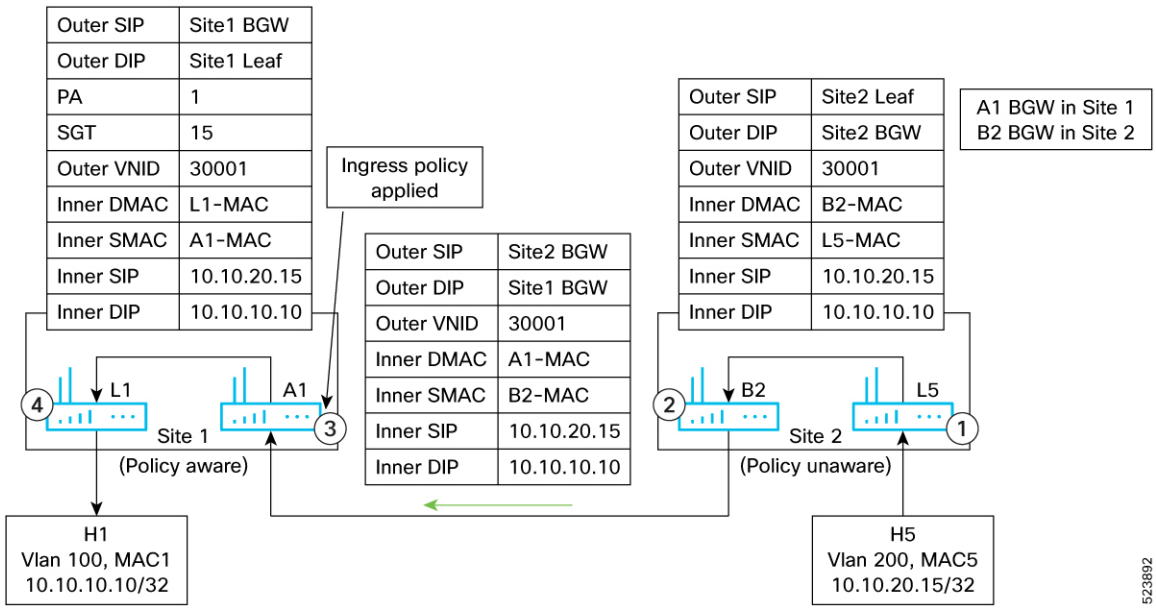
Refer to the following figure to know about the learned endpoint information on the various nodes.

Figure 56: Route Advertisement from a Policy Unaware to Aware Site



Packet flow from Host H5 to Host H1

Figure 57: Packet flow from Host H5 to Host H1



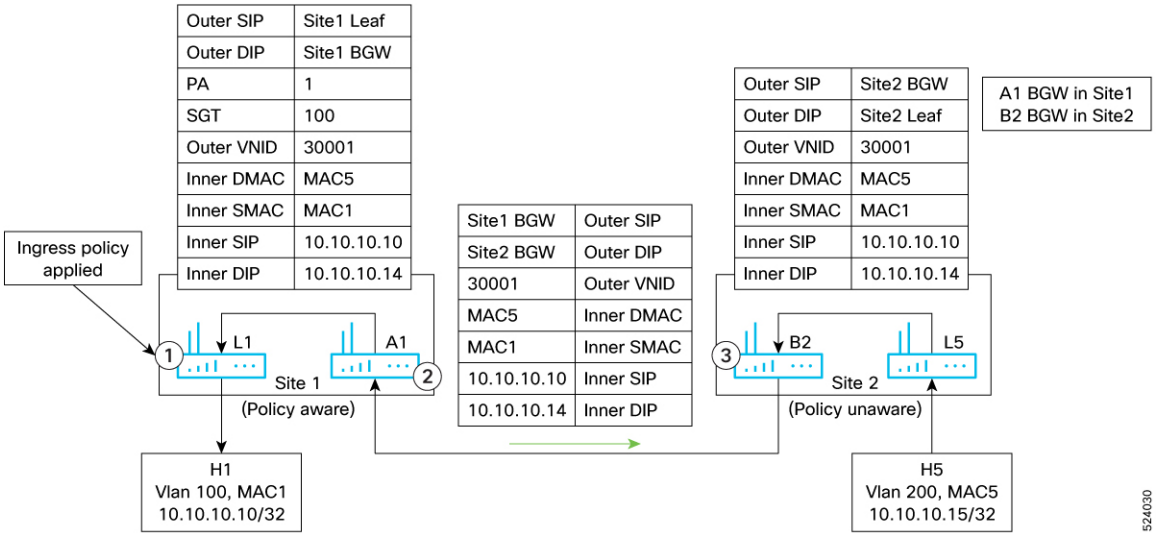
The packet flow between the hosts H5 and H1 would be as follows.

1. Router L5 receives the traffic from the host H5 and routes the traffic to B2 with standard VXLAN header.

2. B2 decapsulates and re-capsulates in a standard VXLAN header.
3. A1 decapsulates the traffic, applies the policy based on source Tag “15” and destination “Tag Y”. Based on the policy, traffic is forward to L1 with VXLAN GPO tunnel and policy-applied bit set.
4. L1 decapsulates the traffic and since PA bit is set, policy is not applied again and forwarded to H1.

Packet flow from Host H1 to Host H5

Figure 58: Packet flow from Host H1 to Host H5



Packet flow from host H1 to host H5 would be as follows.

1. Router L1 receives the traffic from H1, applies the policy based on the source Tag ‘TagY’, destination Tag ‘15’. Based on the policy result, L1 routes the traffic to A1 with VXLAN GPO Tunnel with policy-applied (PA) bit set.
2. A1 decapsulates and since PA bit is set, it does not reapply the policy and forwards the traffic to B2 in a standard VXLAN Tunnel.
3. Traffic flow from B2 to H5 is similar to any Multi-Site deployment.

With the help of micro-segmentation and GPO, NX-OS users can create smaller and isolated segments within a network and enforce security policies. This allows users to have better control over the traffic flow and apply security policies only where they are needed.

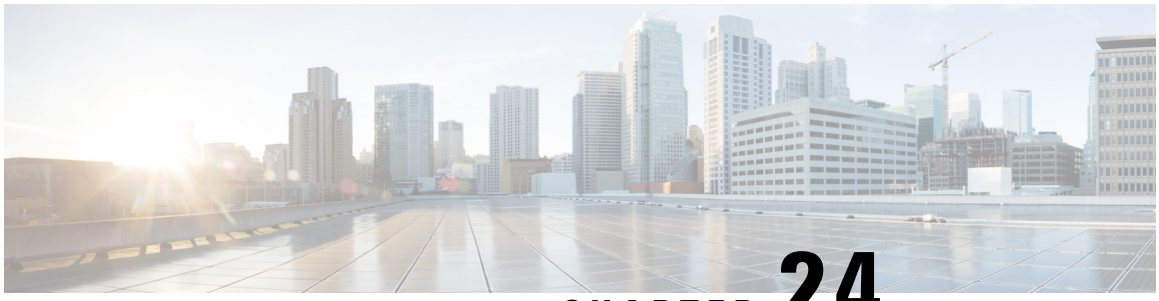
Related Documents

https://www.cisco.com/c/en/us/td/docs/dcn/nx-os/nexus9000/104x/configuration/scalability/cisco-nexus-9000-series-nx-os-verified-scalability-guide-1043.html

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/sw/nx-os/licensing/guide/cisco-nexus-nx-os-smart-licensing-using-policy-user-guide.html

https://www.cisco.com/c/dam/en/us/td/docs/Website/datacenter/platform/platform.html

https://developer.cisco.com/docs/cisco-nexus-3000-and-9000-series-nx-api-rest-sdk-user-guide-and-api-reference-release-10-4-x/



CHAPTER 24

Configuring Layer 4 - Layer 7 Services

This chapter contains these sections:

- About VXLAN Layer 4 - Layer 7 Services, on page 579
- Integrating Layer 3 Firewalls in VXLAN Fabrics, on page 579
- Firewall as Default Gateway, on page 591
- Transparent Firewall Insertion, on page 592
- Firewall Clustering with VXLAN BGP EVPN, on page 598
- Service Redirection in VXLAN EVPN Fabrics, on page 601

About VXLAN Layer 4 - Layer 7 Services

This chapter covers insertion of Layer 4 – Layer 7 services (firewall, load balancer, and so on) in a VXLAN fabric.

As opposed to traditional 3-tier network topologies, in which L4-L7 services are connected to the switches hosting the default gateway (aggregation/distribution), L4-L7 services in VXLAN fabrics are typically connected to the leaf or border switches, often referred to as *services leafs*.

You can attach a L4-L7 services device to a VXLAN fabric in various ways. This chapter addresses the considerations you must take depending on how the L4-L7 services device is attached and the requirements of the device and the network.

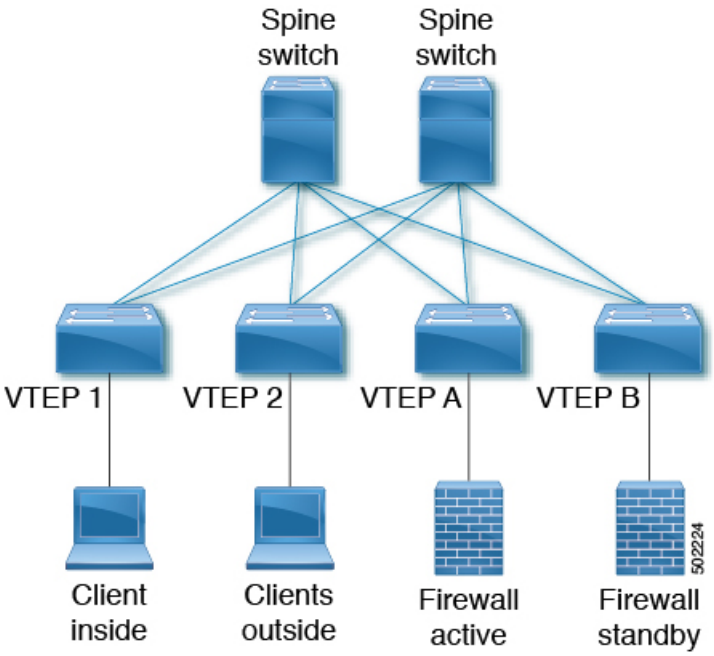
Integrating Layer 3 Firewalls in VXLAN Fabrics

This section provides details on how to integrate a firewall within a VXLAN EVPN fabric. A Layer-3 firewall involves separating different security zones.

When integrating a Layer-3 firewall in a VXLAN EVPN fabric with a distributed Anycast Gateway, each of these zones must correspond to a VRF/tenant on the fabric. The traffic within a tenant is routed by the fabric. Traffic between the tenants is routed by the firewall. This scenario often refers to an inter-tenant or tenant edge firewall.

Consider two zones: an inside zone and an outside zone. This scenario requires a VRF definition on the fabric. You can call the VRFs the inside VRF and the outside VRF. Traffic between subnets within the same VRF is routed on the VXLAN fabric using the distributed gateway. Traffic between VRFs is routed by the firewall where the rules are applied.

Figure 59: Topology Overview with Firewall Attachment



Single-Attached Firewall with Static Routing

If the firewall does not support running a routing protocol, you must have static routes on each VTEP pointing to the firewall as the next hop. The firewall also has static routes pointing to the Anycast Gateway IP as the next hop. The challenge with a static route is that the VTEP with an active firewall must be the one advertising the routes to the fabric. One way to accomplish this is to track the active firewall reachability via HMM and use this tracking to advertise routes into the fabric. When the active firewall is connected to VTEP A, VTEP A has a static route that tracks where the route is advertised if the firewall IP is learned as the HMM route. When the firewall fails and the standby firewall takes over, VTEP A learns the firewall IP using BGP, and VTEP B learns the firewall IP using HMM. VTEP A withdraws the route, and VTEP B advertises the route into the fabric. See the following example.

VTEP A and VTEP B:

```
Vlan 10
 Name inside
 Vn-segment 10010

Vlan 20
 Name outside
 Vn-segment 10020


Interface VLAN 10
 Description inside_vlan
 VRF member INSIDE
 IP address 10.1.1.254/24
 fabric forwarding mode anycast-gateway

Interface VLAN 20
 Description outside_vlan
 VRF member OUTSIDE
```

```
 IP address 20.1.1.254/24
 fabric forwarding mode anycast-gateway

interface nve1
 no shutdown
 host-reachability protocol bgp
 source-interface loopback1
 member vni 10010
  mcastgroup 239.1.1.1
 member vni 10020
  mcastgroup 239.1.1.1
 member vni 1001000 associate-vrf
 member vni 1002000 associate-vrf


track 10 ip route 10.1.1.1/32 reachability hmm
  vrf member INSIDE
!
VRF context INSIDE
 Vni 1001000
 IP route 20.1.1.0/24 10.1.1.1 track 10


track 20 ip route 20.1.1.1/32 reachability hmm
  vrf member OUTSIDE

!
VRF context OUTSIDE
 Vni 1001000
 IP route 10.1.1.0/24 20.1.1.1 track 20


VTEPA# show track 10 Track 10
IP Route 20.1.1.1/32 Reachability Reachability is UP

VTEPA# show ip route 20.1.1.0/24 vrf INSIDE
IP Route Table for VRF "INSIDE"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

20.1.1.0/24, ubest/mbest: 1/0
  *via 10.1.1.1 [1/0], 00:00:08, static

Firewall Failure on VTEP A caused the track to go down causing VTEP A to withdraw the static
 route.

VTEPA# show track 20 Track 20
IP Route 20.1.1.1/32 Reachability Reachability is DOWN

VTEPA# show ip route 20.1.1.0/24 vrf INSIDE
IP Route Table for VRF "RED"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

Route not found
```

Recursive Static Routes Distributed to the Rest of the Fabric

With this approach, the static routes are configured wherever the inside or outside VRF exists. As the next-hop is reachable through host routes (EVPN Route-Type2), the change of the active firewall to standby and vice versa is only seen locally and doesn't introduce any churn to the other VXLAN fabric. This approach can help to better scale and improve convergence.

Any VTEP:

```
VRF context OUTSIDE
 Vni 1002000
 IP route 10.1.1.0/24 20.1.1.1
! static route on VTEP pointing to Firewall next hop
! firewall VIP 20.1.1.1

VRF context INSIDE
 Vni 1001000
 IP route 20.1.1.0/24 10.1.1.1
! static route on VTEP pointing to Firewall next hop
! firewall VIP 10.1.1.1
```

Redistribute Static Routes into BGP and Advertise to the Rest of the Fabric

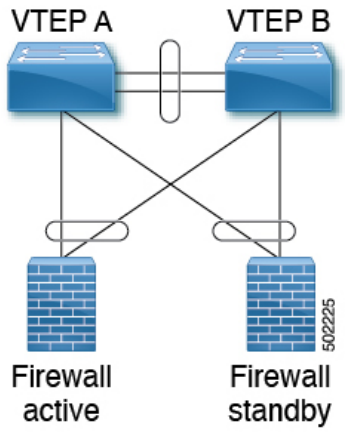
Through redistribution, we make the route toward the active firewall shown to the VTEP where it resides. The route is seen as a prefix route (EVPN Route-Type5), and as such, only the route toward the VTEP with the active firewall is seen. In the case of a firewall active/standby change, the tracking needs to detect the change and inform all of the remote VTEPs of this change. This behavior is equal to a route "delete" followed by an "add." This approach needs to notify all VTEPs with the VRF, and hence a wider churn can be seen.

VTEP A and VTEP B:

```
router bgp 65000
 vrf OUTSIDE
  address-family ipv4 unicast
   redistribute static route-map Static-to-BGP
```

Dual-Attached Firewall with Static Routing

Figure 60: Dual-Attached Firewall with Static Routing



VTEP A and VTEP B:

```
Vlan 10
 Name inside
 Vn-segment 10010

Vlan 20
 Name outside
 Vn-segment 10020

interface nve1
 no shutdown
 host-reachability protocol bgp
 source-interface loopback1
 member vni 10010
  mcastgroup 239.1.1.1
 member vni 10020
  mcastgroup 239.1.1.1
 member vni 1001000 associate-vrf
 member vni 1002000 associate-vrf


Interface VLAN 10
 Description inside_vlan
 VRF member INSIDE
 IP address 10.1.1.254/24
 fabric forwarding mode anycast-gateway

Interface VLAN 20
 Description outside_vlan
 VRF member OUTSIDE
 IP address 20.1.1.254/24
 fabric forwarding mode anycast-gateway

VRF context INSIDE
 Vni 1001000
 IP route 20.1.1.0/24 10.1.1.1
! static route on VTEP pointing to Firewall next hop
! firewall VIP 10.1.1.1
VRF context OUTSIDE
 Vni 1002000
 IP route 10.1.1.0/24 20.1.1.1
! static route on VTEP pointing to Firewall next hop
! firewall VIP 20.1.1.1


router bgp 65000
 vrf INSIDE
  address-family ipv4 unicast
   redistribute static route-map INSIDE-to-BGP
 vrf OUTSIDE
  address-family ipv4 unicast
   redistribute static route-map OUTSIDE-to-BGP
```

Single-Attached Firewall with eBGP Routing

If the firewall supports BGP, one option is to use BGP as a protocol between the firewall and the service VTEP. Peering using the anycast IP is not supported. The recommended design is to use dedicated loopback IPs on each VTEP and peer using the loopback. As long as the loopback interfaces are not advertised via

EVPN, the same IP address could be used on all of the belonging VTEPs. We recommend using individual IP addresses on a per-VTEP basis.

Reachability to the loopback from the firewall can be configured using a static route on the firewall, pointing to the Anycast Gateway IP on the VTEPs.

In the following example, an eBGP peering is established from the VTEPs, which are in AS 65000, and the firewall in AS 65002. The BGP peering with iBGP is not supported.



Note When having eBGP peering to active/standby firewalls connected to different VTEPs, **export-gateway-ip** must be enabled.

Do not use Anycast Gateway for BGP peerings.

VTEP A:

```
Vlan 10
 Name inside
 Vn-segment 10010

Vlan 20
 Name outside
 Vn-segment 10020


Interface VLAN 10
 Description inside_vlan
 VRF member INSIDE
IP address 10.1.1.254/24
fabric forwarding mode anycast-gateway

Interface loopback100
 Vrf member INSIDE
 Ip address 172.16.1.253/32

 Interface VLAN 20
 Description outside_vlan
 VRF member OUTSIDE
 IP address 20.1.1.254/24
 fabric forwarding mode anycast-gateway

 Interface loopback101
 Vrf member OUTSIDE
 Ip address 172.18.1.253/32


router bgp 65000
 vrf INSIDE
 ! peer with Firewall Inside
 neighbor 10.1.1.0/24 remote-as 65123
 update-source loopback100
 ebgp-multihop 5
 address-family ipv4 unicast
  local-as 65051 no-prepend replace-as

 vrf OUTSIDE
 ! peer with Firewall Outside
 neighbor 20.1.1.0/24 remote-as 65123
 update-source loopback101
 ebgp-multihop 5
```

```
 address-family ipv4 unicast
  local-as 65052 no-prepend replace-as
```

VTEP B:

```
Vlan 10
 Name inside
 Vn-segment 10010

Vlan 20
 Name outside
 Vn-segment 10020
Interface VLAN 10
 Description inside_vlan
 VRF member INSIDE
IP address 10.1.1.254/24
fabric forwarding mode anycast-gateway

Interface loopback100
 Vrf member INSIDE
 Ip address 172.16.1.254/32

 Interface VLAN 20
 Description outside_vlan
 VRF member OUTSIDE
 IP address 20.1.1.254/24
 fabric forwarding mode anycast-gateway

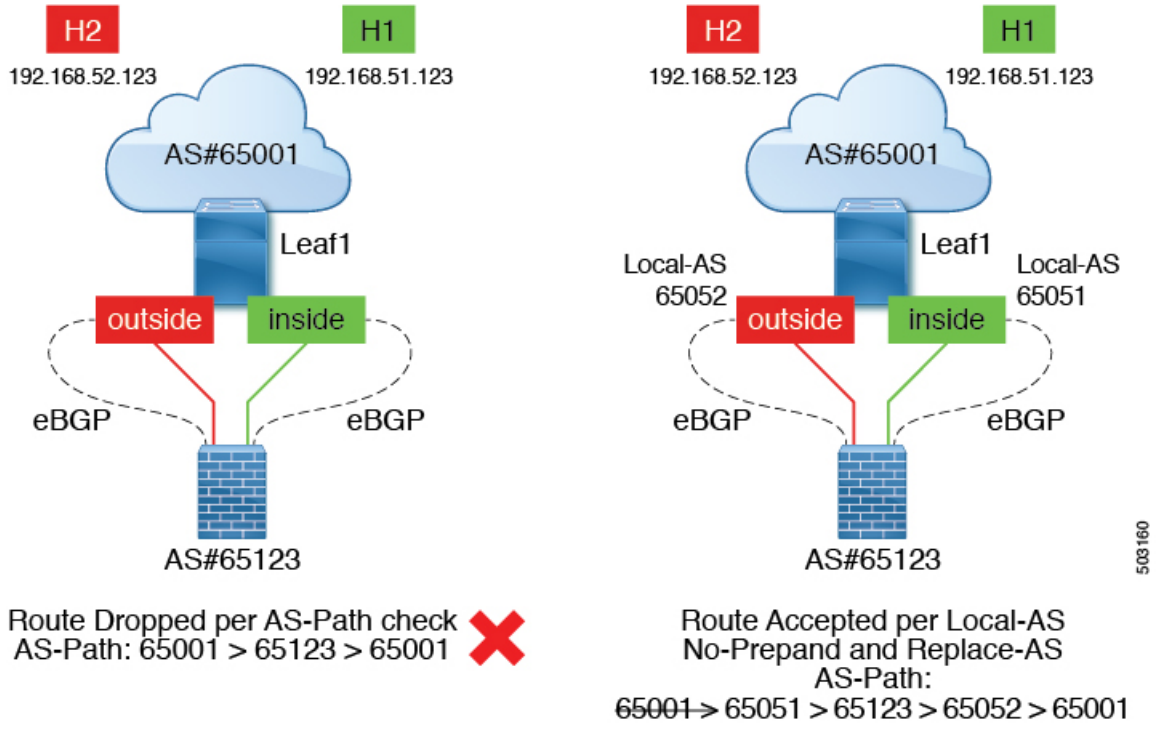
 Interface loopback101
 Vrf member OUTSIDE
 Ip address 172.18.1.254/32


router bgp 65000
 vrf INSIDE
 ! peer with Firewall Inside
 neighbor 10.1.1.0/24 remote-as 65123
 update-source loopback100
 ebgp-multihop 5
 address-family ipv4 unicast
  local-as 65051 no-prepend replace-as

 vrf OUTSIDE
 ! peer with Firewall Outside
 neighbor 20.1.1.0/24 remote-as 65123
 update-source loopback101
 ebgp-multihop 5
 address-family ipv4 unicast
  local-as 65052 no-prepend replace-as
```

With the VXLAN fabric generally being in a single BGP Autonomous System (AS), the AS of the inside VRF and the outside VRF is the same. BGP does not install routes that are received from its own AS. Therefore, we need to adjust the AS-path to override this rule. Various approaches exist, including disabling the rule that BGP drops routes from its own AS, which has further implications to the network. To keep all of the BGP protection mechanics in place, the “local-as” approach allows you to mimic routes being originated from a different AS. We recommend inserting the “local-as #ASN# no-prepend replace-as” on each firewall peering with different “local-as” per VRF.

Figure 61: eBGP AS-Path Check



Dual-Attached Firewall with eBGP Routing

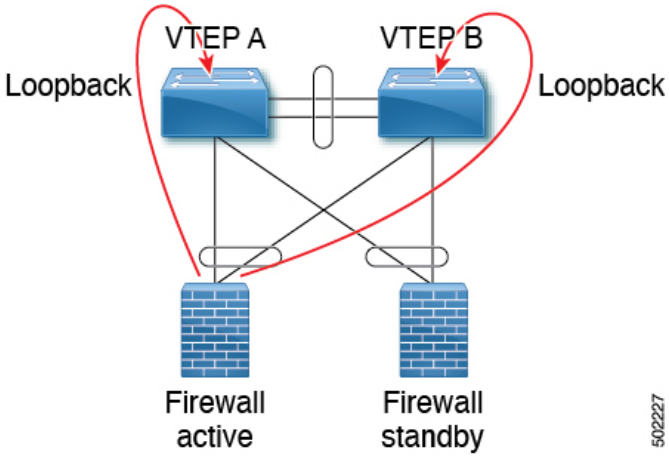
If the firewall supports BGP, one option is to use BGP as a protocol between the firewall and the service VTEP. Peering using the anycast IP is not supported. The recommended design is to use dedicated loopback IPs on each VTEP and peer using the loopback. As long as the loopback interfaces are not advertised via EVPN, the same IP address could be used on all of the belonging VTEPs. We recommend using individual IP addresses on a per-VTEP basis. For vPC environments, it is required.

Reachability to the loopback from the firewall can be configured using a static route on the firewall, pointing to the Anycast Gateway IP on the VTEPs.

In vPC deployments, you must have a per-VRF peering via a vPC peer-link. In addition to the per-VRF peering, you can enable the advertisement of prefix routes (EVPN Route-Type 5) using the **advertise-pip** command. For vPC with fabric peering, the per-VRF peering is not necessary, and the advertisement of prefix routes (EVPN Route-Type5) is required.

In the following example, an eBGP peering is established from the VTEPs, which are in AS 65000, and the firewall in AS 65002. The BGP peering with iBGP is not supported.

Figure 62: Dual-Attached Firewall with eBGP



Note When having eBGP peering to active/standby firewalls connected to different VTEPs, **export-gateway-ip** must be enabled.

Do not use Anycast Gateway for BGP peerings.

VTEP A:

```
Vlan 10
 Name inside
 Vn-segment 10010

Vlan 20
 Name outside
 Vn-segment 10020


Interface VLAN 10
 Description inside_vlan
 VRF member INSIDE
IP address 10.1.1.254/24
fabric forwarding mode anycast-gateway

Interface loopback100
 Vrf member INSIDE
 Ip address 172.16.1.253/32

 Interface VLAN 20
 Description outside_vlan
 VRF member OUTSIDE
 IP address 20.1.1.254/24
 fabric forwarding mode anycast-gateway

 Interface loopback101
 Vrf member OUTSIDE
 Ip address 172.18.1.253/32


router bgp 65000
vrf INSIDE
 ! peer with Firewall Inside
```

```
 neighbor 10.1.1.0/24 remote-as 65123
 update-source loopback100
 ebgp-multihop 5
 address-family ipv4 unicast
  local-as 65051 no-prepend replace-as

 vrf OUTSIDE
 ! peer with Firewall Outside
 neighbor 20.1.1.0/24 remote-as 65123
 update-source loopback101
 ebgp-multihop 5
 address-family ipv4 unicast
  local-as 65052 no-prepend replace-as
```

VTEP B:

```
Vlan 10
 Name inside
 Vn-segment 10010

Vlan 20
 Name outside
 Vn-segment 10020


Interface VLAN 10
 Description inside_vlan
 VRF member INSIDE
IP address 10.1.1.254/24
fabric forwarding mode anycast-gateway

Interface loopback100
 Vrf member INSIDE
 Ip address 172.16.1.254/32

 Interface VLAN 20
 Description outside_vlan
 VRF member OUTSIDE
 IP address 20.1.1.254/24
 fabric forwarding mode anycast-gateway

 Interface loopback101
 Vrf member OUTSIDE
 Ip address 172.18.1.254/32


router bgp 65000
 vrf INSIDE
 ! peer with Firewall Inside
 neighbor 10.1.1.0/24 remote-as 65123
 update-source loopback100
 ebgp-multihop 5
 address-family ipv4 unicast
  local-as 65051 no-prepend replace-as

 vrf OUTSIDE
 ! peer with Firewall Outside
 neighbor 20.1.1.0/24 remote-as 65123
 update-source loopback101
 ebgp-multihop 5
 address-family ipv4 unicast
  local-as 65052 no-prepend replace-as
```

Per-VRF Peering via vPC Peer-Link

VTEP A and VTEP B:

```
vlan 3966
! vlan use for peering between the vPC VTEPS

vlan 3967
! vlan use for peering between the vPC VTEPS

system nve infra-vlans 3966,3967

interface vlan 3966
vrf memner INSIDE
 ip address 100.1.1.1/31

interface vlan 3967
 vrf memner OUTSIDE
 ip address 100.1.2.1/31

router bgp 65000
 vrf INSIDE
 neighbor 100.1.1.0 remote-as 65000
 update-source vlan 3966
 next-hop self
 address-family ipv4 unicast

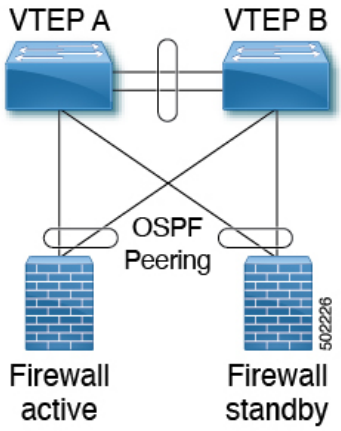
 vrf OUTSIDE
 neighbor 100.1.2.0 remote-as 65000
 update-source vlan 3967
 next-hop self
 address-family ipv4 unicast
```

The routes learned in each VRF are advertised to the rest of the fabric via BGP EVPN updates.

Dual-Attached Firewall with OSPF

Cisco NX-OS supports dynamic OSPF peering over vPC using Layer 3, which enables firewall connectivity using vPC and establishes OSPF peering over this link. The VLAN used to establish peering between the Cisco Nexus 9000 switches and the firewall must be a non-VXLAN-enabled VLAN.

Figure 63: Dual-Attached Firewall with OSPF





Note Do not use Anycast Gateway for OSPF adjacencies.

VTEP A:

```
Vlan 10
 Name inside

Vlan 20
 Name outside

Interface VLAN 10
 Description inside_vlan
 VRF member INSIDE
 IP address 10.1.1.253/24
 Ip router ospf 1 area 0

Interface VLAN 20
 Description outside_vlan
 VRF member OUTSIDE
 IP address 20.1.1.253/24
 Ip router ospf 1 area 0

vpc domain 100
 layer3 peer-router
 peer-gateway
 peer-switch
 peer-keepalive destination x.x.x.x source x.x.x.x peer-gateway
 ipv6 nd synchronize
 ip arp synchronize

router ospf 1
 router-id 192.168.1.1
  vrf INSIDE
  VRF OUTSIDE
```

VTEP B:

```
Vlan 10
 Name inside

Vlan 20
 Name outside

Interface VLAN 10
 Description inside_vlan
 VRF member INSIDE
 IP address 10.1.1.254/24
 Ip router ospf 1 area 0

Interface VLAN 20
 Description outside_vlan
 VRF member OUTSIDE
 IP address 20.1.1.254/24
 Ip router ospf 1 area 0

vpc domain 100
 layer3 peer-router
 peer-gateway
 peer-switch
 peer-keepalive destination x.x.x.x source x.x.x.x peer-gateway
```

```
 ipv6 nd synchronize
 ip arp synchronize

router ospf 1
 router-id 192.168.2.1
  vrf INSIDE
  VRF OUTSIDE

VTEPA# show ip route ospf-1 vrf OUTSIDE
IP Route Table for VRF "OUTSIDE"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

10.1.1.0/24, ubest/mbest: 1/0
  *via 20.1.1.1 Vlan20, [110/41], 1w5d, ospf-1, intra

VTEPA# show ip route ospf-1 vrf INSIDE
IP Route Table for VRF "INSIDE"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

20.1.1.0/24, ubest/mbest: 1/0
  *via 10.1.1.1 Vlan10, [110/41], 1w5d, ospf-1, intra
```

Redistribute OSPF Routes into BGP and Advertise to the Rest of the Fabric

VTEP A and VTEP B:

```
router bgp 65000
 vrf OUTSIDE
  address-family ipv4 unicast
   redistribute ospf 1 route-map OUTSIDEOSPF-to-BGP
 vrf INSIDE
  address-family ipv4 unicast
   redistribute ospf 1 route-map INSIDEOSPF-to-BGP
```

Firewall as Default Gateway

In this deployment model, the VXLAN fabric is a Layer 2 fabric, and the default gateway resides on the firewall.

For example:

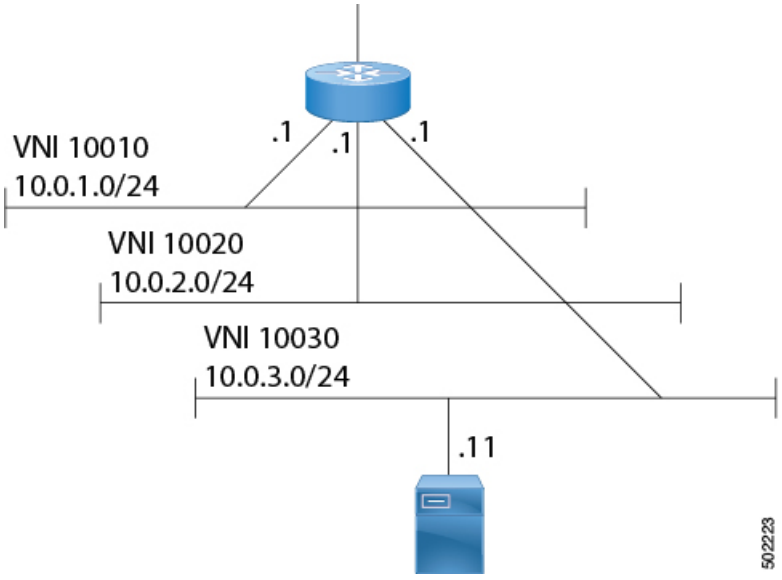
```
vlan 10
  name WEB
  vn-segment 10010
vlan 20
  name APPLICATION
  vn-segment 10020
vlan 30
  name DATABASE
  vn-segment 10030

interface nve1
  no shutdown
```

```
host-reachability protocol bgp
source-interface loopback1
member vni 10010
 mcastgroup 239.1.1.1
member vni 10020
 mcastgroup 239.1.1.1
member vni 10030
 mcastgroup 239.1.1.1
```

The firewall has a logical interface in each VNI and is the default gateway for all endpoints. Every inter-VNI communication flows through the firewall. Take special care with the sizing of the firewall so that it does not become a bottleneck. Therefore, use this design in environments with low-bandwidth requirements.

Figure 64: Firewall as Default Gateway with a Layer-2 VXLAN Fabric



Transparent Firewall Insertion

Transparent firewalls or Layer 2 firewalls (including IPS/IDS) typically bridge between an inside VLAN and outside VLAN and inspect traffic as it traverses through them. VLAN stitching is done by placing the default gateway for the service on the inside VLAN. The Layer 2 reachability to this gateway is done on the outside VLAN.

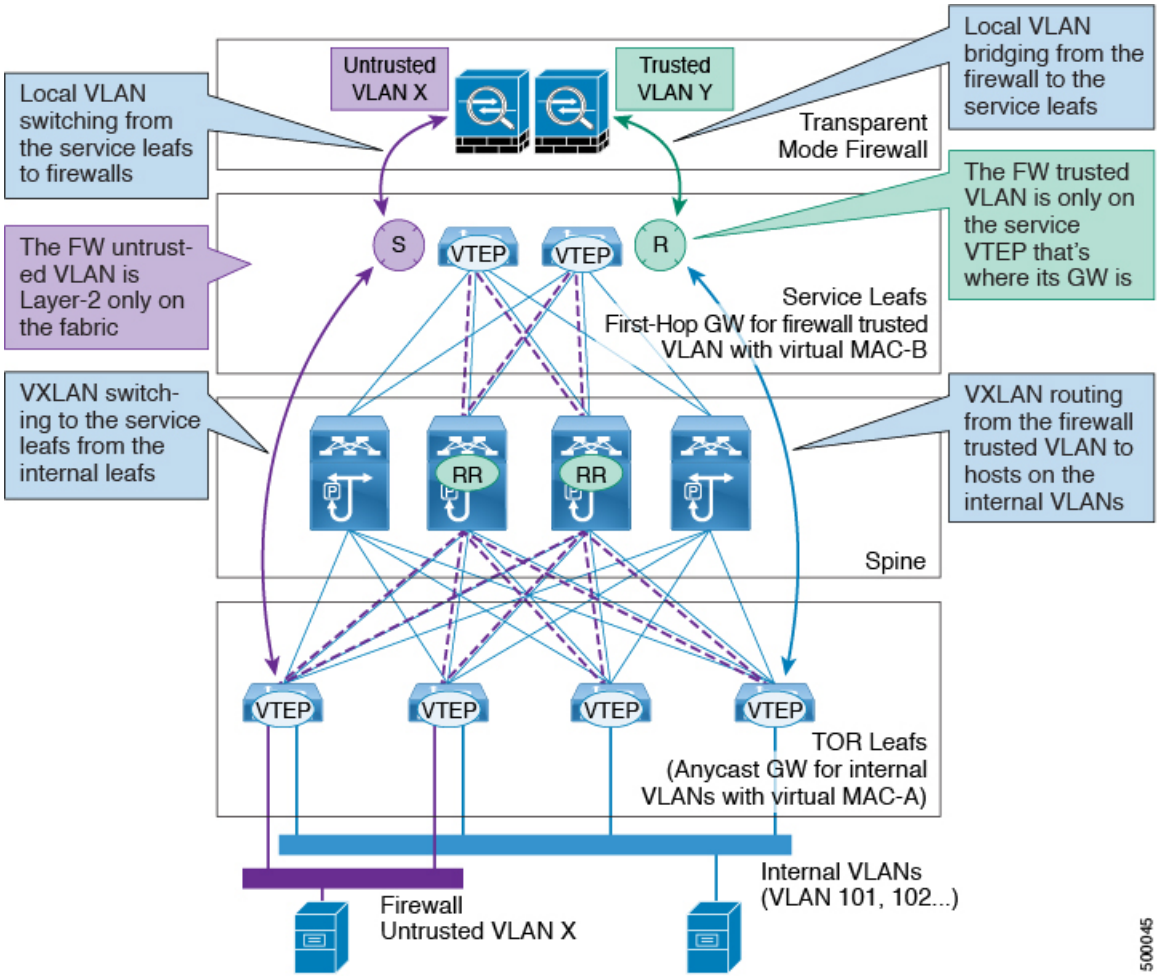
Overview of EVPN with Transparent Firewall Insertion

The topology contains the following types of VLANs:

- Internal VLAN (a regular VXLAN on ToR leafs with Anycast Gateway)
- Firewall untrusted VLAN X
- Firewall trusted VLAN Y

In this topology, the traffic that goes from VLAN X to other VLANs must go through a transparent Layer 2 firewall that is attached to the service leafs. This topology utilizes an approach of an untrusted VLAN X and a trusted VLAN Y. All ToR leafs have a Layer 2 VNI VLAN X. There is no SVI for VLAN X. The service leafs that are connected to the firewall have Layer 2 VNI VLAN X, non-VXLAN VLAN Y, and SVI Y with an HSRP gateway.

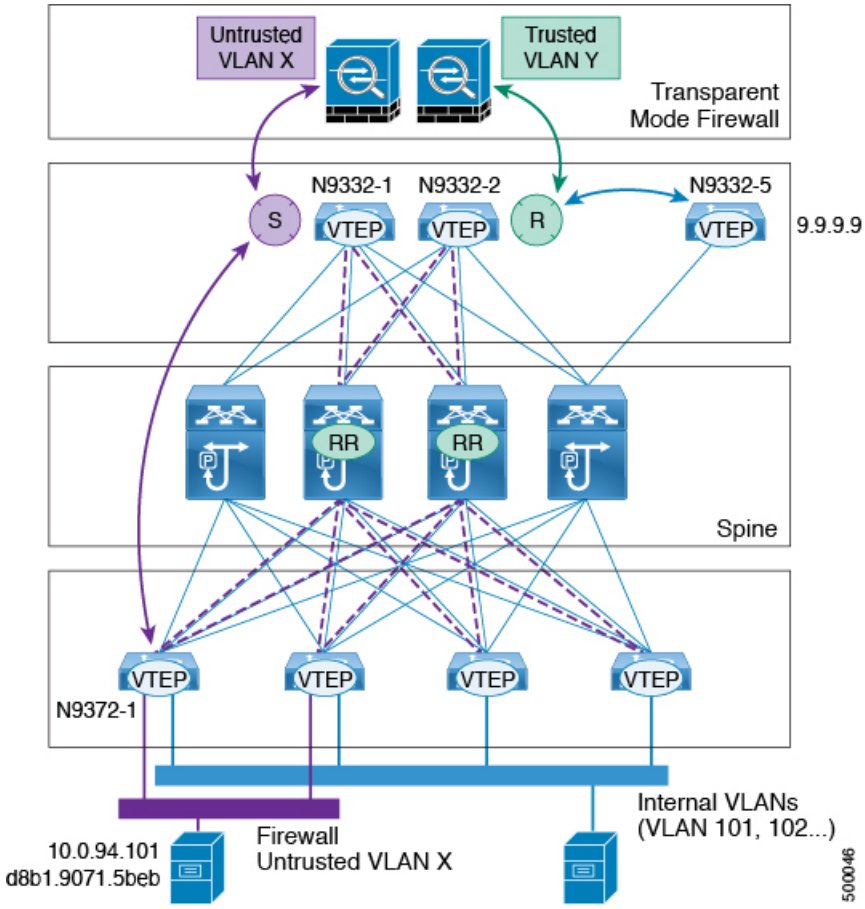
Overview of EVPN with Transparent Firewall Insertion



Note For VXLAN EVPN, we recommend using the distributed Anycast Gateway with transparent firewall insertion. Doing so allows all VLANs to be VXLAN enabled. When using an HSRP/VRRP-based First-Hop Gateway, the VLAN for the SVI can't be VXLAN enabled and should reside on a vPC pair for redundancy.

EVPN with Transparent Firewall Insertion Example

Example of EVPN with Transparent Firewall Insertion



- Host in VLAN X: 10.1.94.101
- ToR leaf: N9372-1
- Service leaf in vPC: N9332-1 and N9332-2
- Border leaf: N9332-5

ToR Leaf Configuration

```
vlan 94
vn-segment 100094

interface nve1
  member vni 100094
   mcastgroup 239.1.1.1

router bgp 64500
  routerid 1.1.2.1
  neighbor 1.1.1.1 remote-as 64500
  address-family l2vpn evpn
    send-community extended
```

```
  neighbor 1.1.1.2 remote-as 64500
  address-family l2vpn evpn
    send-community extended
  vrf Ten1
   address-family ipv4 unicast
     advertise l2vpn evpn

evpn
vni 100094 l2
  rd auto
  route-target import auto
  route-target export auto
```

Service Leaf 1 Configuration Using HSRP

```
vlan 94
description untrusted_vlan
  vn-segment 100094

vlan 95
  description trusted_vlan

vpc domain 10
  peer-switch
  peer-keepalive destination 10.1.59.160
  peer-gateway
  auto-recovery
  ip arp synchronize

interface Vlan2
description vpc_backup_svi_for_overlay
  no shutdown
  no ip redirects
  ip address 10.10.60.17/30
  no ipv6 redirects
  ip router ospf 100 area 0.0.0.0
  ip ospf bfd
  ip pim sparsemode

interface Vlan95
description SVI_for_trusted_vlan
  no shutdown
  mtu 9216
  vrf member Ten-1
  no ip redirects
  ip address 10.0.94.2/24
  hsrp 0
   preempt priority 255
   ip 10.0.94.1

interface nve1
  member vni 100094
   mcast-group 239.1.1.1

router bgp 64500
  routerid 1.1.2.1
  neighbor 1.1.1.1 remote-as 64500
  address-family l2vpn evpn
   send-community extended
  neighbor 1.1.1.2 remote-as 64500
   address-family l2vpn evpn
    send-community extended
  vrf Ten-1
   address-family ipv4 unicast
```

```
    network 10.0.94.0/24 /*advertise /24 for SVI 95 subnet; it is not VXLAN anymore*/
    advertise l2vpn evpn

evpn
vni 100094 l2
 rd auto
 route-target import auto
 route-target export auto
```

Service Leaf 2 Configuration Using HSRP

```
vlan 94
  description untrusted_vlan
  vnsegment 100094

vlan 95
  description trusted_vlan

vpc domain 10
  peer-switch
  peer-keepalive destination 10.1.59.159
  peer-gateway
  auto-recovery
  ip arp synchronize

interface Vlan2
description vpc_backup_svi_for_overlay
  no shutdown
  no ip redirects
  ip address 10.10.60.18/30
  no ipv6 redirects
  ip router ospf 100 area 0.0.0.0
  ip pim sparsemode

interface Vlan95
  description SVI_for_trusted_vlan
  no shutdown
  mtu 9216
  vrf member Ten-1
  no ip redirects
  ip address 10.0.94.3/24
  hsrp 0
   preempt priority 255
   ip 10.0.94.1

interface nve1
  member vni 100094
   mcastgroup 239.1.1.1

router bgp 64500
  router-id 1.1.2.1
  neighbor 1.1.1.1 remote-as 64500
  address-family l2vpn evpn
    send-community extended
  neighbor 1.1.1.2 remote-as 64500
  address-family l2vpn evpn
   send-community extended
  vrf Ten-1
    address-family ipv4 unicast
     network 10.0.94.0/24 /*advertise /24 for SVI 95 subnet; it is not VXLAN anymore*/
     advertise l2vpn evpn

evpn
vni 100094 l2
```

```
  rd auto
  route-target import auto
  route-target export auto
```

Show Command Examples

Display information about the ingress leaf learned local MAC from host:

```
switch# sh mac add vl 94 | i 5b|MAC
* primary entry, G - Gateway MAC, (R) Routed - MAC, O - Overlay MAC
VLAN MAC Address Type age Secure NTFY Ports
* 94 d8b1.9071.5beb dynamic 0 F F Eth1/1
```

Display information about the service leaf found MAC of host:



Note In VLAN 94, the service leaf learned the host MAC from the remote peer by BGP.

```
switch# sh mac add vl 94 | i VLAN|eb

VLAN MAC Address Type age Secure NTFY Ports
* 94 d8b1.9071.5beb dynamic 0 F F nve1(1.1.2.1)

switch#  sh mac add vl 94 | i VLAN|eb

VLAN MAC Address Type age Secure NTFY Ports
* 94 d8b1.9071.5beb dynamic 0 F F nve1(1.1.2.1)

switch#  sh mac add vl 95 | i VLAN|eb

VLAN MAC Address Type age Secure NTFY Ports
+ 95 d8b1.9071.5beb dynamic 0 F F Po300

switch#  sh mac add vl 95 | i VLAN|eb

VLAN MAC Address Type age Secure NTFY Ports
+ 95 d8b1.9071.5beb dynamic 0 F F Po300
```

Display information about service leaf learned ARP for host on VLAN 95:

```
switch# sh ip arp vrf ten-1
Address     Age      MAC Address     Interface
10.0.94.101 00:00:26 d8b1.9071.5beb  Vlan95
```

Service Leaf learns 9.9.9.9 from EVPN.

```
switch# sh ip route vrf ten-1 9.9.9.9
IP Route Table for VRF "Ten-1"
'*' denotes best ucast nexthop
'**' denotes best mcast nexthop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

9.9.9.9/32, ubest/mbest: 1/0
  *via 1.1.2.7%default, [200/0], 02:57:27, bgp64500,internal, tag 65000 (evpn) segid: 10011
tunnelid: 0x1
010207 encap: VXLAN
```

Display information about the border leaf learned host routes by BGP:

```
switch# sh ip route 10.0.94.101

IP Route Table for VRF "default"
'*' denotes best ucast nexthop
'**' denotes best mcast nexthop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

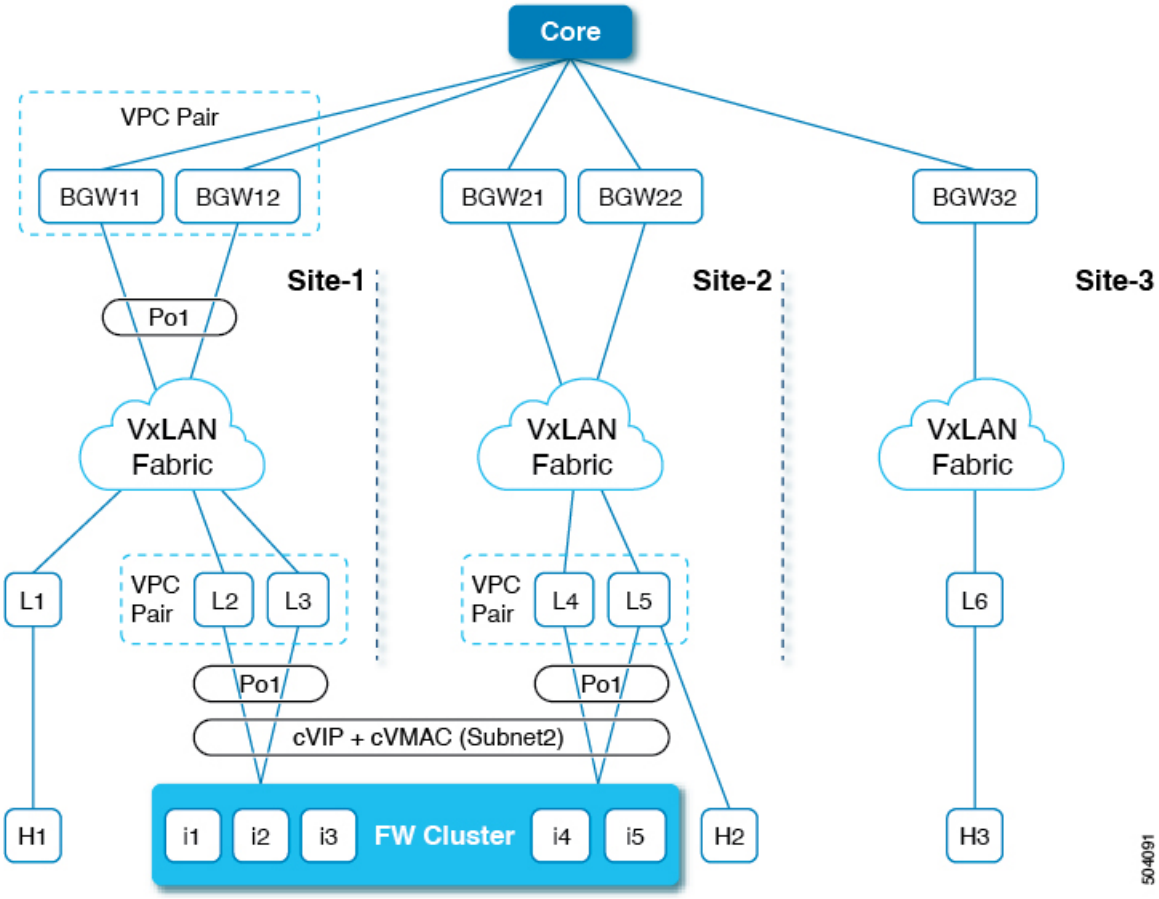
10.0.94.0/24, ubest/mbest: 1/0
  *via 10.100.5.0, [20/0], 03:14:27, bgp65000,external, tag 6450
```

Firewall Clustering with VXLAN BGP EVPN

This section provides details on how to configure a firewall cluster that spans across multiple sites running a VXLAN fabric with a BGP EVPN control plane.

The following topology illustrates the firewall clustering with VXLAN EVPN.

Figure 65: Firewall clustering with VXLAN EVPN



This topology covers the following:

- Firewall cluster consists of multiple instances that act as a single device.
- Routed Access to firewall can be through a different or same subnet.
- Firewall employs a L2 port-channel spanned across all instances.
- A common ESI represents all vPC port-channels that connect to the firewall cluster.
- Single VIP/VMAC is present across all instances.
- BGP-EVPN VXLAN overlay per site is stitched at Border Gateways.
- Anycast forwarding of Active-to-Active instances within the same site and Active-to-Backup access to firewall across sites for traffic flows is supported.
- Each site has a single vPC pair connected to the cluster with a port-channel interface assigned to it.
- The cluster VIP and cluster VMAC are advertised into the VXLAN EVPN fabric as BGP EVPN Route Target-2s (with the ESI set to the configured value on each vPC's port-channel interface). The next hop of the Route Target-2 is the vPC pair's VTEP VIP address.
- Each site may have multiple clusters. The clusters are attached to the vPC pair with their individual port-channels with unique ESIs.
- Each cluster has its own cVIP and cVMAC that are advertised into the VXLAN EVPN fabric as BGP EVPN Route Target -2s (with the ESI set to the configured value on its vPC's Port-channel interface).
- A cluster may have multiple VLANs on the port-channel connected to the vPC pair. Each cVIP/cVMAC learnt on a VLAN is advertised with its corresponding L2VNI as a Route T-2 EVPN route.
- VIP and VMAC (Firewall Hosts) are attached to a single spanned Ether-channel.
- Spanned Ether-channel extends across sites.
- Anycast forwarding to VIP is determined by leverage of existing BGP path attributes and best-path selection.

On the VTEP leafs attached to the firewall cluster, BGP uses a route-map to attach a community to firewall cluster-related EVPN EAD/ES (Type-1) and MAC/IP (Type-2) routes.

```
router bgp 12000
 address-family l2vpn evpn
originate-map set_esi
 template peer SITE-BGW
   remote-as 12000
   update-source loopback1
   address-family l2vpn evpn
     send-community
     send-community extended
 template peer VTEP-PEERS
   remote-as 12000
   update-source loopback1
   address-family l2vpn evpn
     send-community
     send-community extended
```

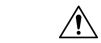
On the border gateways, BGP uses a route-map to match the firewall clustering community attached to EVPN EAD/ES (Type-1) and MAC/IP (Type-2) routes.

```
router bgp 11000
```

```
  bestpath as-path multipath-relax
  neighbor 111.111.10.1 remote-as 12000
   peer-type fabric-external
   address-family l2vpn evpn
     send-community
     send-community extended
     route-map preserve_esi out
     rewrite-evpn-rt-asn
```

On the VTEP leafs attached to the firewall cluster, you need to configure a route-map to attach a community to firewall cluster-related EVPN EAD/ES (Type-1) and MAC/IP (Type-2) routes.

```
route-map set_esi permit 10
  match tag 100000
  match evpn route-type 1 2
  set community 23456:12345
route-map set_esi permit 15
```



Caution The **match tag** command in a route-map associated with route-map <*name*> out BGP command under neighbor address-family mode is only supported if configured under address-family l2vpn evpn.

On the border gateways, you need to configure separate route-maps for fabric-internal and fabric-external peers to match the firewall clustering community attached to EVPN EAD/ES (Type-1) and MAC/IP (Type-2) routes.

Matching outbound L2VPN/EVPN route-map to fabric-internal peers:

```
route-map preserve_esi permit 10
  match community preserve_esi
  match evpn route-type 2
  set esi unchanged
route-map preserve_esi permit 15
route-map preserve_esi permit 30
```

Matching outbound L2VPN/EVPN route-map to fabric-external peers:

```
route-map preserve_esi_external permit 10
  match community preserve_esi
  match evpn route-type 2
  set esi unchanged
route-map preserve_esi_external permit 15
  match community preserve_esi
  match evpn route-type 1
route-map preserve_esi_external permit 20
  match evpn route-type 1
  match route-type local
route-map preserve_esi_external deny 25
  match evpn route-type 1
route-map preserve_esi_external permit 30
```

The ethernet-segment can be configured only under vPC port-channel.

```
interface port-channel 100
  ethernet-segment vpc
  esi <esi> [ tag <uint >]
interface port-channel 200
  ethernet-segment vpc
  esi system-mac <system-mac> <local-identifier> [tag <uint>]
```

A common ESI represents all vPC port-channels that connect to the firewall cluster. You can configure ESI under a vPC port-channel.

```
evpn esi multihoming
port-channel 100
  ethernet-segment 1
     system-mac aa.bb.cc <anycast-host>
```

Keep the same system-mac for all vPC port-channels that host the same firewall cluster.

For more firewall information, see Integrating Layer 3 Firewalls in VXLAN Fabrics.

Service Redirection in VXLAN EVPN Fabrics

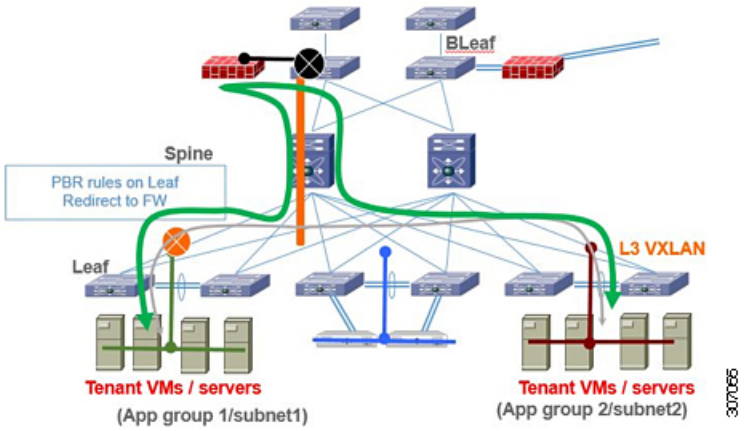
Today, insertion of service appliances (also referred to as service nodes or service endpoints) such as firewalls, load-balancers, etc are needed to secure and optimize applications within a data center. This section describes the Layer 4-Layer 7 service insertion and redirection features offered on VXLAN EVPN fabrics that provides sophisticated mechanisms to onboard and selectively redirect traffic to these services.

Use of Policy-Based Redirect for Services Insertion

Policy-based redirect (PBR) provides a mechanism to bypass a routing table lookup and redirect traffic to a next-hop IP reachable over VXLAN. The feature enables service redirection to Layer 4-Layer 7 devices such as firewalls and load balancers.

PBR involves configuring a route-map with rules that dictate where traffic must be forwarded. The route map is applied on the tenant SVI to influence traffic coming from the host-facing interfaces to a next hop reachable via the fabric.

In scenarios where traffic is coming to a VTEP from the overlay and needs to be redirected to another next hop, the PBR policy must be applied on the fabric facing Layer-3 VNI Interface.



In the previous figure, communication between App group 1 and App group 2 takes place via inter-VLAN/VNI routing in the tenant VRF by default. If there is a requirement where traffic from App group 1 to App group 2 must go through a firewall, a PBR policy can be used to redirect traffic. The example in section “Configuration Example for Policy-Based Redirect” provides the necessary configuration that redirects the traffic flow.

This VXLAN PBR functionality is very basic and lacks many of the required functionality for proper insertion of services in VXLAN fabric. Hence the recommendation is to instead look at ePBR for all the reasons explained in Enhanced-Policy Based Redirect (ePBR), on page 606 section.

Guidelines and Limitations for Policy-Based Redirect

The following guidelines and limitations apply to PBR over VXLAN.

- The following platforms support PBR over VXLAN:
  - Cisco Nexus 9300-FX/FX2/FX3 switches
  - Cisco Nexus 9300-GX switches
  - Cisco Nexus 9300-GX2 switches
  - Cisco Nexus 9332D-H2R switches
  - Cisco Nexus 93400LD-H1 switches
  - Cisco Nexus 9364C-H1 switches
  - Cisco Nexus 9504 and 9508 switches with X97160YC-EX, 9700-FX/FX3 line cards
- Beginning with Cisco NX-OS Release 10.2(3)F, the VXLAN PBR feature is supported with VXLANv6 on all TOR switches.
- PBR over VXLAN doesn't support the following features:VTEP ECMP, and the **load-share** keyword in the **set {ip | ipv6} next-hop** *ip-address* command.
- When you configure **bestpath as-path multipath-relax**, BGP installs all the multi-paths for IPv4 as best-path in URIB with least metric available among the paths.
- When you configure **bestpath as-path multipath-relax**, BGP doesn't install all the multi-paths for IPv6 as best-path in U6RIB. It will still have the individual metric available for those paths.

Enabling the Policy-Based Redirect Feature

To configure basic PBR, in cases where the advanced (and recommended) ePBR functions are not deployed, see the following sections:

- Enabling the Policy-Based Redirect Feature, on page 602
- Configuring a Route Policy, on page 603
- Verifying the Policy-Based Redirect Configuration, on page 605
- Configuration Example for Policy-Based Redirect, on page 605

Before you begin

Enable the policy-based redirect feature before you can configure a route policy.

SUMMARY STEPS

1. **configure terminal**
2. **[no] feature pbr**

3. (Optional) **show feature**
4. (Optional) **copy running-config startup-config**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**[no] feature pbr** **Example:** `switch(config)# feature pbr`	Enables the policy-based routing feature.
Step 3	(Optional) **show feature** **Example:** `switch(config)# show feature`	Displays enabled and disabled features.
Step 4	(Optional) **copy running-config startup-config** **Example:** `switch(config)# copy running-config startup-config`	Saves this configuration change.

Configuring a Route Policy

You can use route maps in policy-based routing to assign routing policies to the inbound interface. Cisco NX-OS routes the packets when it finds a next hop and an interface.



Note The switch has a RACL TCAM region by default for IPv4 traffic.

Before you begin

Configure the RACL TCAM region (using TCAM carving) before you apply the policy-based routing policy. For instructions, see the "Configuring ACL TCAM Region Sizes" section in the Cisco Nexus 9000 Series NX-OS Security Configuration Guide, Release 9.2(x).

SUMMARY STEPS

1. **configure terminal**
2. **interface** *type slot/port*
3. **{ip | ipv6} policy route-map** *map-name*
4. **route-map** *map-name* **[permit | deny]** [*seq*]
5. **match {ip | ipv6} address access-list-name** *name* [*name...*]
6. **set ip next-hop** *address1*

7. **set ipv6 next-hop** *address1*
8. (Optional) **set interface null0**
9. (Optional) **copy running-config startup-config**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**interface** *type slot/port* **Example:** `switch(config)# interface ethernet 1/2`	Enters interface configuration mode.
Step 3	**{ip | ipv6} policy route-map** *map-name* **Example:** `switch(config-inf)# ip policy route-map Testmap`	Assigns a route map for IPv4 or IPv6 policy-based routing to the interface.
Step 4	**route-map** *map-name* **[permit | deny]** [*seq*] **Example:** `switch(config-inf)# route-map Testmap`	Creates a route map or enters route-map configuration mode for an existing route map. Use *seq* to order the entries in a route map.
Step 5	**match {ip | ipv6} address access-list-name** *name* [*name...*] **Example:** `switch(config-route-map)# match ip address access-list-name ACL1`	Matches an IPv4 or IPv6 address against one or more IPv4 or IPv6 access control lists (ACLs). This command is used for policy-based routing and is ignored by route filtering or redistribution.
Step 6	**set ip next-hop** *address1* **Example:** `switch(config-route-map)# set ip next-hop 192.0.2.1`	Sets the IPv4 next-hop address for policy-based routing.
Step 7	**set ipv6 next-hop** *address1* **Example:** `switch(config-route-map)# set ipv6 next-hop 2001:0DB8::1`	Sets the IPv6 next-hop address for policy-based routing.
Step 8	(Optional) **set interface null0** **Example:** `switch(config-route-map)# set interface null0`	Sets the interface that is used for routing. Use the **null0** interface to drop packets.
Step 9	(Optional) **copy running-config startup-config** **Example:**	Saves this configuration change.

	Command or Action	Purpose
	`switch(config-route-map)# copy running-config startup-config`	

Verifying the Policy-Based Redirect Configuration

To display the policy-based redirect configuration information, perform one of the following tasks:

Command	Purpose
show [ip | ipv6] policy [*name*]	Displays information about an IPv4 or IPv6 policy.
show route-map [*name*] **pbr-statistics**	Displays policy statistics.

Use the **route-map** *map-name* **pbr-statistics** command to enable policy statistics. Use the **clear route-map** *map-name* **pbr-statistics** command to clear these policy statistics.

Configuration Example for Policy-Based Redirect

Perform the following configuration on all tenant VTEPs, excluding the service VTEP.

```
feature pbr

ipv6 access-list IPV6_App_group_1
10 permit ipv6 any 2001:10:1:1::0/64

ip access-list IPV4_App_group_1
10 permit ip any 10.1.1.0/24

ipv6 access-list IPV6_App_group_2
10 permit ipv6 any 2001:20:1:1::0/64

ip access-list IPV4_App_group_2
10 permit ip any 20.1.1.0/24

route-map IPV6_PBR_Appgroup1 permit 10
  match ipv6 address IPV6_App_group_2
  set ipv6 next-hop 2001:100:1:1::20  (next hop is that of the firewall)

route-map IPV4_ PBR_Appgroup1 permit 10
  match ip address IPV4_App_group_2
  set ip next-hop 10.100.1.20 (next hop is that of the firewall)

route-map IPV6_PBR_Appgroup2 permit 10
  match ipv6 address IPV6_App_group1
  set ipv6 next-hop 2001:100:1:1::20  (next hop is that of the firewall)

route-map IPV4_ PBR_Appgroup2 permit 10
  match ip address IPV4_App_group_1
  set ip next-hop 10.100.1.20 (next hop is that of the firewall)


interface Vlan10
! tenant SVI appgroup 1
vrf member appgroup
 ip address 10.1.1.1/24
 no ip redirect
 ipv6 address 2001:10:1:1::1/64
 no ipv6 redirects
 fabric forwarding mode anycast-gateway
```

```
ip policy route-map IPV4_ PBR_Appgroup1
ipv6 policy route-map IPV6_PBR_Appgroup1
interface Vlan20
! tenant SVI appgroup 2
vrf member appgroup
 ip address 20.1.1.1/24
 no ip redirect
 ipv6 address 2001:20:1:1::1/64
 no ipv6 redirects
 fabric forwarding mode anycast-gateway
ip policy route-map IPV4_ PBR_Appgroup2
ipv6 policy route-map IPV6_PBR_Appgroup2

On the service VTEP, the PBR policy is applied on the tenant VRF SVI. This ensures the
traffic post decapsulation will be redirected to firewall.
feature pbr

ipv6 access-list IPV6_App_group_1
10 permit ipv6 any 2001:10:1:1::0/64

ip access-list IPV4_App_group_1
10 permit ip any 10.1.1.0/24

ipv6 access-list IPV6_App_group_2
10 permit ipv6 any 2001:20:1:1::0/64

ip access-list IPV4_App_group_2
10 permit ip any 20.1.1.0/24

route-map IPV6_PBR_Appgroup1 permit 10
  match ipv6 address IPV6_App_group_2
  set ipv6 next-hop 2001:100:1:1::20  (next hop is that of the firewall)

route-map IPV6_PBR_Appgroup permit 20
  match ipv6 address IPV6_App_group1
  set ipv6 next-hop 2001:100:1:1::20  (next hop is that of the firewall)

route-map IPV4_ PBR_Appgroup permit 10
  match ip address IPV4_App_group_2
  set ip next-hop 10.100.1.20 (next hop is that of the firewall)

route-map IPV4_ PBR_Appgroup permit 20
  match ip address IPV4_App_group_1
  set ip next-hop 10.100.1.20 (next hop is that of the firewall)

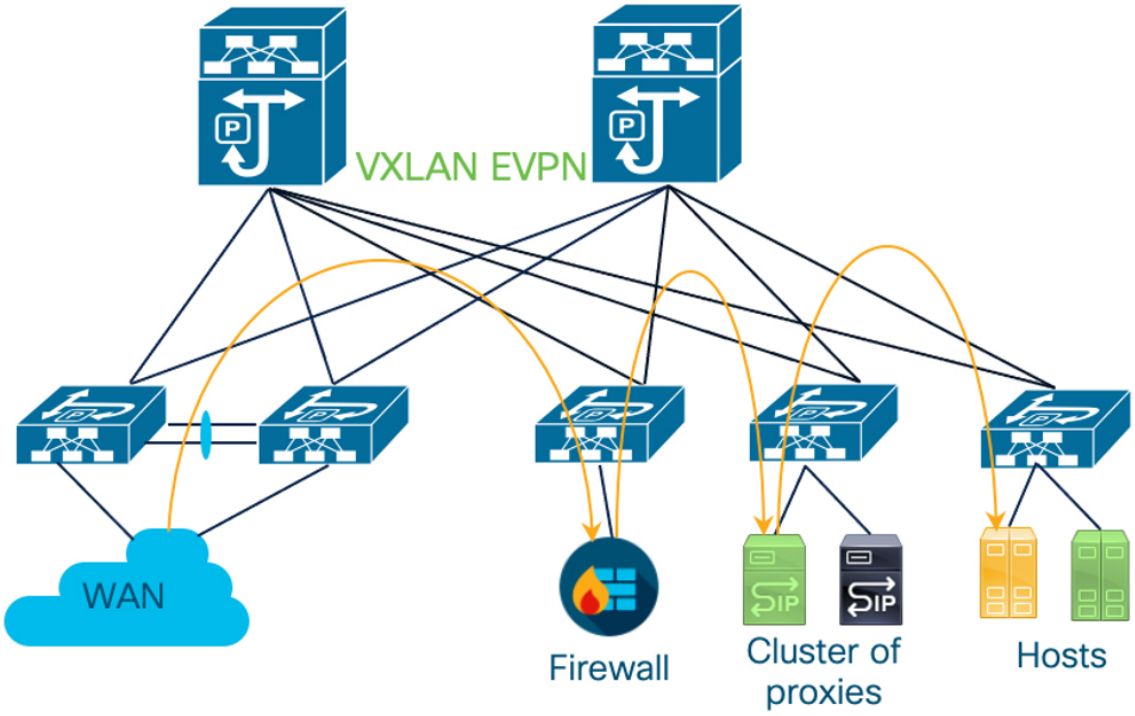

interface vlan1000
!L3VNI SVI for Tenant VRF
vrf member appgroup
ip forward
ipv6 forward
ipv6 ipv6 address use-link-local-only
ip policy route-map IPV4_ PBR_Appgroup
ipv6 policy route-map IPV6_PBR_Appgroup
```

Enhanced-Policy Based Redirect (ePBR)

VXLAN PBR as a solution to selectively redirect traffic can only cater to simple traffic redirection requirements. For more complex use cases like service chaining, symmetric load-balancing, or tracking health of service appliances, usage of PBR becomes difficult. The challenge with service chaining using PBR is that it requires the user to create unique policies per node and manage the redirection rules manually across all the nodes in

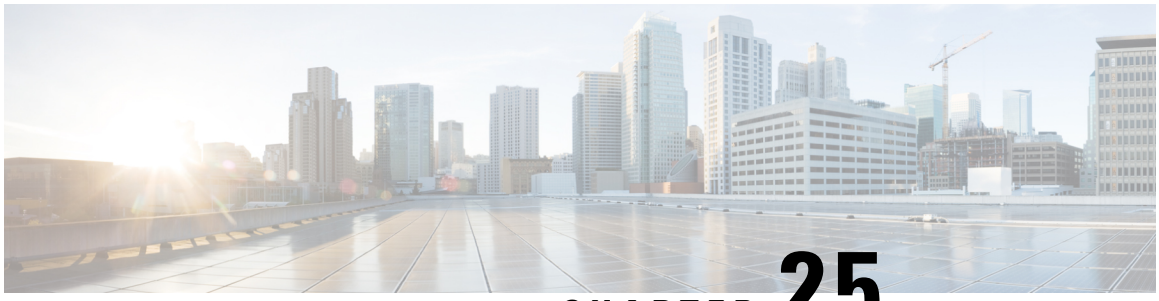
the chain. Also, given the stateful nature of the service nodes, the PBR rules must ensure symmetry for the reverse traffic, and this adds additional complexity to the configuration and management of the PBR policies.

Enhanced Policy-Based Redirect (ePBR) provides a comprehensive solution to insert service nodes, selectively redirect and load-balance traffic. ePBR provides a simplified workflow to create traffic chains and load-balancing rules along with providing options for probing/monitoring the health of service appliances and taking corrective action in the event of failure. ePBR is supported in both single and multi-site VXLAN EVPN deployments.



In this Figure, selective traffic originating from WAN is chained to a firewall and then the traffic is load-balanced across a cluster of proxies before forwarding toward the destination hosts. ePBR ensures symmetry is maintained for a given flow by making sure that traffic in both forward and reverse direction is redirected to the same service endpoint in the cluster of TCP proxies.

For more detailed information, guidelines and configuration examples on ePBR, see Cisco Nexus 9000 Series NX-OS ePBR Configuration Guide and Layer 4 to Layer 7 Service Redirection with Enhanced Policy-Based Redirect White Paper.



CHAPTER 25

Configuring Proportional Multipath for VNF

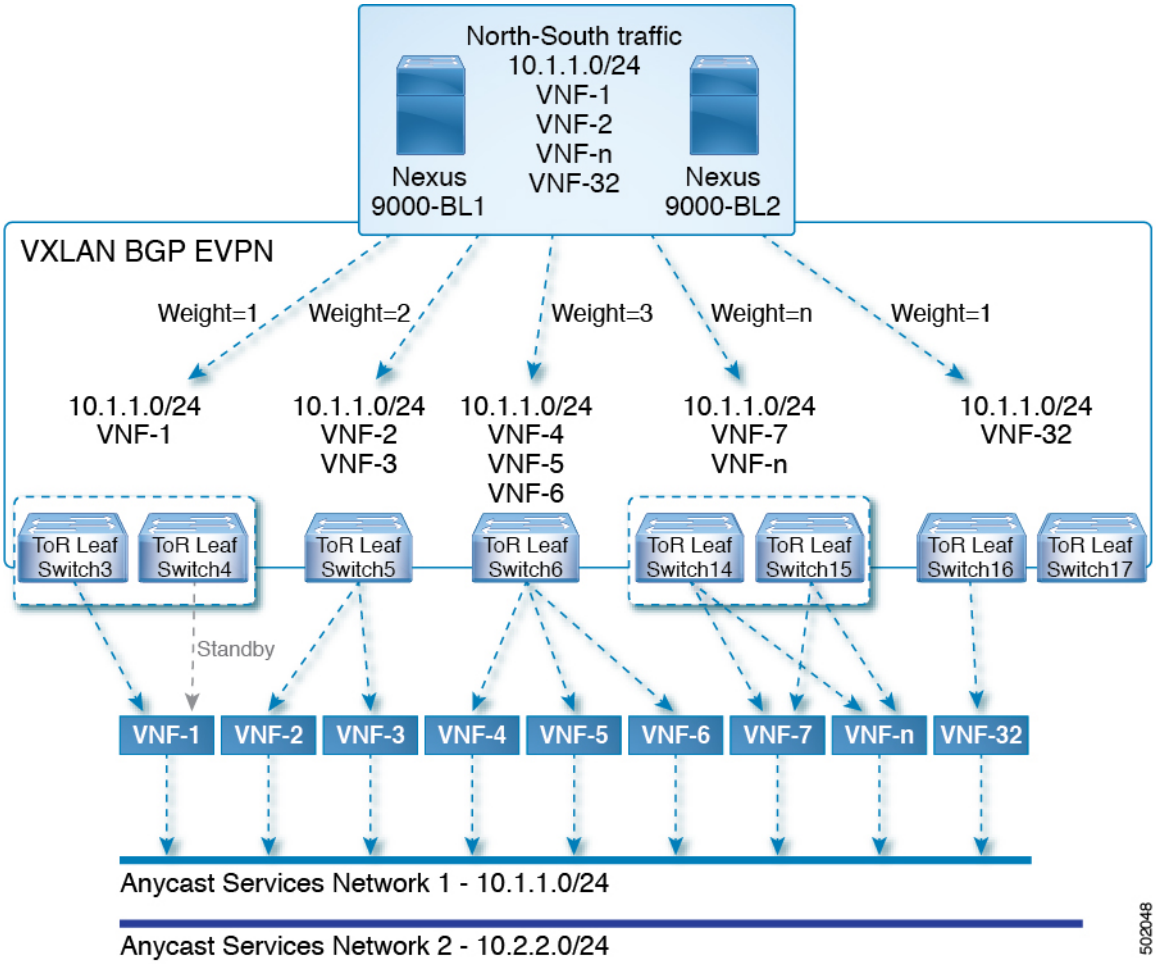
This chapter contains the following sections:

- About proportional multipath for VNF, on page 609
- Proportional Multipath for VNF with Multi-Site, on page 613
- Prerequisites for Proportional Multipath for VNF, on page 614
- Guidelines and limitations for proportional multipath for VNF, on page 614
- Configuring the Spine or Route Reflector, on page 616
- Configuring the Leaf or ToR, on page 617
- Configuring the Border Leaf or Border Gateway, on page 623
- Configuring the BGP Legacy Peer, on page 629
- Configure next-hop filtering, on page 630
- Configuring a User-Defined Profile for Maintenance Mode, on page 634
- Configuring a User-Defined Profile for Normal Mode, on page 634
- Configuring a Default Route Map, on page 635
- Applying a Route Map to a Route Reflector, on page 635
- Verifying Proportional Multipath for VNF, on page 636
- Configuration Example for Proportional Multipath for VNF with Multi-Site, on page 640

About proportional multipath for VNF

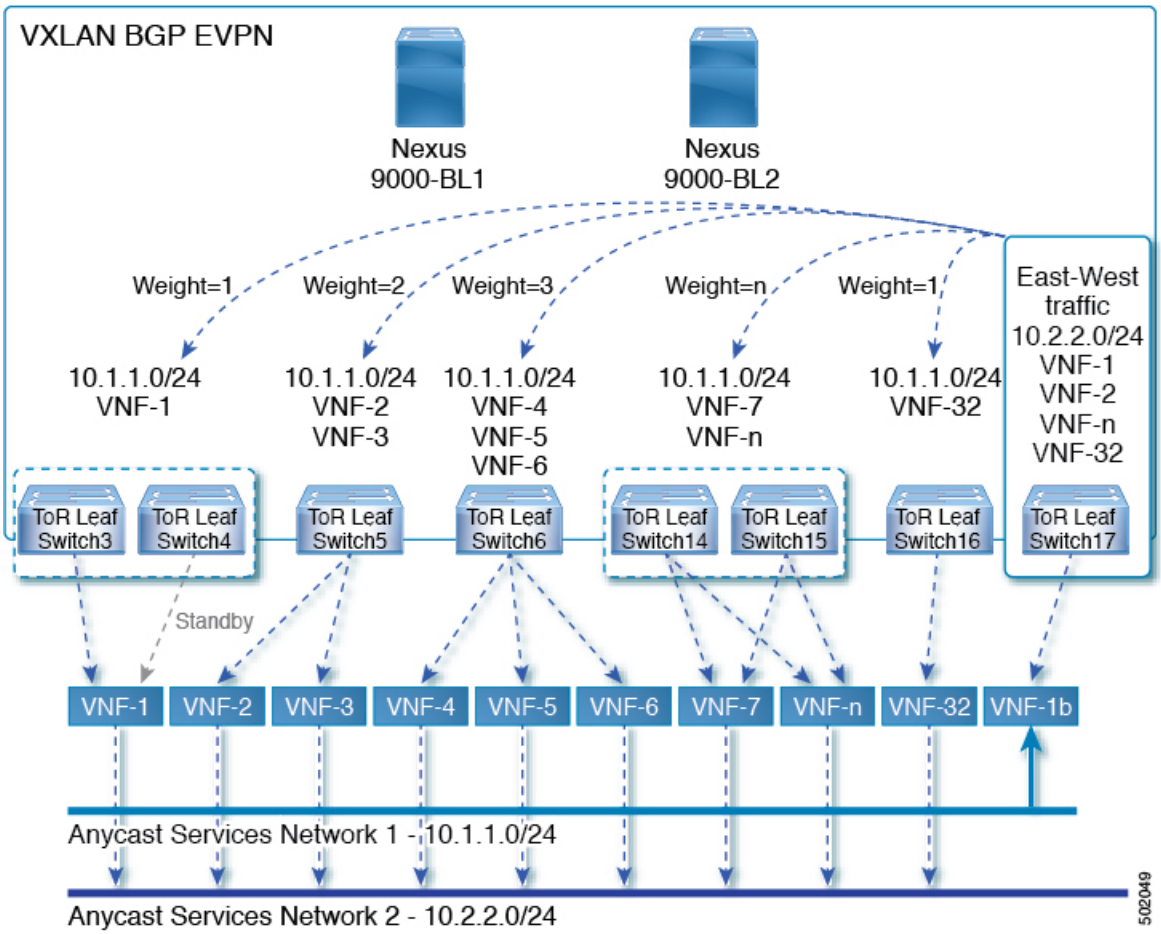
In Network Function Virtualization Infrastructures (NFVi), anycast services networks are advertised from multiple Virtual Network Functions (VNFs). The Proportional Multipath for VNF feature enables advertising of all the available next hops to a given destination network. This feature enables the switch to consider all paths to a given route as equal cost multipath (ECMP) allowing the traffic to be forwarded using all the available links stretched across multiple ToRs.

Figure 66: Sample Topology (North-South Traffic)



In the preceding diagram, North-South traffic that enters the VXLAN fabric at a border leaf is sent across all egress endpoints with the traffic forwarded proportional to the number of links from the egress top of rack (ToR) to the destination network.

Figure 67: Sample Topology (East-West Traffic)



East-West traffic is forwarded between the VXLAN Tunnel Endpoints (VTEPs) proportional to the number of next hops advertised by each ToR switch to the destination network.

The switch uses BGP to advertise reachability within the fabric using the Layer 2 VPN (L2VPN)/Ethernet VPN (EVPN) address family. If all ToR switches and border leafs are within the same Autonomous System (AS), a full internal BGP (iBGP) mesh is configured by using route reflectors or by having each BGP router peer with every other router.

Each ToR and border leaf constitutes a VTEP in the VXLAN fabric. You can use a BGP route reflector to reduce the full mesh BGP sessions across the VTEPs to a single BGP session between a VTEP and the route reflector. Virtual Network Identifiers (VNIs) are globally unique within the overlay. Each Virtual Routing and Forwarding (VRF) instance is mapped to a unique VNI. The inner destination MAC address in the VXLAN header belongs to the receiving VTEP that does the routing of the VXLAN payload. This MAC address is distributed as a BGP attribute along with the EVPN routes.

Advertisement of customer networks

Customer networks are configured statically or learned locally by using an interior gateway protocol, (IGP) or external BGP (eBGP), over a Provider Edge(PE)-Customer Edge(CE) link. These networks are redistributed into BGP and advertised to the VXLAN fabric.

The networks advertised to the ToRs by the virtual machines (VMs) attached to them are advertised to the VXLAN fabric as EVPN Type-5 routes with the following:

- The route distinguisher (RD) will be the Layer 3 VNI's configured RD.
- The gateway IP field will be populated with the next hop.
- The next hop of the EVPN route will continue to be the VTEP IP.
- The export route targets of the routes will be derived from the configured export route targets of the associated Layer 3 VNI.

Multiple VRF routes may generate the same Type-5 Network Layer Reachability Information (NLRI) differentiated only by the gateway IP field. The routes are advertised with the L3VNI's RD, and the gateway IP isn't part of the Type-5 NLRI's key. The NLRI is exchanged between BGP routers using update messages. These routes are advertised to the EVPN AF by extending the BGP export mechanism to include ECMPs and using the addpath BGP feature in the EVPN AF.

Each Type-5 route within the EVPN AF that is created by using the Proportional Multipath for VNF feature may have multiple paths that are imported into the corresponding VRF based on the matching of the received route targets and by having ECMP enabled within the VRF and in the EVPN AF. Within the VRF, the route is a single prefix with multiple paths. Each path represents a Type-5 EVPN path or those learned locally within the VRF. The EVPN Type-5 routes that are enabled for the Proportional Multipath for VNF feature will have their next hop in the VRF derived from their gateway IP field. Use the **export-gateway-ip** command to enable BGP to advertise the gateway IP in the EVPN Type-5 routes.

Use the **maximum-paths mixed** command to enable BGP and the Unicast Routing Information Base (URIB) to consider the following paths as ECMP:

- iBGP paths
- eBGP paths
- Paths from other protocols (such as static) that are redistributed or injected into BGP

The paths can be either local to the device (static, iBGP, or eBGP) or remote (eBGP or iBGP learned over BGP-EVPN). This overrides the default route selection behavior in which local routes are preferred over remote routes. URIB downloads all next hops of the route, including locally learned and user-configured routes, to the Unicast FIB Distribution Module (uFDM)/Forwarding Information Base (FIB).

Beginning with Cisco NX-OS Release 9.3(5), you don't need to use mixed paths. You can choose to have only eBGP or iBGP filter the ECMP paths.

When you enter the **maximum-paths mixed** command beginning with Cisco NX-OS Release 9.3(5), BGP checks for the AS-path length by default. If you want to ignore the AS-path length (for example, on nodes that participate in packet forwarding such as BGWs and VTEPs), you must enter the **bestpath as-path ignore** command. When the **maximum-paths mixed** command is enabled for earlier releases, BGP ignores the AS-path length, and URIB ignores the administrative distance when choosing ECMPs. To ensure that no impact is observed, we recommend upgrading to Cisco NX-OS Release 9.3(5) prior to entering this command.

Legacy peer support

Use the **advertise-gw-ip** command to advertise EVPN Type-5 routes with the gateway IP set. ToRs then advertise the gateway IP in the Type-5 NLRI. However, legacy peers running on NX-OS version older than Cisco NX-OS Release 9.2(1) can't process the gateway IP which might lead to unexpected behavior. To prevent this scenario from occurring, use the **no advertise-gw-ip** command to disable the Proportional

Multipath for VNF feature for a legacy peer. BGP sets the gateway IP field of the Type-5 NLRI to zero even if the path being advertised has a valid gateway IP.

The **no advertise-gw-ip** command flaps the specified peer session as gracefully as possible. The remote peer triggers a graceful restart if the peer supports this capability. When the session is re-established, the local peer advertises EVPN Type-5 routes with the gateway IP set or with the gateway IP as zero depending on whether the **advertise-gw-ip** command has been used. By default, this knob is enabled and the gateway IP field is populated with the appropriate next hop value.

Next-hop filtering

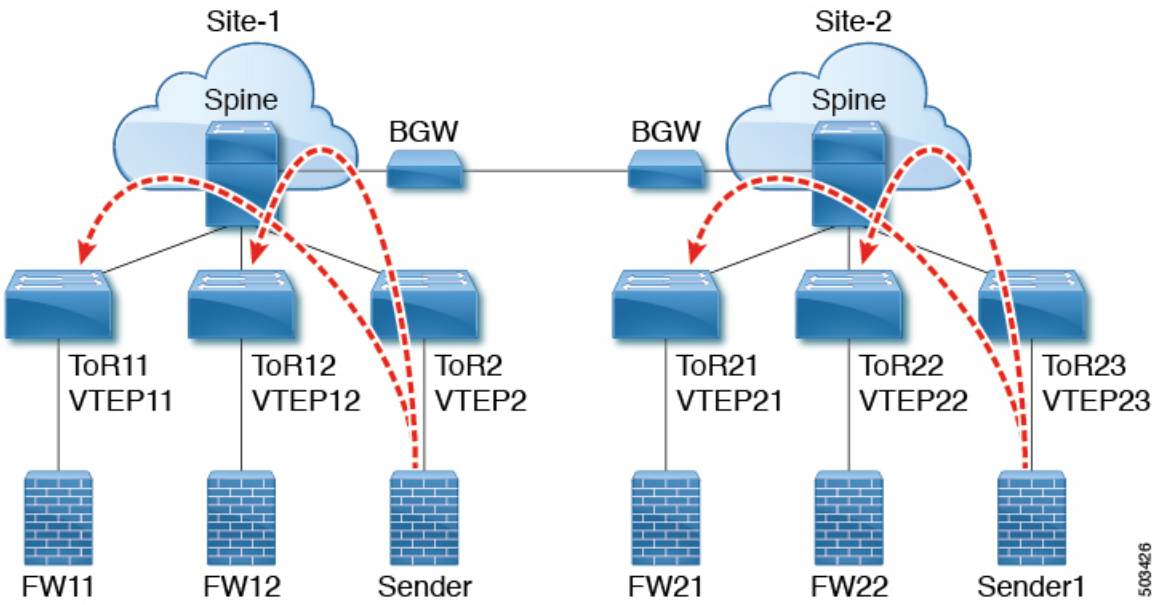
You can configure next-hop filtering to exclude temporarily silent next-hops, but you should know the subnet IPs, SVI, and interfaces to which the VNF is attached. You can configure next-hop filtering and use this as a safeguard when a VNF does not exist. However, activate the next-hop filtering only when the VNF's IP ARP is resolved.

Proportional Multipath for VNF with Multi-Site

Cisco NX-OS Release 9.3(6) and later releases support Proportional Multipath for VNF with Multi-Site. This feature allows traffic to be sent across sites if a local VNF isn't available.

ToRs prefer to use local VNFs. However, if local VNFs aren't available, they can use VNFs in a different site. In the following topology, the ToRs in site 2 would use VNFs 21 and 22. However, if these VNFs aren't available, sender 1 in site 2 could send traffic to VNFs 11 and 12 in site 1.

Figure 68: Proportional Multipath for VNF with Multi-Site Topology



To use this feature, simply configure Proportional Multipath for VNF and enable Multi-Site. For a sample configuration, see Configuration Example for Proportional Multipath for VNF with Multi-Site, on page 640.

Prerequisites for Proportional Multipath for VNF

If desired, take the following actions before upgrading to Cisco NX-OS Release 9.3(5):

- Configure a route map for redistributed paths and use the **set ip next-hop redist-unchanged** command when using locally redistributed paths to export the gateway IP address. This command preserves the next hop for locally redistributed paths. For example:

```
route-map redist-rtmap permit 10
match ip prefix-list vm-pfx-list
set ip next-hop redist-unchanged
```

- Enter the **bestpath as-path ignore** command on nodes that participate in packet forwarding, such as BGWs and VTEPs. This command causes BGP to ignore the AS-path length.

Guidelines and limitations for proportional multipath for VNF

Proportional Multipath for VNF has the following guidelines and limitations:

- If the Proportional Multipath for VNF feature is enabled, maintenance mode isolation doesn't work because BGP installs all the paths in mixed multipath mode. Alternatively, a route-map is used to deny outbound BPG updates when a switch goes into maintenance mode by using user-defined profiles.
- This feature is supported for Cisco Nexus 9300-FX/FX2/FX3 platform switches and Cisco Nexus 9500 platform switches with the N9K-C9508-FM-E2 fabric module and an X97160YC-EX and 9700-FX line card.
- Beginning with Cisco NX-OS Release 10.2(3)F, the Proportional Multipath for VNF feature is supported on Cisco Nexus 9300-GX/GX2B platform switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, the Proportional Multipath for VNF feature is supported on Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, the Proportional Multipath for VNF feature is supported on Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, the Proportional Multipath for VNF feature is supported on Cisco Nexus 9364C-H1 switches.
- Beginning with Cisco NX-OS Release 10.5(2)F, the Proportional Multipath for VNF feature is supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.
- Static and direct routes have to be redistributed into the BGP when the Proportional Multipath for VNF feature is enabled.
- If OSPF or EIGRP is being used as an IGP, routes can't be redistributed into BGP.
- If Proportional Multipath for VNF is enabled and routes aren't redistributed into BGP, asymmetric load balancing of traffic may occur as the local routes from URIB may not show up in BGP and on remote TORs as EVPN paths.
- Devices on which mixed-multipath is enabled must support the same load-balancing algorithm.

- If a VNF instance is multi-homed to multiple TORs, policies have to be configured or BGP routes have to be originated using a network command. As a result, each TOR connection to the VNF is displayed in the BGP routing table. Each TOR can now see the VNF's direct routes to the other TORs in which the VNF is multi-homed. Consequently, each TOR can advertise paths to the Gateway IPs through other TORs leading to a next hop resolution loop.

  Consider a scenario in which a VNF is multi-homed to two TORs, TOR1 and TOR2. Individual links to the TORs are addressed as 1.1.1.1 and 2.2.2.2. If the VNF advertises a service 192.168.1.0/24 through the TORs, the TORs advertise EVPN routes to 192.168.1.0/24 with Gateway IPs of 1.1.1.1 and 2.2.2.2 respectively.

  As a result, an issue occurs with the Recursive Next Hop (RNH) resolution on a remote TOR (for example, TOR3). The gateway IP is resolved to a /24 route pointing to another gateway IP. That second gateway IP is resolved by a route pointing to the first gateway IP. So, in our scenario, the gateway IP 1.1.1.1 is resolved by 1.1.1.0/24 which points to 2.2.2.2. And 2.2.2.2 is resolved by 2.2.2.0/24 which points to 1.1.1.1.

  This condition occurs as both TORs connected to the VNF are advertising the VNF’s connected routes. TOR1 is advertising 1.1.1.0/24 and 2.2.2.0/24. However, 1.1.1.0 is advertised without a gateway IP as it's a connected subnet on TOR1. Also, 2.2.2.0 is an OSPF route pointing to 1.1.1.1 which is the VNF’s address connected to TOR1.

  Similarly, TOR2 advertises both subnets and 2.2.2.0/24 is sent without a gateway IP as it is directly connected to TOR2. 1.1.1.0 is learned via OSPF and is sent with a gateway IP of 2.2.2.2 which is the VNF’s address connected to TOR2. 1.1.1.1/32 and 2.2.2.2/32 won't be advertised as they are Adjacency Manager (AM) routes on each TOR.

  This issue doesn't have a resolution when Type-5 routes are involved. However, this scenario can be avoided if the TORs advertise the gateway IP’s /32 address using a network command. And if the gateway IPs are being resolved by Type-2 EVPN MAC/IP routes, this scenario can be avoided as the gateway IP will be resolved by the /32 IP route.

- The following guidelines and limitations apply to Proportional Multipath for VNF with Multi-Site:
  - This feature is supported for Cisco Nexus 9300-FX/FX2/FX3 platform switches and Cisco Nexus 9500 platform switches with the N9K-C9508-FM-E2 fabric module and X97160YC-EX, 9700-FX or -FX3 line card.
  - VNF moves across sites aren't supported.

- Proportional multipath with max-path mixed configuration is not supported for VNFs attached to vPC leaf switches, and vPC is supported when the max-path mixed configuration is not used. However, beginning with Cisco NX-OS Release 10.5(2)F, proportional multipath with max-path mixed configuration is supported, and these are the guidelines and limitations.
  - Multi-hop recursion is not supported for mixed-path with silent host. Next hop filtering can be used, but you should resolve the VNF before sending service IP over silent host.
  - Maximum mixed-path config should be the same across all the VTEPs.
  - Configure the number or value for the maximum mixed-path so that it is more than the available VNFs.
  - Ensure that the VNF is reachable or resolvable. If a VNF does not exist but is still used by service IPs or VM routes, then traffic loss occurs.

- For VNF learned over L2VNI, Border Gateways should configure the L2VNI, if **export-gw-ip** is configured in the Border Gateway.
- BGP Peering to learn VNF as GWIP for a service route or VM route cannot be across multi-site.
- VNF cannot be attached to BGW.
- If the VNFs are learned from remote VTEP, static routes and IGP routes cannot use next hops as VNFs .
- For a vPC setup, enable **advertise-pip**.
- Any one of the two options is possible at a time:
  - Advertise same prefix advertised to two VRFs
  - Leaking between the VRFs
- This feature does not support
  - fabric peering or vMCT with mixed-path,
  - BGP underlay with eBGP fabric,
  - VxLANv6, and
  - ESI multihoming.

Configuring the Spine or Route Reflector

SUMMARY STEPS

1. **configure terminal**
2. **router bgp** *number*
3. **address-family l2vpn evpn**
4. **additional-paths send**
5. **additional-paths receive**
6. **additional-paths selection route-map passall**
7. **route-map passall permit** *seq-num*
8. **set path-selection all advertise**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.

	Command or Action	Purpose
Step 2	**router bgp** *number* **Example:** `switch(config)# router bgp 2`	Configure BGP.
Step 3	**address-family l2vpn evpn** **Example:** `switch(config-router)# address-family l2vpn evpn`	Configure address family Layer 2 VPN EVPN under **router bgp** context.
Step 4	**additional-paths send** **Example:** `switch(config-router-af)# additional-paths send`	The additional-paths configuration for sending..
Step 5	**additional-paths receive** **Example:** `switch(config-router-af)# additional-paths receive`	The additional-paths configuration for receiving.
Step 6	**additional-paths selection route-map passall** **Example:** `switch(config-router-af)# additional-paths selection route-map passall`	The additional-paths configuration applied the route map.
Step 7	**route-map passall permit** *seq-num* **Example:** `switch(config)# route-map passall permit 10`	Configure the route map.
Step 8	**set path-selection all advertise** **Example:** `switch(config-route-map)# set path-selection all advertise`	Sets the route-map related to the additional-paths feature.

Configuring the Leaf or ToR

This procedure describes how to configure the ToR.

SUMMARY STEPS

1. **configure terminal**
2. **router bgp** *number*
3. **address-family l2vpn evpn**
4. [**no**] **maximum-paths** [*eBGP max-paths* |**mixed** | **ibgp** |**local** | **eibgp**] *mpath-count*
5. **additional-paths send**
6. **additional-paths receive**
7. **additional-paths selection route-map passall**

8. **exit**
9. **vrf** *vrf-name*
10. **address-family ipv4 unicast**
11. **export-gateway-ip**
12. [**no**] **maximum-paths** [*eBGP max-paths* |**mixed** | **ibgp** |**local** | **eibgp**] *mpath-count*
13. **redistribute static route-map** *route-map-name*
14. **maximum-paths local** *number*
15. **exit**
16. **address-family ipv6 unicast**
17. **export-gateway-ip**
18. [**no**] **maximum-paths** [*eBGP max-paths* |**mixed** | **ibgp** |**local** | **eibgp**] *mpath-count*
19. **redistribute static route-map** *route-map-name*
20. **maximum-paths local** *number*
21. **exit**
22. **route-map passall permit** *seq-num*
23. **set path-selection all advertise**
24. **ip load-sharing address source-destination rotate** *rotate* **universal-id** *seed*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.
Step 2	**router bgp** *number* **Example:** `switch(config)# router bgp 2`	Configure BGP.
Step 3	**address-family l2vpn evpn** **Example:** `switch(config-router)# address-family l2vpn evpn`	Configure address family Layer 2 VPN EVPN under **router bgp** context.

Step 4	[**no**] **maximum-paths** [*eBGP max-paths* |**mixed**	**ibgp** |**local**	**eibgp**] *mpath-count*
			Example:
			`switch(config-router-af)# maximum-paths ?`
			`  <1-64>  Number of parallel paths`
			`  *Default value is 1`
			`  eibgp  Configure multipath for both EBGP and IBGP paths`
			`  ibgp  Configure multipath for IBGP paths`
			`  local  Configure multipath for local paths`
				The following options are available:
				• *eBGP max-path*–Enables the eBGP maximum paths. The range is from 1 to 64 parallel paths. The default value is 1.
				• **mixed**–Enables BGP and the Unicast Routing Information Base (URIB) to consider the following paths as Equal Cost Multi Path (ECMP):
				  • eBGP paths

	Command or Action	Purpose
	mixed Configure multipath for local and remote paths switch(config-router-af)# **maximum-paths mixed 32** **Example:** switch(config-router-af)# **maximum-paths ibgp 32**	• eiBGP paths • iBGP paths • Paths from other protocols (such as static) that are redistributed or injected into BGP • **ibgp**–Uses iBGP to filter the ECMP paths. • **local**–Enables the multipath for local paths. • If you enter the command without the **mixed** or **ibgp** option, eBGP is used to filter the ECMP paths. **Note** Use the **no** form of this command if you want to use a single path instead of maximum paths.
Step 5	**additional-paths send** **Example:** switch(config-router-af)# **additional-paths send**	The additional-paths configuration for sending.
Step 6	**additional-paths receive** **Example:** switch(config-router-af)# **additional-paths receive**	The additional-paths configuration for receiving.
Step 7	**additional-paths selection route-map passall** **Example:** switch(config-router-af)# **additional-paths selection route-map passall**	The additional-paths configuration applied the route map.
Step 8	**exit** **Example:** switch(config-router-af)# **exit**	Exits command mode.
Step 9	**vrf** *vrf-name* **Example:** switch(config-router)# **vrf evpn-tenant-1001**	Switch to the VRF configuration mode.
Step 10	**address-family ipv4 unicast** **Example:** switch(config-router)# **address-family ipv4 unicast**	Configure address family for IPv4.
Step 11	**export-gateway-ip** **Example:** switch(config-router-vrf-af)# **export-gateway-ip**	Enables BGP to advertise the gateway IP in the EVPN Type-5 routes. It exports the gateway IP for all prefixes in that VRF. **Note**

	Command or Action	Purpose

		If you want choose specific prefixes for which to export the gateway IP, use the following configuration instead of the **export-gateway-ip** command:
		`route-map name permit sequence`
		`  match ip address prefix-list name`
		`  set evpn gateway-ip use-next-hop`
		
		`vrf context vrf`
		`  address-family ipv4 unicast`
		`  export map name`

Step 12	[**no**] **maximum-paths** [*eBGP max-paths* |**mixed** | **ibgp** |**local** | **eibgp**] *mpath-count*
	Example:
	`switch(config-router-vrf-af)# maximum-paths ?`
	`  <1-64>  Number of parallel paths`
	`  *Default value is 1`
	`  eibgp  Configure multipath for both EBGP and IBGP paths`
	`  ibgp  Configure multipath for IBGP paths`
	`  local  Configure multipath for local paths`
	`  mixed  Configure multipath for local and remote paths`
	`switch(config-router-vrf-af)# maximum-paths mixed 32`
	Example:
	`switch(config-router-vrf-af)# maximum-paths ibgp 32`
		The following options are available:
		• *eBGP max-path*–Enables the eBGP maximum paths. The range is from 1 to 64 parallel paths. The default value is 1.
		• **mixed**–Enables BGP and the Unicast Routing Information Base (URIB) to consider the following paths as Equal Cost Multi Path (ECMP):
		  • eBGP paths
		  • eiBGP paths
		  • iBGP paths
		  • Paths from other protocols (such as static) that are redistributed or injected into BGP
		• **ibgp**–Uses iBGP to filter the ECMP paths.
		• **local**–Enables the multipath for local paths.
		• If you enter the command without the **mixed** or **ibgp** option, eBGP is used to filter the ECMP paths.
		Note
		Use the **no** form of this command if you want to use a single path instead of maximum paths.

Step 13	**redistribute static route-map** *route-map-name* **Example:** `switch(config-router-vrf-af)# redistribute static route-map redist-rtmap`	Preserves the next-hop of the redistributed paths.
Step 14	**maximum-paths local** *number* **Example:** `switch(config-router-vrf-af)# maximum-paths local 32`	Specifies the number of local paths to be redistributed as the BGP best path for a route. The range is from 0 to 32. The default value is 1. **Note**

	Command or Action	Purpose
		This command isn't supported with the **maximum-paths mixed** *mpath-count* command. An error message appears if you try to configure them together. **Note** The **set ip next-hop redist-unchanged** command is required in order for the **maximum-paths local** command to work.
Step 15	**exit** **Example:** `switch(config-router-vrf-af)# exit`	Exits command mode.
Step 16	**address-family ipv6 unicast** **Example:** `switch(config-router-vrf)# address-family ipv6 unicast`	Configure address family for IPv6.

Step 17	**export-gateway-ip**
	Example:
	`switch(config-router-vrf-af)# export-gateway-ip`
		Enables BGP to advertise the gateway IP in the EVPN Type-5 routes. It exports the gateway IP for all prefixes in that VRF.
		Note
		If you want choose specific prefixes for which to export the gateway IP, use the following configuration instead of the **export-gateway-ip** command:
		`route-map name permit sequence`
		`  match ip address prefix-list name`
		`  set evpn gateway-ip use-next-hop`
		
		`vrf context vrf`
		`  address-family ipv4 unicast`
		`  export map name`

Step 18	[**no**] **maximum-paths** [*eBGP max-paths* |**mixed**	**ibgp** |**local**	**eibgp**] *mpath-count*
			Example:
			`switch(config-router-vrf-af)# maximum-paths ?`
			`  <1-64>  Number of parallel paths`
			`  *Default value is 1`
			`  eibgp  Configure multipath for both EBGP and IBGP paths`
			`  ibgp  Configure multipath for IBGP paths`
			`  local  Configure multipath for local paths`
			`  mixed  Configure multipath for local and remote paths`
			`switch(config-router-vrf-af)# maximum-paths mixed 32`
			Example:
				The following options are available:
				• *eBGP max-path*–Enables the eBGP maximum paths. The range is from 1 to 64 parallel paths. The default value is 1.
				• **mixed**–Enables BGP and the Unicast Routing Information Base (URIB) to consider the following paths as Equal Cost Multi Path (ECMP):
				  • eBGP paths
				  • eiBGP paths
				  • iBGP paths
				  • Paths from other protocols (such as static) that are redistributed or injected into BGP

	Command or Action	Purpose
	`switch(config-router-vrf-af)# maximum-paths ibgp 32`	• **ibgp**–Uses iBGP to filter the ECMP paths. • **local**–Enables the multipath for local paths. • If you enter the command without the **mixed** or **ibgp** option, eBGP is used to filter the ECMP paths. **Note** Use the **no** form of this command if you want to use a single path instead of maximum paths.
Step 19	**redistribute static route-map** *route-map-name* **Example:** `switch(config-router-vrf-af)# redistribute static route-map redist-rtmap`	Preserves the next-hop of the redistributed paths.
Step 20	**maximum-paths local** *number* **Example:** `switch(config-router-vrf-af)# maximum-paths local 32`	Specifies the number of local paths to be redistributed as the BGP best path for a route. The range is from 0 to 32. The default value is 1. **Note** This command isn't supported with the **maximum-paths mixed** *mpath-count* command. An error message appears if you try to configure them together.
Step 21	**exit** **Example:** `switch(config-router-vrf-af)# exit`	Exits command mode.
Step 22	**route-map passall permit** *seq-num* **Example:** `switch(config)# route-map passall permit 10`	Configure the route map.
Step 23	**set path-selection all advertise** **Example:** `switch(config-route-map)# set path-selection all advertise`	Sets the route-map related to the additional-paths feature.
Step 24	**ip load-sharing address source-destination rotate** *rotate* **universal-id** *seed* **Example:** `ip load-sharing address source-destination rotate 32 universal-id 1`	Configures the unicast FIB load-sharing algorithm for data traffic. • The **universal-id** option sets the random seed for the hash algorithm and shifts the flow from one link to another. You do not need to configure the universal ID. Cisco NX-OS chooses the Universal ID if you do not configure it. The *seed* range is from 1 to 4294967295.

	Command or Action	Purpose
		• The **rotate** option causes the hash algorithm to rotate the link picking selection so that it does not continually choose the same link across all nodes in the network. It does so by influencing the bit pattern for the hash algorithm. This option shifts the flow from one link to another and load balances the already load-balanced (polarized) traffic from the first ECMP level across multiple links. If you specify a **rotate** value, the 64-bit stream is interpreted starting from that bit position in a cyclic rotation. The **rotate** range is from 1 to 63, and the default is 32. **Note** With multi-tier Layer 3 topology, polarization is possible. To avoid polarization, use a different rotate bit at each tier of the topology. **Note** To configure a rotation value for port channels, use the **port-channel load-balance src-dst ip-l4port rotate** *rotate* command. For more information on this command, see the *Cisco Nexus 9000 Series NX-OS Interfaces Configuration Guide, Release 9.x*.

Configuring the Border Leaf or Border Gateway

This procedure describes how to configure the border leaf.

SUMMARY STEPS

1. **configure terminal**
2. **router bgp** *number*
3. **address-family l2vpn evpn**
4. [**no**] **maximum-paths** [*eBGP max-paths* |**mixed** | **ibgp** |**local** | **eibgp**] *mpath-count*
5. **additional-paths send**
6. **additional-paths receive**
7. **additional-paths selection route-map passall**
8. **exit**
9. **vrf** *vrf-name*
10. **address-family ipv4 unicast**
11. **export-gateway-ip**
12. [**no**] **maximum-paths** [*eBGP max-paths* |**mixed** | **ibgp** |**local** | **eibgp**] *mpath-count*
13. **redistribute static route-map** *route-map-name*
14. **maximum-paths local** *number*

15. **address-family ipv6 unicast**
16. **export-gateway-ip**
17. [**no**] **maximum-paths** [*eBGP max-paths* |**mixed** | **ibgp** |**local** | **eibgp**] *mpath-count*
18. **redistribute static route-map** *route-map-name*
19. **maximum-paths local** *number*
20. **exit**
21. **route-map passall permit** *seq-num*
22. **set path-selection all advertise**
23. **ip load-sharing address source-destination rotate** *rotate* **universal-id** *seed*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.
Step 2	**router bgp** *number* **Example:** `switch(config)# router bgp 2`	Configure BGP.
Step 3	**address-family l2vpn evpn** **Example:** `switch(config-router)# address-family l2vpn evpn`	Configure address family Layer 2 VPN EVPN under **router bgp** context.

Step 4	[**no**] **maximum-paths** [*eBGP max-paths* |**mixed**	**ibgp** |**local**	**eibgp**] *mpath-count*
			Example:
			`switch(config-router-af)# maximum-paths ?`
			`  <1-64>  Number of parallel paths`
			`  *Default value is 1`
			`  eibgp   Configure multipath for both EBGP and IBGP paths`
			`  ibgp    Configure multipath for IBGP paths`
			`  local   Configure multipath for local paths`
			`  mixed   Configure multipath for local and remote paths`
			`switch(config-router-af)# maximum-paths mixed 32`
			Example:
			`switch(config-router-af)# maximum-paths ibgp 32`
				The following options are available:
				• *eBGP max-path*–Enables the eBGP maximum paths. The range is from 1 to 64 parallel paths. The default value is 1.
				• **mixed**–Enables BGP and the Unicast Routing Information Base (URIB) to consider the following paths as Equal Cost Multi Path (ECMP):
				  • eBGP paths
				  • eiBGP paths
				  • iBGP paths
				  • Paths from other protocols (such as static) that are redistributed or injected into BGP
				• **ibgp**–Uses iBGP to filter the ECMP paths.
				• **local**–Enables the multipath for local paths.

	Command or Action	Purpose
		• If you enter the command without the **mixed** or **ibgp** option, eBGP is used to filter the ECMP paths. **Note** Use the **no** form of this command if you want to use a single path instead of maximum paths.
Step 5	**additional-paths send** **Example:** `switch(config-router-af)# additional-paths send`	The additional-paths configuration for sending.
Step 6	**additional-paths receive** **Example:** `switch(config-router-af)# additional-paths receive`	The additional-paths configuration for receiving.
Step 7	**additional-paths selection route-map passall** **Example:** `switch(config-router-af)# additional-paths selection route-map passall`	The additional-paths configuration enables the additional-paths feature.
Step 8	**exit** **Example:** `switch(config-router-af)# exit`	Exits command mode.
Step 9	**vrf** *vrf-name* **Example:** `switch(config-router)# vrf evpn-tenant-1001`	Switch to the VRF configuration mode.
Step 10	**address-family ipv4 unicast** **Example:** `switch(config-router)# address-family ipv4 unicast`	Configure address family for IPv4.

Step 11	**export-gateway-ip**
	Example:
	`switch(config-router-vrf-af)# export-gateway-ip`
		Enables BGP to advertise the gateway IP in the EVPN Type-5 routes. It exports the gateway IP for all prefixes in that VRF.
		Note
		If you want choose specific prefixes for which to export the gateway IP, use the following configuration instead of the **export-gateway-ip** command:
		`route-map name permit sequence`
		`  match ip address prefix-list name`
		`  set evpn gateway-ip use-next-hop`
		
		`vrf context vrf`
		`  address-family ipv4 unicast`
		`  export map name`

	Command or Action	Purpose

Step 12	[**no**] **maximum-paths** [*eBGP max-paths* |**mixed**	**ibgp** |**local**	**eibgp**] *mpath-count*
			Example:
			`switch(config-router-af)# maximum-paths ?`
			`<1-64> Number of parallel paths`
			`*Default value is 1`
			`eibgp Configure multipath for both EBGP and IBGP paths`
			`ibgp Configure multipath for IBGP paths`
			`local Configure multipath for local paths`
			`mixed Configure multipath for local and remote paths`
			`switch(config-router-vrf-af)# maximum-paths mixed 32`
			Example:
			`switch(config-router-vrf-af)# maximum-paths ibgp 32`
				The following options are available:
				• *eBGP max-path*–Enables the eBGP maximum paths. The range is from 1 to 64 parallel paths. The default value is 1.
				• **mixed**–Enables BGP and the Unicast Routing Information Base (URIB) to consider the following paths as Equal Cost Multi Path (ECMP):
				  • eBGP paths
				  • eiBGP paths
				  • iBGP paths
				  • Paths from other protocols (such as static) that are redistributed or injected into BGP
				• **ibgp**–Uses iBGP to filter the ECMP paths.
				• **local**–Enables the multipath for local paths.
				• If you enter the command without the **mixed** or **ibgp** option, eBGP is used to filter the ECMP paths.
				Note
				Use the **no** form of this command if you want to use a single path instead of maximum paths.

Step 13	**redistribute static route-map** *route-map-name* **Example:** `switch(config-router-vrf-af)# redistribute static route-map redist-rtmap`	Preserves the next-hop of the redistributed paths.
Step 14	**maximum-paths local** *number* **Example:** `switch(config-router-vrf-af)# maximum-paths local 32`	Specifies the number of local paths to be redistributed as the BGP best path for a route. The range is from 0 to 32. The default value is 1. **Note** This command isn't supported with the **maximum-paths mixed** *mpath-count* command. An error message appears if you try to configure them together.
Step 15	**address-family ipv6 unicast** **Example:** `switch(config-router-vrf)# address-family ipv6 unicast`	Configure address family for IPv6.

	Command or Action	Purpose

Step 16	**export-gateway-ip**
	Example:
	`switch(config-router-vrf-af)# export-gateway-ip`
		Enables BGP to advertise the gateway IP in the EVPN Type-5 routes. It exports the gateway IP for all prefixes in that VRF.
		Note
		If you want choose specific prefixes for which to export the gateway IP, use the following configuration instead of the **export-gateway-ip** command:
		`route-map name permit sequence`
		`  match ip address prefix-list name`
		`  set evpn gateway-ip use-next-hop`
		`vrf context vrf`
		`  address-family ipv4 unicast`
		`    export map name`

Step 17	[**no**] **maximum-paths** [*eBGP max-paths* |**mixed** | **ibgp** |**local** | **eibgp**] *mpath-count*
	Example:
	`switch(config-router-vrf-af)# maximum-paths ?`
	`  <1-64>  Number of parallel paths`
	`  *Default value is 1`
	`  eibgp   Configure multipath for both EBGP and IBGP paths`
	`  ibgp    Configure multipath for IBGP paths`
	`  local   Configure multipath for local paths`
	`  mixed   Configure multipath for local and remote paths`
	`switch(config-router-vrf-af)# maximum-paths mixed 32`
	Example:
	`switch(config-router-vrf-af)# maximum-paths ibgp 32`
		The following options are available:
		• *eBGP max-path*–Enables the eBGP maximum paths. The range is from 1 to 64 parallel paths. The default value is 1.
		• **mixed**–Enables BGP and the Unicast Routing Information Base (URIB) to consider the following paths as Equal Cost Multi Path (ECMP):
		  • eBGP paths
		  • eiBGP paths
		  • iBGP paths
		  • Paths from other protocols (such as static) that are redistributed or injected into BGP
		• **ibgp**–Uses iBGP to filter the ECMP paths.
		• **local**–Enables the multipath for local paths.
		• If you enter the command without the **mixed** or **ibgp** option, eBGP is used to filter the ECMP paths.
		Note
		Use the **no** form of this command if you want to use a single path instead of maximum paths.

Step 18	**redistribute static route-map** *route-map-name* **Example:** `switch(config-router-vrf-af)# redistribute static route-map redist-rtmap`	Preserves the next-hop of the redistributed paths.

	Command or Action	Purpose
Step 19	**maximum-paths local** *number* **Example:** `switch(config-router-vrf-af)# maximum-paths local 32`	Specifies the number of local paths to be redistributed as the BGP best path for a route. The range is from 0 to 32. The default value is 1. **Note** This command isn't supported with the **maximum-paths mixed** *mpath-count* command. An error message appears if you try to configure them together.
Step 20	**exit** **Example:** `switch(config-router-vrf-af)# exit`	Exits command mode.
Step 21	**route-map passall permit** *seq-num* **Example:** `switch(config)# route-map passall permit 10`	Configure the route map.
Step 22	**set path-selection all advertise** **Example:** `switch(config-route-map)# set path-selection all advertise`	Sets the route-map related to the additional-paths feature.
Step 23	**ip load-sharing address source-destination rotate** *rotate* **universal-id** *seed* **Example:** `ip load-sharing address source-destination rotate 32 universal-id 1`	Configures the unicast FIB load-sharing algorithm for data traffic. • The **universal-id** option sets the random seed for the hash algorithm and shifts the flow from one link to another. You do not need to configure the universal ID. Cisco NX-OS chooses the Universal ID if you do not configure it. The *seed* range is from 1 to 4294967295. • The **rotate** option causes the hash algorithm to rotate the link picking selection so that it does not continually choose the same link across all nodes in the network. It does so by influencing the bit pattern for the hash algorithm. This option shifts the flow from one link to another and load balances the already load-balanced (polarized) traffic from the first ECMP level across multiple links. If you specify a **rotate** value, the 64-bit stream is interpreted starting from that bit position in a cyclic rotation. The **rotate** range is from 1 to 63, and the default is 32. **Note**

	Command or Action	Purpose
		With multi-tier Layer 3 topology, polarization is possible. To avoid polarization, use a different rotate bit at each tier of the topology. **Note** To configure a rotation value for port channels, use the **port-channel load-balance src-dst ip-l4port rotate** *rotate* command. For more information on this command, see the *Cisco Nexus 9000 Series NX-OS Interfaces Configuration Guide, Release 9.x*.

Configuring the BGP Legacy Peer

If you are running a Cisco Nexus Release prior to 9.2(1), follow this procedure to disable sending the gateway IP address to that peer.

SUMMARY STEPS

1. **configure terminal**
2. **router bgp** *number*
3. **neighbor** *address* **remote-as** *number*
4. **address-family l2vpn evpn**
5. **no advertise-gw-ip**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.
Step 2	**router bgp** *number* **Example:** `switch(config)# router bgp 2000000`	Configure BGP.
Step 3	**neighbor** *address* **remote-as** *number* **Example:** `switch(config-router)# neighbor 8.8.8.8 remote-as 2000000`	Define neighbor.
Step 4	**address-family l2vpn evpn** **Example:**	Configure address family Layer 2 VPN EVPN.

	Command or Action	Purpose
	`switch(config-router-neighbor)# address-family l2vpn evpn`	
Step 5	**no advertise-gw-ip** **Example:** `switch(config-router-neighbor-af)# no advertise-gw-ip`	Disables the BGP EVPN Mixed-path and Proportional Layer-3 Multipath feature for a legacy peer.

Configure next-hop filtering

Use this procedure to set up BGP next-hop route filtering for a specific VRF. This configuration is supported on both IPv4 and IPv6. This configuration ensures that any BGP routes with next-hop addresses in the configured IP address and defined range are denied, while all other routes are permitted. The configured interface is part of the specified VRF and has the specified IP address.

SUMMARY STEPS

1. **configure terminal**
2. **ip prefix-list nexthopfilter** [**seq** *seq-value*] [**permit** *network/length*] [**le** *le-value*]
3. **route-map nexthopfilter deny** *sequence-number*
4. **match ip address prefix-list nexthopfilter**
5. **route-map nexthopfilter permit** *sequence-number*
6. **router bgp** *autonomous-system-number*
7. **vrf** *vrf-name*
8. **address-family ip unicast**
9. **next-hop route-map nexthopfilter**
10. **interface** *interface-type*
11. **vrf member** *vrf-name*
12. **ip address** *ip-prefix/length*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.
Step 2	**ip prefix-list nexthopfilter** [**seq** *seq-value*] [**permit** *network/length*] [**le** *le-value*] **Example:** `switch(config)# ip prefix-list nexthopfilter seq 10 permit 111.0.2.0/24 le 31`	Creates a prefix list for next hop filtering. • Selective next-hop route filtering supports prefix length matching or source protocol matching on a per address-family basis.

	Command or Action	Purpose
	`switch(config)# ipv6 prefix-list nexthopfilter6 seq 10 permit 101:1:1::/64 le 127`	• In this example, the nexthopfilter prefix list matches IP addresses in the range **111.0.2.0/24** with a prefix length of up to **/31**.

Step 3	**route-map nexthopfilter deny** *sequence-number*
	Example:
	`switch(config)# route-map nexthopfilter deny 10`
	`switch(config)# route-map nexthopfilter6 deny 10`
		Configures a route map and enters route map configuration mode.
		• In this example, a route map named **nexthopfilter** is created. If there is an IP address match in the next **match** command, the IP address is denied.

Step 4	**match ip address prefix-list nexthopfilter**
	Example:
	`switch(config-route-map)# match ip address prefix-list nexthopfilter`
	`switch(config-route-map)# match ipv6 address prefix-list nexthopfilter6`
		Matches the IP addresses, IPv4 or IPv6, in the specified prefix list.

Step 5	**route-map nexthopfilter permit** *sequence-number*
	Example:
	`switch(config)# route-map nexthopfilter permit 20`
	`switch(config)# route-map nexthopfilter6 permit 20`
		Configures a route map named **nexthopfilter** and enters route map configuration mode.
		• In this example, all other IP addresses are permitted by route map nexthopfilter.

Step 6	**router bgp** *autonomous-system-number* **Example:** `switch(config)# router bgp 100`	Enters router configuration mode and creates a BGP routing process.
Step 7	**vrf** *vrf-name* **Example:** `switch(config-router)# vrf vrf_flow1_1`	Enters VRF configuration mode.

Step 8	**address-family ip unicast**
	Example:
	`switch(config-router-vrf)# address-family ipv4 unicast`
	`switch(config-router-vrf)# address-family ipv6 unicast`
		Specifies the IP address family and enters address family configuration mode.
		• The **unicast** keyword specifies the IPv4 or IPv6 unicast address family.

Step 9	**next-hop route-map nexthopfilter**
	Example:
	`switch (config-router-vrf-af)# nexthop route-map nexthopfilter`
	`switch (config-router-vrf-af)# nexthop route-map nexthopfilter6`
		Permits the next-hop filter route map to selectively define routes to help resolve the next hop.
		Step 6 to Step 9 configurations result in BGP configuration applying the next-hop filter route map to the next-hop addresses for the vrf_flow1_1 VRF in IPv4 or IPv6 unicast address family.

	Command or Action	Purpose
Step 10	**interface** *interface-type* **Example:** `switch (config)# interface vlan11`	Enters interface configuration mode.
Step 11	**vrf member** *vrf-name* **Example:** `switch(config-if)# vrf vrf_flow1_1`	Adds this interface to a VRF.

Step 12	**ip address** *ip-prefix/length*
	Example:
	`switch(config-if)# ip address 111.0.2.2/24`
	`switch(config-if)# ipv6 address 101:1:1::254/64`
		Configures an IPv4 or IPv6 address for this interface. You must do this step after you assign this interface to a VRF.

Verify next-hop filtering

This section provides information about the show commands required to verify next-hop filtering along with the examples. Perform the relevant show commands listed in the table to display the required information about the next-hop filtering feature.

Command	Purpose
show bgp ipv4 unicast <*prefix*> **vrf** <*vrf*> **show bgp ipv6 unicast** <*prefix*> **vrf** <*vrf*>	Displays Border Gateway Protocol (BGP) information for the specific IPv4 or IPv6 unicast prefix within a specified Virtual Routing and Forwarding (VRF) instance.
show bgp ipv4 unicast nexthop-database vrf <*vrf*> **show bgp ipv6 unicast nexthop-database vrf** <*vrf*>	Displays the next-hop database information for Border Gateway Protocol (BGP) routes for IPv4 or IPv6 unicast routes within a specified VRF instance.

Examples of show commands

This section includes examples of show commands for the next-hop filtering feature.

- **show bgp ipv4 unicast** <*prefix*> **vrf** <*vrf*>

```
show bgp ipv4 unicast 12.34.56.100 vrf org-1-internal-vrf
BGP routing table information for VRF org-1-internal-vrf, address family IPv4 Unicast
BGP routing table entry for 12.34.56.100/32, version 1007
Paths: (19 available, best #8)
Flags: (0x80c021a) (high32 0x000060) on xmit-list, is in urib, is best urib route, is
in HW, exported
  vpn: version 3649, (0x00000000100002) on xmit-list
Multipath: Mixed

  Path type: external, path is invalid(rnh not resolved), no labeled nexthop
  AS-Path: 65001 , path sourced external to AS
    25.1.0.101 (inaccessible, metric 4294967295) from 25.0.0.11 (192.0.0.1)
      Origin IGP, MED not set, localpref 100, weight 0
      Received path-id 2
      Extcommunity: RT:65201:2021
```

```
      Discarded Attribute Length 7

       1. Discarded Local-pref Attribute: type 0x5 flag 0x40 length 0x4
        value 0000 0000
```

- **show bgp ipv6 unicast** <*prefix*> **vrf** <*vrf*>

```
show bgp ipv6 unicast 12:34:56::100/128 vrf org-1-internal-vrf
BGP routing table information for VRF org-1-internal-vrf, address family IPv6 Unicast
BGP routing table entry for 12:34:56::100/128, version 807
Paths: (19 available, best #8)
Flags: (0x80c021a) (high32 0x000060) on xmit-list, is in u6rib, is best u6rib route,
is in HW, exported
  vpn: version 3005, (0x00000000100002) on xmit-list
Multipath: Mixed

  Path type: external, path is invalid(rnh not resolved), no labeled nexthop
  AS-Path: 65001 , path sourced external to AS
    2001:25:0:1::101 (inaccessible, metric 4294967295) from 2001:25::11 (194.0.0.1)
      Origin IGP, MED not set, localpref 100, weight 0
      Received path-id 2
      Extcommunity: RT:65201:2021
      Discarded Attribute Length 7

       1. Discarded Local-pref Attribute: type 0x5 flag 0x40 length 0x4
        value 0000 0000
```

- **show bgp ipv4 unicast nexthop-database vrf** <*vrf*>

```
show bgp ipv4 unicast nexthop-database vrf org-1-internal-vrf

Next Hop table for VRF org-1-internal-vrf, address family IPv4 Unicast:
Next-hop trigger-delay(miliseconds)
  Critical: 3000 Non-critical: 10000
IPv4 Next-hop table

IPv4 Unicast Next-hops:

Nexthop: 25.1.0.101, Flags: 0x20, Refcount: 10, IGP cost: 4294967295, Multipath: Yes
IGP Route type: 0, IGP preference: 0
Nexthop is not-attached not-local unreachable not-labeled filtered
Nexthop last resolved: 00:00:25, using 25.1.0.0/16
Metric next advertise: Never
RNH epoch: 6
```

- **show bgp ipv6 unicast nexthop-database vrf** <*vrf*>

```
show bgp ipv6 unicast nexthop-database vrf org-1-internal-vrf

Next Hop table for VRF org-1-internal-vrf, address family IPv6 Unicast:
Next-hop trigger-delay(miliseconds)
  Critical: 3000 Non-critical: 10000
IPv6 Next-hop table

IPv6 Unicast Next-hops:
Nexthop: 2001:25:0:1::101, Flags: 0x20, Refcount: 10, IGP cost: 4294967295, Multipath:
 Yes
IGP Route type: 0, IGP preference: 0
Nexthop is not-attached not-local unreachable not-labeled filtered
Nexthop last resolved: 00:33:32, using 2001:25:0:1::/64
Metric next advertise: Never
RNH epoch: 1
```

Configuring a User-Defined Profile for Maintenance Mode

SUMMARY STEPS

1. **configure terminal**
2. **configure maintenance profile maintenance-mode**
3. **route-map** *name* **deny** *sequence*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.
Step 2	**configure maintenance profile maintenance-mode** **Example:** `switch(config)# configure maintenance profile maintenance-mode`	Configure maintenance mode profile.
Step 3	**route-map** *name* **deny** *sequence* **Example:** `switch(config-mm-profile)# route-map GIR deny 5`	Configure route map. The value of *sequence* is from 0 to 65535. Default is 10.

Configuring a User-Defined Profile for Normal Mode

SUMMARY STEPS

1. **configure terminal**
2. **configure maintenance profile normal-mode**
3. **route-map** *name* **permit** *sequence*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.

	Command or Action	Purpose
Step 2	**configure maintenance profile normal-mode** **Example:** `switch(config)# configure maintenance profile normal-mode`	Configure maintenance mode.
Step 3	**route-map** *name* **permit** *sequence* **Example:** `switch(config-mm-profile)# route-map GIR permit 5`	Configure route map. The value of *sequence* is from 0 to 65535. Default is 10.

Configuring a Default Route Map

SUMMARY STEPS

1. **configure terminal**
2. **route-map** *name* **permit** *sequence*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.
Step 2	**route-map** *name* **permit** *sequence* **Example:** `switch(config-mm-profile)# route-map GIR permit 5`	Configure route map. The value of *sequence* is from 0 to 65535. Default is 10.

Applying a Route Map to a Route Reflector

SUMMARY STEPS

1. **configure terminal**
2. **router bgp** *number*
3. **neighbor** *ip-address*
4. **address-family l2vpn evpn**
5. **route-map** *name* **out**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enter global configuration mode.
Step 2	**router bgp** *number* **Example:** `switch(config)# router bgp 2`	Configure BGP.
Step 3	**neighbor** *ip-address* **Example:** `switch(config-router)# neighbor 10.1.1.1`	Configure the IP address of a BGP neighbor which is the route reflector. *ip-address* can be an IPv4 or IPv6 address or prefix.
Step 4	**address-family l2vpn evpn** **Example:** `switch(config-router-neighbor)# address-family l2vpn evpn`	Configure a Layer 2 VPN EVPN address family.
Step 5	**route-map** *name* **out** **Example:** `switch(config-router-neighbor-af)# route-map GIR out`	Apply the route map to the neighbor route reflector.

Verifying Proportional Multipath for VNF

Command	Purpose
show bgp ipv4 unicast	Displays Border Gateway Protocol (BGP) information for the IPv4 unicast address family.
show bgp l2vpn evpn	Displays BGP information for the Layer-2 Virtual Private Network (L2VPN) Ethernet Virtual Private Network (EVPN) address family.
show ip route	Displays routes from the unicast RIB.
show maintenance profile maintenance-mode	Displays the GIR user-defined profile for the maintenance mode.
show maintenance profile normal-mode	Displays the GIR user-defined profile for the normal mode.

The following example shows how to display BGP information for the L2VPN EVPN address family:

```
switch# show bgp l2vpn evpn 11.1.1.0
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 13.13.13.13:3       //     Remote route
BGP routing table entry for [5]:[0]:[0]:[24]:[11.1.1.0]/224, version 1341
Paths: (3 available, best #1)
Flags: (0x000002) on xmit-list, is not in l2rib/evpn, is not in HW
Multipath: eBGP

  Advertised path-id 1
  Path type: external, path is valid, is best path
             Imported to 2 destination(s)
  Gateway IP: 11.1.1.133
  AS-Path: 2000000 100000 , path sourced external to AS
    11.11.11.11 (metric 5) from 102.102.102.102 (102.102.102.102)
      Origin incomplete, MED not set, localpref 100, weight 0
      Received label 22001
      Received path-id 3
      Extcommunity: RT:23456:22001 Route-Import:11.11.11.11:2001 ENCAP:8
          Router MAC:003a.7d7d.1dbd

  Path type: external, path is valid, not best reason: Neighbor Address, multipath
             Imported to 2 destination(s)
  Gateway IP: 11.1.1.233
  AS-Path: 2000000 100 , path sourced external to AS
    33.33.33.33 (metric 5) from 102.102.102.102 (102.102.102.102)
      Origin incomplete, MED not set, localpref 100, weight 0
      Received label 22001
      Received path-id 2
      Extcommunity: RT:23456:22001 Route-Import:33.33.33.33:2001 ENCAP:8
          Router MAC:e00e.da4a.589d

  Path type: external, path is valid, not best reason: Neighbor Address, multipath
             Imported to 2 destination(s)
  Gateway IP: 11.1.1.100
  AS-Path: 2000000 500000 , path sourced external to AS
    22.22.22.22 (metric 5) from 102.102.102.102 (102.102.102.102)
      Origin incomplete, MED not set, localpref 100, weight 0
      Received label 22001
      Received path-id 1
      Extcommunity: RT:23456:22001 Route-Import:22.22.22.22:2001 ENCAP:8
          Router MAC:e00e.da4a.62a5

  Path-id 1 not advertised to any peer

Route Distinguisher: 4.4.4.4:3    (L3VNI 22001)       //   Local L3VNI
BGP routing table entry for [5]:[0]:[0]:[24]:[11.1.1.0]/224, version 3465
Paths: (3 available, best #1)
Flags: (0x000002) on xmit-list, is not in l2rib/evpn, is not in HW
Multipath: eBGP

  Advertised path-id 1
  Path type: external, path is valid, is best path
             Imported from 13.13.13.13:3:[5]:[0]:[0]:[24]:[11.1.1.0]/224
  Gateway IP: 11.1.1.100
  AS-Path: 2000000 500000 , path sourced external to AS
    22.22.22.22 (metric 5) from 102.102.102.102 (102.102.102.102)
      Origin incomplete, MED not set, localpref 100, weight 0
      Received label 22001
      Received path-id 1
      Extcommunity: RT:23456:22001 Route-Import:22.22.22.22:2001 ENCAP:8
          Router MAC:e00e.da4a.62a5
```

```
  Path type: external, path is valid, not best reason: newer EBGP path, multipat
h
             Imported from 13.13.13.13:3:[5]:[0]:[0]:[24]:[11.1.1.0]/224
  Gateway IP: 11.1.1.233
  AS-Path: 2000000 100 , path sourced external to AS
    33.33.33.33 (metric 5) from 102.102.102.102 (102.102.102.102)
      Origin incomplete, MED not set, localpref 100, weight 0
      Received label 22001
      Received path-id 2
      Extcommunity: RT:23456:22001 Route-Import:33.33.33.33:2001 ENCAP:8
          Router MAC:e00e.da4a.589d

  Path type: external, path is valid, not best reason: newer EBGP path, multipat
h
             Imported from 13.13.13.13:3:[5]:[0]:[0]:[24]:[11.1.1.0]/224
  Gateway IP: 11.1.1.133
  AS-Path: 2000000 100000 , path sourced external to AS
    11.11.11.11 (metric 5) from 102.102.102.102 (102.102.102.102)
      Origin incomplete, MED not set, localpref 100, weight 0
      Received label 22001
      Received path-id 3
      Extcommunity: RT:23456:22001 Route-Import:11.11.11.11:2001 ENCAP:8
          Router MAC:003a.7d7d.1dbd

  Path-id 1 not advertised to any peer
```

The following example shows how to display BGP information for the IPv4 unicast address family:

```
switch# show bgp ipv4 unicast 11.1.1.0 vrf cust_1
BGP routing table information for VRF cust_1, address family IPv4 Unicast
BGP routing table entry for 11.1.1.0/24, version 4
Paths: (3 available, best #1)
Flags: (0x80080012) on xmit-list, is in urib, is backup urib route, is in HW
  vpn: version 1093, (0x100002) on xmit-list
Multipath: eBGP iBGP

  Advertised path-id 1, VPN AF advertised path-id 1
  Path type: external, path is valid, is best path, in rib
             Imported from 13.13.13.13:3:[5]:[0]:[0]:[24]:[11.1.1.0]/224
  AS-Path: 2000000 500000 , path sourced external to AS
    11.1.1.100 (metric 5) from 102.102.102.102 (102.102.102.102)
      Origin incomplete, MED not set, localpref 100, weight 0
      Received label 22001
      Received path-id 1
      Extcommunity: RT:23456:22001 Route-Import:22.22.22.22:2001 ENCAP:8
          Router MAC:e00e.da4a.62a5

  Path type: external, path is valid, not best reason: Neighbor Address, multipath, in rib
             Imported from 13.13.13.13:3:[5]:[0]:[0]:[24]:[11.1.1.0]/224
  AS-Path: 2000000 100 , path sourced external to AS
    11.1.1.233 (metric 5) from 102.102.102.102 (102.102.102.102)
      Origin incomplete, MED not set, localpref 100, weight 0
      Received label 22001
      Received path-id 2
      Extcommunity: RT:23456:22001 Route-Import:33.33.33.33:2001 ENCAP:8
          Router MAC:e00e.da4a.589d

  Path type: external, path is valid, not best reason: Neighbor Address, multipath, in rib
             Imported from 13.13.13.13:3:[5]:[0]:[0]:[24]:[11.1.1.0]/224
  AS-Path: 2000000 100000 , path sourced external to AS
    11.1.1.133 (metric 5) from 102.102.102.102 (102.102.102.102)
      Origin incomplete, MED not set, localpref 100, weight 0
      Received label 22001
```

```
      Received path-id 3
      Extcommunity: RT:23456:22001 Route-Import:11.11.11.11:2001 ENCAP:8
          Router MAC:003a.7d7d.1dbd

  VRF advertise information:
  Path-id 1 not advertised to any peer

  VPN AF advertise information:
  Path-id 1 not advertised to any peer
```

The following example shows how to display routes from the unicast RIB after the Proportional Multipath for VNF feature has been configured:

```
switch# show ip route 1.1.1.0 vrf cust_1
IP Route Table for VRF "cust_1"
…
1.1.1.0/24, ubest/mbest: 22/0, all-best (0x300003d)
    *via 3.0.0.1, [1/0], 08:13:17, static
         recursive next hop: 3.0.0.1/32
    *via 3.0.0.2, [1/0], 08:13:17, static
         recursive next hop: 3.0.0.2/32
    *via 3.0.0.3, [1/0], 08:13:16, static
         recursive next hop: 3.0.0.3/32
    *via 3.0.0.4, [1/0], 08:13:16, static
         recursive next hop: 3.0.0.4/32
   *via 2.0.0.1, [200/0], 06:09:19, bgp-2, internal, tag 2 (evpn) segid: 3003802 tunnelid:
 0x300003e encap: VXLAN
          BGP-EVPN: VNI=3003802 (EVPN)
         client-specific data: 3b
         recursive next hop: 2.0.0.1/32
         extended route information: BGP origin AS 2 BGP peer AS 2
   *via 2.0.0.2, [200/0], 06:09:19, bgp-2, internal, tag 2 (evpn) segid: 3003802 tunnelid:
 0x300003e encap: VXLAN
          BGP-EVPN: VNI=3003802 (EVPN)
         client-specific data: 3b
         recursive next hop: 2.0.0.2/32
         extended route information: BGP origin AS 2 BGP peer AS 2
```

The following example shows how to display the GIR user-defined profile for the maintenance mode:

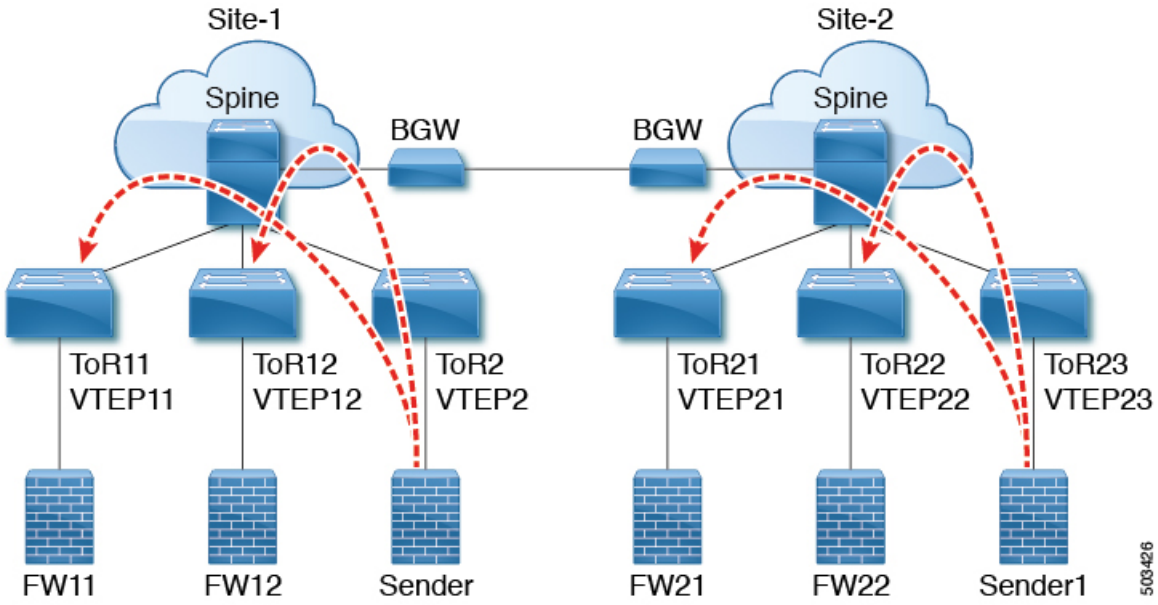
```
switch# show maintenance profile maintenance-mode
  [Maintenance Mode]
  ip pim isolate
  router bgp 2
   isolate
  router isis 1
   isolate
  route-map GIR deny 5
```

The following example shows how to display the GIR user-defined profile for the normal mode:

```
switch# show maintenance profile normal-mode
  [Normal Mode]
  no ip pim isolate
  router bgp 2
   no isolate
  router isis 1
   no isolate
  route-map GIR permit 5
```

Configuration Example for Proportional Multipath for VNF with Multi-Site

Figure 69: Proportional Multipath for VNF with Multi-Site Topology



The following configuration example allows traffic to be sent across sites if a local VNF isn't available.

```
feature telnet
feature nxapi
feature bash-shell
feature scp-server
nv overlay evpn
feature ospf
feature bgp
feature pim
feature interface-vlan
feature vn-segment-vlan-based
feature bfd
feature nv overlay

no password strength-check
username admin password 5 password role network-admin
ip domain-lookup
copp profile strict
evpn multisite border-gateway 1
  delay-restore time 30
snmp-server user admin network-admin auth md5 0x66a8185ad28d9df13d9214f6e19aad37 priv
0x66a8185ad28d9df13d9214f6e19aad37 localizedkey

fabric forwarding anycast-gateway-mac 0000.2222.3333
ip pim ssm range 232.0.0.0/8
vlan 1,14,24,100-110,120-150,1000-1010,1100-1110,2000-2010,2100-2110,3000-3010
vlan 100
  name l2-vni-vlan-0-for-vrf100
  vn-segment 2000100
```

```
vlan 101
  name l2-vni-vlan-0-for-vrf101
  vn-segment 2000101
vlan 1100
  name l2-vni-vlan-1-for-vrf100
  vn-segment 2001100
vlan 1101
  name l2-vni-vlan-1-for-vrf101
  vn-segment 2001101
vlan 2100
  name l3-vni-vlan-for-vrf100
  vn-segment 3000100
vlan 2101
  name l3-vni-vlan-for-vrf101
  vn-segment 3000101

route-map passall permit 10
  set path-selection all advertise
route-map permit-all permit 10
  set path-selection all advertise
route-map permit-all-v6 permit 10

vrf context vrf100
  vni 3000100
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
  address-family ipv6 unicast
    route-target both auto
    route-target both auto evpn
vrf context vrf101
 vni 3000101
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
  address-family ipv6 unicast
    route-target both auto
    route-target both auto evpn

interface Vlan14
  no shutdown
  vrf member vrf100
  ip address 192.14.0.1/24
  ipv6 address 192:14::1/64

interface Vlan24
  no shutdown
  vrf member vrf101
  ip address 192.24.0.1/24
  ipv6 address 192:24::1/64

interface Vlan100
  description "L3VRF.VLANNUM.0.222"
  no shutdown
  vrf member vrf100
  ip address 100.0.0.222/24
  ipv6 address 100::222/64
  fabric forwarding mode anycast-gateway

interface Vlan101
  description "L3VRF.VLANNUM.0.222"
  no shutdown
```

```
  vrf member vrf101
  ip address 101.0.0.222/24
  ipv6 address 101::222/64
  fabric forwarding mode anycast-gateway

interface Vlan1100
  description "L3VRF.VLANNUM.0.222"
  no shutdown
  vrf member vrf100
  ip address 100.1.0.222/16
  ipv6 address 100:1::222/64
  fabric forwarding mode anycast-gateway

interface Vlan1101
  description "L3VRF.VLANNUM.0.222"
  no shutdown
  vrf member vrf101
  ip address 101.1.0.222/16
  ipv6 address 101:1::222/64
  fabric forwarding mode anycast-gateway

interface Vlan2100
  no shutdown
  vrf member vrf100
  ip forward
  ipv6 address use-link-local-only

interface Vlan2101
  no shutdown
  vrf member vrf101
  ip forward
  ipv6 address use-link-local-only

interface nve1
  no shutdown
  host-reachability protocol bgp
  source-interface loopback1
  multisite border-gateway interface loopback2
  member vni 2000100-2000110
    suppress-arp
    mcast-group 227.1.1.1
  member vni 2000120-2000150
    suppress-arp
    mcast-group 227.1.1.1
  member vni 2001100-2001110
    suppress-arp
    mcast-group 227.1.1.1
  member vni 3000100-3000110 associate-vrf
  member vni 3100100-3100110 associate-vrf

interface Ethernet1/22
  description "BGW11 to BGW2"
  medium p2p
  ip unnumbered loopback0
  ip ospf cost 40
  ip ospf network point-to-point
  ip router ospf 12 area 0.0.0.0
  no shutdown
  evpn multisite dci-tracking

interface Ethernet1/25
  description "BGW11 to Spine11"
  medium p2p
  ip unnumbered loopback0
```

```
  ip ospf cost 40
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  no shutdown
  evpn multisite fabric-tracking

interface Ethernet1/27
  description "BGW11 to Spine12"
  medium p2p
  ip unnumbered loopback0
  ip ospf cost 40
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  no shutdown
  evpn multisite fabric-tracking

interface Ethernet1/34
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 14,24
  no shutdown

interface loopback0
  ip address 1.1.11.0/32
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode

interface loopback1
  ip address 1.1.11.1/32
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode

interface loopback2
  ip address 11.11.11.11/32
  ip router ospf 12 area 0.0.0.0
  ip pim sparse-mode

router ospf 1
  redistribute direct route-map permit-all
router ospf 12
  redistribute direct route-map permit-all
ip load-sharing address source-destination rotate 32 universal-id 1

router bgp 1
  log-neighbor-changes
  address-family l2vpn evpn
    maximum-paths 8
    maximum-paths ibgp 8
    additional-paths send
    additional-paths receive
    additional-paths selection route-map passall
  neighbor 1.2.11.1
    remote-as 1
    description "SPINE-11"
    update-source loopback1
    address-family l2vpn evpn
      send-community extended
  neighbor 1.2.12.1
    remote-as 1
    description "SPINE-12"
    update-source loopback1
    address-family l2vpn evpn
      send-community extended
  neighbor 2.1.2.1
```

```
    remote-as 2
    description "BGW-2"
    update-source loopback1
    ebgp-multihop 3
    peer-type fabric-external
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
      rewrite-evpn-rt-asn
  vrf vrf100
    address-family ipv4 unicast
      redistribute direct route-map permit-all
      maximum-paths 8
      maximum-paths ibgp 8
      export-gateway-ip
    address-family ipv6 unicast
      redistribute direct route-map permit-all
      maximum-paths 8
      maximum-paths ibgp 8
      export-gateway-ip
  vrf vrf101
    address-family ipv4 unicast
      redistribute direct route-map permit-all
      maximum-paths 8
      maximum-paths ibgp 8
      export-gateway-ip
    address-family ipv6 unicast
      redistribute direct route-map permit-all
      maximum-paths 8
      maximum-paths ibgp 8
      export-gateway-ip
evpn
  vni 2000100 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 2000101 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 2001100 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 2001101 l2
    rd auto
    route-target import auto
    route-target export auto
```

The following example shows that the VTEP in site 1 prefers the local VNF (FW).

```
leaf1# show bgp l2vpn evpn 200.100.1.1
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 1.3.12.0:3
BGP routing table entry for [5]:[0]:[0]:[32]:[200.100.1.1]/224, version 77902
Paths: (4 available, best #2)
Flags: (0x000002) (high32 00000000) on xmit-list, is not in l2rib/evpn, is not in HW
Multipath: eBGP iBGP Local

  Path type: internal, path is valid, not best reason: Neighbor Address, no labeled nexthop

  Gateway IP: 100.0.0.12
  AS-Path: 99 100 , path sourced external to AS
    1.3.12.1 (metric 81) from 1.2.12.1 (1.2.12.0)
```

```
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 3000100
      Received path-id 2
      Extcommunity: RT:1:3000100 ENCAP:8 Router MAC:00be.7547.13bf
      Originator: 1.3.12.0 Cluster list: 1.2.12.0

  Advertised path-id 2
  Path type: local, path is valid, not best reason: Locally originated, multipath, no labeled
 nexthop
  Gateway IP: 100.0.0.11
  AS-Path: 99 100 , path sourced external to AS
    1.3.11.1 (metric 0) from 0.0.0.0 (1.3.11.0)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 3000100
      Received path-id 1
      Extcommunity: RT:1:3000100 ENCAP:8 Router MAC:d478.9bb3.c1a1
```

The following example shows how the local VNF is disabled so that the VNF from site 2 is used. The BGP adjacency is shut down between site 1's VTEP11 to FW11 and between VTEP12 to FW12.

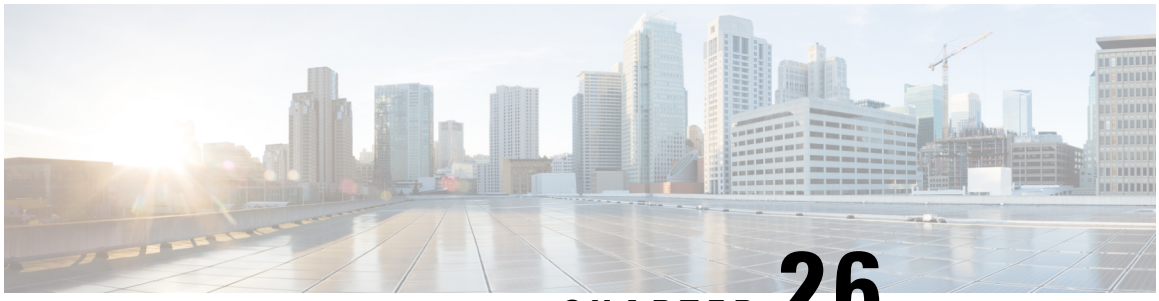
```
leaf1(config-router)# vrf vrf100
leaf1(config-router-vrf)# neighbor 100::11
leaf1(config-router-vrf-neighbor)# shut
leaf1(config-router-vrf-neighbor)# neighbor 100::12
leaf1(config-router-vrf-neighbor)# shut
leaf1(config-router-vrf-neighbor)# neighbor 100:1::11
leaf1(config-router-vrf-neighbor)# shut
leaf1(config-router-vrf-neighbor)# neighbor 100:1::12
leaf1(config-router-vrf-neighbor)# shut
leaf1(config-router-vrf-neighbor)# neighbor 100.0.0.11
leaf1(config-router-vrf-neighbor)# shut
leaf1(config-router-vrf-neighbor)# neighbor 100.0.0.12
leaf1(config-router-vrf-neighbor)# shut
leaf1(config-router-vrf-neighbor)# neighbor 100.1.0.11
leaf1(config-router-vrf-neighbor)# shut
leaf1(config-router-vrf-neighbor)# neighbor 100.1.0.12
leaf1(config-router-vrf-neighbor)# shut
leaf1(config-router-vrf-neighbor)# end
```

The following example shows that the prefix now uses the VNF (FW) from site 2.

```
leaf1# show bgp l2vpn evpn 200.100.1.1
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 1:3000100
BGP routing table entry for [5]:[0]:[0]:[32]:[200.100.1.1]/224, version 97269
Paths: (3 available, best #3)
Flags: (0x000002) (high32 00000000) on xmit-list, is not in l2rib/evpn, is not in HW
Multipath: eBGP iBGP Local

  Path type: internal, path is valid, not best reason: Neighbor Address, no labeled nexthop

  Gateway IP: 100.1.0.21
  AS-Path: 2 99 100 , path sourced external to AS
    11.11.11.11 (metric 20) from 1.2.12.1 (1.2.12.0)
      Origin IGP, MED 2000, localpref 100, weight 0
      Received label 3000100
      Received path-id 2
      Extcommunity: RT:1:3000100 SOO:03030100:00000000 ENCAP:8
          Router MAC:0200.0b0b.0b0b
      Originator: 1.1.12.0 Cluster list: 1.2.12.0
```



CHAPTER 26

EVPN Distributed NAT

- EVPN Distributed NAT , on page 647

EVPN Distributed NAT

Beginning with Cisco NX-OS Release 10.2(1)F, EVPN Distributed NAT feature is supported on N9K-C9336C-FX2, N9K-C93240YC-FX2, N9K-C93360YC-FX2 TOR switches. The Distributed Elastic NAT feature enables NAT on the leaf and spine in the VXLAN topology.

Guidelines and Limitations of EVPN Distributed NAT

EVPN Distributed NAT supports the following:

- Up to 8192 NAT translations
- Static NAT
- IPv4 NAT
- Match in VRF-aware NAT
- Add-route for static inside configuration

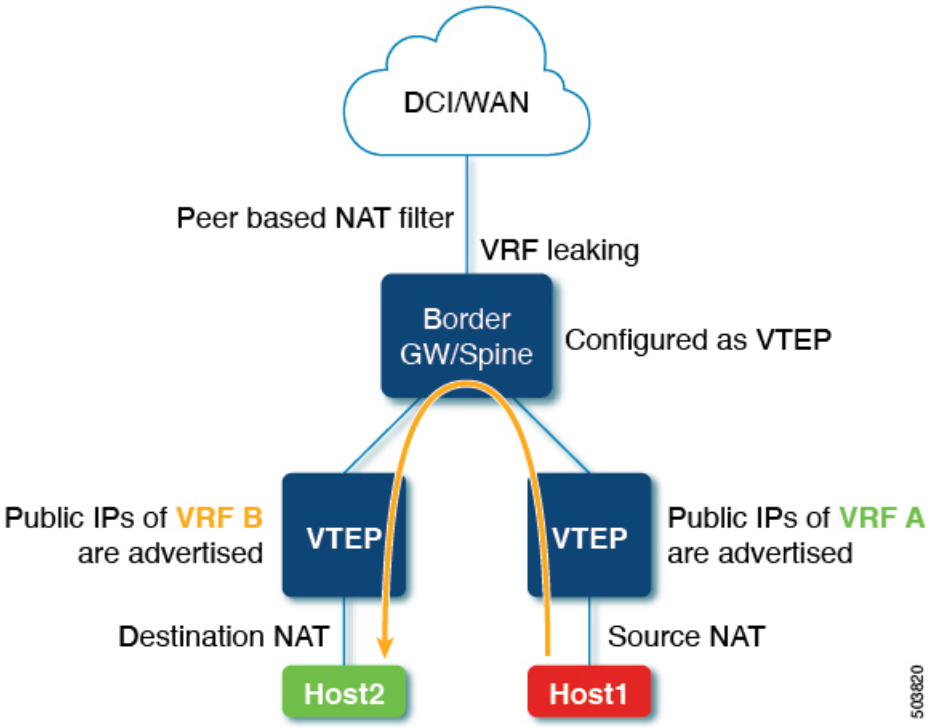
EVPN Distributed NAT does not support the following:

- IPv6 NAT
- Dynamic NAT
- NAT mobility
- Subnet-based filtering
- Per rule statistics
- NAT is unaware of vPC. NAT configuration should be identical on both vPC peers.
- Within a fabric if source and destination hosts are in same VRF, regular NAT can be used. EVPN Distributed NAT is not supported within same VRF. It is supported between different VRF's.

EVPN Distributed NAT Topology

The following topology illustrates the EVPN Distributed NAT configuration on VTEPs.

Figure 70: EVPN Distributed NAT Configuration Topology



In the above topology:

- EVPN Distributed NAT is configured only on the VTEPs.
- The spine does not require any EVPN Distributed NAT related configuration.
- Spine is configured as a VTEP.
- Only the routes are leaked in the spine for reachability using VxLAN underlay routing protocols.
- The Source and Destination NAT are configured on both the leaf.
- Source NAT is performed on the switch directly connected to the Source.
- Destination NAT is performed on the switch directly connected to the Destination.
- If both Source and Destination are on the same switch, Source NAT is performed first. The packet is then looped through Spine, and the Destination NAT is performed.
- Hosts can send traffic using private IP address or public IP address, depending on the requirement.
- VXLAN Peer-based NAT filtering is configured.

Peer-based NAT Filter

- The peer-based NAT filter allows NAT only for the flows that are destined to the configured tunnel endpoints and the rest of the flows remain unaffected.

- Peer-based NAT filter is useful in cases where large number of prefixes needs to be NATed.
- NAT ACL region must be carved first so that the peer-based NAT filter can work.
- You can configure peer-based filters on the border nodes.
- Peer-based NAT filter is useful for inter-VRF cases such as a service leaf where centralized VRF leak is configured.
- You can configure peer-based NAT filter using the **system nve nat peer-ip** <*peer-ip*> command.

VRF-Aware NAT

- The VRF aware NAT enables a switch to understand an address space in a VRF (virtual routing and forwarding instances) and to translate the packet. This allows the NAT feature to translate traffic in an overlapping address space that is used between two VRFs.
- You can enable FP Tile-based NAT using **system routing vrf-aware-nat** command.
- For more details on VRF aware NAT, see Cisco Nexus 9000 NX-OS Interfaces Configuration Guide.

Configuring EVPN Distributed NAT

The following is the EVPN Distributed NAT configuration in Leaf-1.

```
feature bgp
feature interface-vlan
feature vn-segment-vlan-based
feature nat
feature nv overlay

hardware access-list tcam region nat 512  (Carves NAT TCAM)


system routing vrf-aware-nat
system nve nat peer-ip 100.100.100.3       (peer-ip is the Spine address which is leaking
the route)

ip nat inside source static 21.1.1.10 172.21.1.10 vrf vrf1 match-in-vrf add-route

ip nat inside source static 31.1.1.10 172.31.1.10 vrf vrf2 match-in-vrf add-route

vlan 202
  vn-segment 20202

vlan 301
  vn-segment 20301

vlan 3200
  vn-segment 33200

vlan 3300
 vn-segment 33300


interface Vlan202
  no shutdown
  vrf member vrf1
  ip address 22.1.1.1/24
  fabric forwarding mode anycast-gateway
  ip nat inside
```

```
interface Vlan3200
  no shutdown
  vrf member vrf1
  ip forward
  ip nat outside

interface Vlan301
  no shutdown
  vrf member vrf2
  ip address 31.1.1.1/24
  fabric forwarding mode anycast-gateway
  ip nat inside

interface Ethernet1/11
  switchport mode trunk

interface Ethernet1/35
  switchport mode trunk

vrf context vrf1
  vni 33200
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn

vrf context vrf2
  vni 33300
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn

router bgp 100
  vrf vrf1
    address-family ipv4 unicast
      network 172.21.1.10/32
      advertise l2vpn evpn
vrf vrf2
    address-family ipv4 unicast
      network 172.31.1.10/32
      advertise l2vpn evpn
```

The following is the EVPN Distributed NAT configuration in Leaf-2.

```
feature bgp
feature interface-vlan
feature vn-segment-vlan-based
feature nat
feature nv overlay

system routing vrf-aware-nat
system nve nat peer-ip 100.100.100.3   (peer-ip is the spine address which is leaking the
route)

ip nat inside source static 21.1.1.20 172.21.1.20 vrf vrf1 match-in-vrf add-route

ip nat inside source static 31.1.1.20 172.31.1.20 vrf vrf2 match-in-vrf add-route

vlan 202
  vn-segment 20202

vlan 301
  vn-segment 20301
```

```
vlan 3200
  vn-segment 33200

vlan 3300
 vn-segment 33300

interface Vlan202
  no shutdown
  vrf member vrf1
  ip address 22.1.1.1/24
  fabric forwarding mode anycast-gateway
  ip nat inside


interface Vlan3200
  no shutdown
  vrf member vrf1
  ip forward
  ip nat outside

interface Vlan301
  no shutdown
  vrf member vrf2
  ip address 31.1.1.1/24
  fabric forwarding mode anycast-gateway
  ip nat inside

interface Vlan3300
  no shutdown
  vrf member vrf2
  ip forward
  ip nat outside

interface Ethernet1/16
  switchport
  switchport mode trunk

interface Ethernet1/43
  switchport
  switchport mode trunk

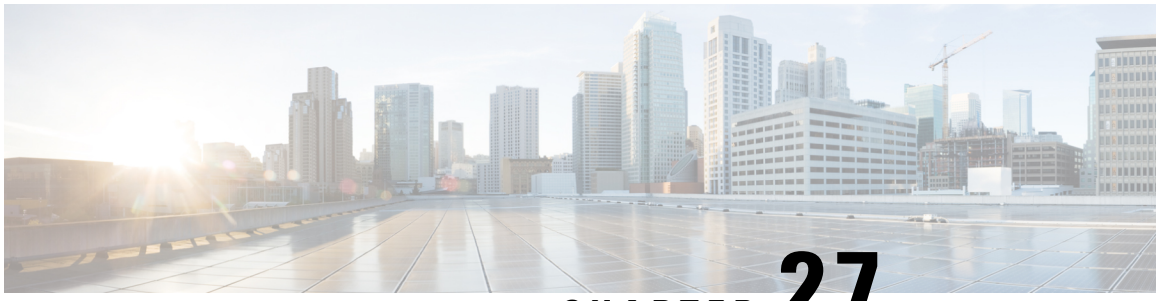
vrf context vrf1
  vni 33200
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
vrf context vrf2
  vni 33300
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn


router bgp 100
  vrf vrf1
    address-family ipv4 unicast
      network 172.21.1.20/32
      advertise l2vpn evpn
vrf vrf2
    address-family ipv4 unicast
      network 172.31.1.20/32
```

```
    advertise l2vpn evpn
```

The following show command provides the display of insulation policies configured in the switch for EVPN Distributed NAT.

```
show ip nat translations
Pro Inside global Inside local Outside local Outside global
any 174.2.216.2 42.2.216.2 --- ---
any 174.3.217.2 42.3.217.2 --- ---
```



CHAPTER 27

DHCP Relay in VXLAN BGP EVPN

This chapter contains these sections:

- DHCP Relay in VXLAN BGP EVPN Overview, on page 653
- Guidelines and Limitations for DHCP Relay in VXLAN BGP EVPN, on page 654
- DHCP Relay in VXLAN BGP EVPN Example, on page 655
- DHCP Relay on VTEPs, on page 655
- Client on Tenant VRF and Server on Layer 3 Default VRF, on page 656
- Client on Tenant VRF (SVI X) and Server on the Same Tenant VRF (SVI Y), on page 659
- Client on Tenant VRF (VRF X) and Server on Different Tenant VRF (VRF Y), on page 663
- Client on Tenant VRF and Server on Non-Default Non-VXLAN VRF, on page 666
- Configuring vPC Peers Example, on page 668
- vPC VTEP DHCP Relay Configuration Example, on page 670

DHCP Relay in VXLAN BGP EVPN Overview

DHCP relay is utilized to forward DHCP packets between the hosts and DHCP server. The VXLAN VTEP can act as a relay agent, providing DHCP relay services in a multi-tenant VXLAN environment.

With DHCP Relay, DHCP messages require to be sent through the same Switch in both directions. GiAddr (Gateway IP Address) for DHCP Relay is commonly used for Scope Selection and DHCP response messages. In any VXLAN fabric with Distributed IP Anycast Gateway, DHCP messages can be returned to ANY Switch hosting the respective Gateway IP Address (GiAddr).

Solution requires a different way of Scope Selection and Unique IP Address for each Switch. Unique Loopback Interface per Switch will become GiAddr for responding to correct Switch. Option 82 (dhcp option vpn) will be used for Scope Selection based on L2VNI.

In a multi-tenant EVPN environment, DHCP relay uses the following sub-options of Option 82:

- Sub-option 151(0x97) - Virtual Subnet Selection (Defined in RFC#6607)

  Used to convey VRF related information to the DHCP server in an MPLS-VPN and VXLAN EVPN multi-tenant environment.

- Sub-option 11(0xb) - Server ID Override (Defined in RFC#5107)

  The server identifier (server ID) override sub-option allows the DHCP relay agent to specify a new value for the server ID option, which is inserted by the DHCP server in the reply packet. This sub-option allows the DHCP relay agent to act as the actual DHCP server such that the renew requests will come to the

relay agent rather than the DHCP server directly. The server ID override sub-option contains the incoming interface IP address, which is the IP address on the relay agent that is accessible from the client. Using this information, the DHCP client sends all renew and release request packets to the relay agent. The relay agent adds all of the appropriate sub-options and then forwards the renew and release request packets to the original DHCP server. For this function, Cisco's proprietary implementation is sub-option 152(0x98). You can use the **ip dhcp relay sub-option type cisco** command to manage the function.

- Sub-option 5(0x5) - Link Selection (Defined in RFC#3527)

  The link selection sub-option provides a mechanism to separate the subnet/link on which the DHCP client resides from the gateway address (giaddr), which can be used to communicate with the relay agent by the DHCP server. The relay agent will set the sub-option to the correct subscriber subnet and the DHCP server will use that value to assign an IP address rather than the giaddr value. The relay agent will set the giaddr to its own IP address so that DHCP messages are able to be forwarded over the network. For this function, Cisco's proprietary implementation is sub-option 150(0x96). You can use the **ip dhcp relay sub-option type cisco** command to manage the function.

Guidelines and Limitations for DHCP Relay in VXLAN BGP EVPN

DHCP Relay in VXLAN BGP EVPN has the following configuration guideline and limitation:

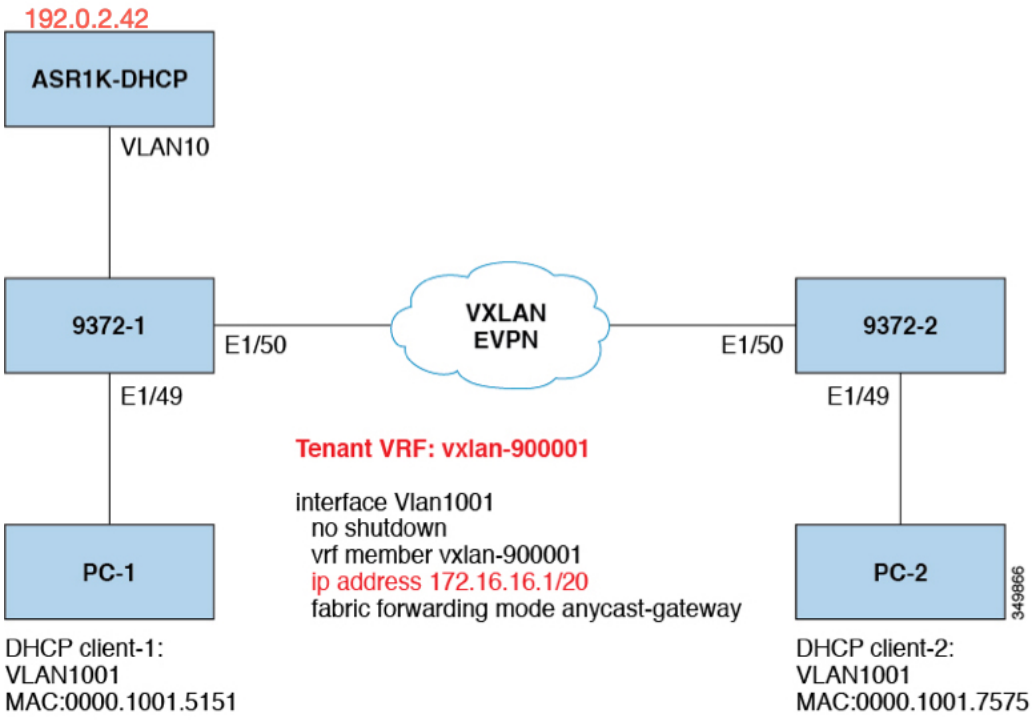
- DHCP Relay in VXLAN BGP EVPN feature is not supported on EoR switches.

DHCP Relay in VXLAN BGP EVPN supported release and platform

Release	Platform
7.x	Cisco Nexus 9300-FX/FX2
9.2(x)	Cisco Nexus 9300-GX
9.3(x)	Cisco Nexus 9300-FX3
10.2(3)F	Cisco Nexus 9300-GX2
10.4(1)F	Cisco Nexus 9332D-H2R
10.4(2)F	Cisco Nexus 93400LD-H1
10.4(3)F	Cisco Nexus 9364C-H1
10.5(3)F and later	Cisco Nexus 9364E-SG2-Q and 9364E-SG2-O switches
10.6(1)F and later	Cisco Nexus 9336C-SE1 switches

DHCP Relay in VXLAN BGP EVPN Example

Figure 71: Example Topology



Topology characteristics:

- Switches 9372-1 and 9372-2 are VTEPs connected to the VXLAN fabric.
- Client1 and client2 are DHCP clients in vlan1001. They belong to tenant VRF vxlan-900001.
- The DHCP server is ASR1K, a router that sits in vlan10.
- DHCP server configuration

```
ip vrf vxlan900001
ip dhcp excluded-address vrf vxlan900001 172.16.16.1 172.16.16.9
ip dhcp pool one
 vrf vxlan900001
 network 172.16.16.0 255.240.0.0
 defaultrouter 172.16.16.1
```

DHCP Relay on VTEPs

The following are common deployment scenarios:

- Client on tenant VRF and server on Layer 3 default VRF.

- Client on tenant VRF (SVI X) and server on the same tenant VRF (SVI Y).
- Client on tenant VRF (VRF X) and server on different tenant VRF (VRF Y).
- Client on tenant VRF and server on non-default non-VXLAN VRF.

The following sections below move vlan10 to different VRFs to depict different scenarios.

Client on Tenant VRF and Server on Layer 3 Default VRF

Put DHCP server (192.0.2.42) into the default VRF and make sure it is reachable from both 9372-1 and 9372-2 through the default VRF.

```
9372-1# sh run int vl 10

!Command: show running-config interface Vlan10
!Time: Mon Aug 24 07:51:16 2018

version 7.0(3)I1(3)

interface Vlan10
  no shutdown
  ip address 192.0.2.25/24
  ip router ospf 1 area 0.0.0.0

9372-1# ping 192.0.2.42 cou 1

PING 192.0.2.42 (192.0.2.42): 56 data bytes
64 bytes from 192.0.2.42: icmp_seq=0 ttl=254 time=0.593 ms
- 192.0.2.42 ping statistics -
1 packets transmitted, 1 packets received, 0.00% packet loss
roundtrip min/avg/max = 0.593/0.592/0.593 ms

9372-2# ping 192.0.2.42 cou 1
PING 192.0.2.42 (192.0.2.42): 56 data bytes
64 bytes from 192.0.2.42: icmp_seq=0 ttl=252 time=0.609 ms
- 192.0.2.42 ping statistics -
1 packets transmitted, 1 packets received, 0.00% packet loss
round-trip min/avg/max = 0.609/0.608/0.609 ms
```

DHCP Relay Configuration

- 9372-1

```
9372—1# sh run dhcp

!Command: show running—config dhcp
!Time: Mon Aug 24 08:26:00 2018

version 7.0(3) I1(3)
feature dhcp

service dhcp
ip dhcp relay
ip dhcp relay information option
ip dhcp relay information option vpn
```

```
ipv6 dhcp relay

interface Vlan1001
  ip dhcp relay address 192.0.2.42 use—vrf default
```

- 9372-2

```
9372-2# sh run dhcp

!Command: show running—config dhcp
!Time: Mon Aug 24 08:26:16 2018

version 7.0(3)11(3)
feature dhcp

service dhcp
ip dhcp relay
ip dhcp relay information option
ip dhcp relay information option vpn
ipv6 dhcp relay

interfaoe Vlan1001
  ip dhcp relay address 192.0.2.42 use—vrf default
```

Debug Output

- The following is a packet dump for DHCP interact sequences.

```
9372-1# ethanalyzer local interface inband display-filter
"udp.srcport==67 or udp.dstport==67" limit-captured frames 0

Capturing on inband
20180824 08:35:25.066530 0.0.0.0 -> 255.255.255.0 DHCP DHCP Discover - Transaction ID
0x636a38fd
20180824 08:35:25.068141 192.0.2.25 -> 192.0.2.42 DHCP DHCP Discover - Transaction ID
0x636a38fd
20180824 08:35:27.069494 192.0.2.42 -> 192.0.2.25 DHCP DHCP Offer Transaction - ID
0x636a38fd
20180824 08:35:27.071029 172.16.16.1 -> 172.16.16.11 DHCP DHCP Offer Transaction - ID
0x636a38fd
20180824 08:35:27.071488 0.0.0.0 -> 255.255.255.0 DHCP DHCP Request Transaction - ID
0x636a38fd
20180824 08:35:27.072447 192.0.2.25 -> 192.0.2.42 DHCP DHCP Request Transaction - ID
0x636a38fd
20180824 08:35:27.073008 192.0.2.42 -> 192.0.2.25 DHCP DHCP ACK Transaction - ID
0x636a38fd
20180824 08:35:27.073692 172.16.16.1 -> 172.16.16.11 DHCP DHCP ACK Transaction - ID
0x636a38fd
```



Note Ethanalyzer might not capture all DHCP packets because of inband interpretation issues when you use the filter. You can avoid this by using SPAN.

- DHCP Discover packet 9372-1 sent to DHCP server.

giaddr is set to 192.0.2.25 (ip address of vlan10) and suboptions 5/11/151 are set accordingly.

```
Bootp flags: 0x0000 (unicast)
client IP address: 0.0.0.0 (0.0.0.0)
Your (client) IP address: 0.0.0.0 (0.0.0.0)
Next server IP address: 0.0.0.0 (0.0.0.0)
Relay agent IP address: 192.0.2.25 (192.0.2.25)
client MAC address Hughes_01:51:51 (00:00:10:01:51:51)
client hardware address padding: 00000000000000000000
Server host name not given
Boot file name not given
Magic cookie: DHCP
Option: (53) DHCP Message Type
  Length: 1
  DHCP: Discover (1)
Option: (55) Parameter Request List
  Length: 4
  Parameter Request List Item: (1) Subnet Mask
  Parameter Request List Item: (3) Router
  Parameter Request List Item: (58) Renewal Time Value
  Parameter Request List Item: (59) Rebinding Time Value
Option: (61) client identifier
  Length: 7
  Hardware type: Ethernet (0x01)
  Client MAC address: Hughes_01:51:51 (00:00:10:01:51:51)
Option: (82) Agent Information Option
  Length: 47
Option 82 Suboption: (1) Agent Circuit ID
  Length: 10
  Agent Circuit ID: 01080006001e88690030
Option 82 Suboption: (2) Agent Remote ID
  Length: 6
  Agent Remote ID: f8c2882333a5
Option 82 Suboption: (151) VRF name/VPN ID
Option 82 Suboption: (11) Server ID Override
  Length: 4
  Server ID Override: 172.16.16.1 (172.16.16.1)
Option 82 Suboption: (5) Link selection
  Length: 4
  Link selection: 172.16.16.0 (172.16.16.0)
```

```
ASR1K-DHCP# sh ip dhcp bin
Bindings from all pools not associated with VRF:
IP address ClientID/ Lease expiration Type State Interface
         Hardware address/
         User name

Bindings from VRF pool vxlan900001:
IP address ClientID/ Lease expiration Type State Interface
         Hardware address/
         User name
172.16.16.10 0100.0010.0175.75 Aug 25 2018 09:21 AM Automatic Active GigabitEthernet2/1/0
172.16.16.11 0100.0010.0151.51 Aug 25 2018 08:54 AM Automatic Active GigabitEthernet2/1/0

9372-1# sh ip route vrf vxlan900001
IP Route Table for VRF "vxlan900001"
'*' denotes best ucast nexthop
'**' denotes best mcast nexthop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>
```

```
10.11.11.11/8, ubest/mbest: 2/0, attached
  *via 10.11.11.11, Lo1, [0/0], 18:31:57, local
  *via 10.11.11.11, Lo1, [0/0], 18:31:57, direct
10.22.22.22/8, ubest/mbest: 1/0
  *via 1.2.2.2%default, [200/0], 18:31:57, bgp65535,internal, tag 65535 (evpn)segid:
900001 tunnelid: 0x2020202
encap: VXLAN

172.16.16.0/20, ubest/mbest: 1/0, attached
 *via 172.16.16.1, Vlan1001, [0/0], 18:31:57, direct
172.16.16.1/32, ubest/mbest: 1/0, attached
 *via 172.16.16.1, Vlan1001, [0/0], 18:31:57, local
172.16.16.10/32, ubest/mbest: 1/0
 *via 1.2.2.2%default, [200/0], 00:00:47, bgp65535,internal, tag 65535 (evpn)segid:
900001 tunnelid: 0x2020202
encap: VXLAN

172.16.16.11/32, ubest/mbest: 1/0, attached
 *via 172.16.16.11, Vlan1001, [190/0], 00:28:10, hmm

9372-1# ping 172.16.16.11 vrf vxlan900001 count 1
PING 172.16.16.11 (172.16.16.11): 56 data bytes
64 bytes from 172.16.16.11: icmp_seq=0 ttl=63 time=0.846 ms
- 172.16.16.11 ping statistics -
1 packets transmitted, 1 packets received, 0.00% packet loss
round-trip min/avg/max = 0.846/0.845/0.846 ms

9372-1# ping 172.16.16.10 vrf vxlan900001 count 1
PING 172.16.16.10 (172.16.16.10): 56 data bytes
64 bytes from 172.16.16.10: icmp_seq=0 ttl=62 time=0.874 ms
- 172.16.16.10 ping statistics -
1 packets transmitted, 1 packets received, 0.00% packet loss
round-trip min/avg/max = 0.874/0.873/0.874 ms
```

Client on Tenant VRF (SVI X) and Server on the Same Tenant VRF (SVI Y)

Put DHCP server (192.0.2.42) into VRF of vxlan-900001 and make sure it is reachable from both 9372-1 and 9372-2 through VRF of vxlan-900001.

```
9372-1# sh run int vl 10

!Command: show running-config interface Vlan10
!Time: Mon Aug 24 09:10:26 2018

version 7.0(3)I1(3)

interface Vlan10
  no shutdown
  vrf member vxlan-900001
  ip address 192.0.2.25/24
```

Because 172.16.16.1 is an anycast address for vlan1001 configured on all the VTEPs, we need to pick up a unique address as the DHCP relay packet's source address to make sure the DHCP server can deliver a response

to the original DHCP Relay agent. In this scenario, we use loopback1 and we need to make sure loopback1 is reachable from everywhere of VRF vxlan-900001.

```
9372-1# sh run int lo1

!Command: show running-config interface loopback1
!Time: Mon Aug 24 09:18:53 2018

version 7.0(3)I1(3)

interface loopback1
  vrf member vxlan-900001
  ip address 10.11.11.11/8

9372-1# ping 192.0.2.42 vrf vxlan900001 source 10.11.11.11 cou 1
PING 192.0.2.42 (192.0.2.42) from 10.11.11.11: 56 data bytes
64 bytes from 192.0.2.42: icmp_seq=0 ttl=254 time=0.575 ms
- 192.0.2.42 ping statistics -
1 packets transmitted, 1 packets received, 0.00% packet loss
round-trip min/avg/max = 0.575/0.574/0.575 ms

9372-2# sh run int lo1

!Command: show running-config interface loopback1
!Time: Mon Aug 24 09:19:30 2018

version 7.0(3)I1(3)

interface loopback1
  vrf member vxlan900001
  ip address 10.22.22.22/8

9372-2# ping 192.0.2.42 vrf vxlan-900001 source 10.22.22.22 cou 1
PING 192.0.2.42 (192.0.2.42) from 10.22.22.22: 56 data bytes
64 bytes from 192.0.2.42: icmp_seq=0 ttl=253 time=0.662 ms
- 192.0.2.42 ping statistics -
1 packets transmitted, 1 packets received, 0.00% packet loss
round-trip min/avg/max = 0.662/0.662/0.662 ms
```

DHCP Relay Configuration

- 9372-1

```
9372—1# sh run dhcp

!Command: show running-config dhcp
!Time: Mon Aug 24 08:26:00 2018

version 7.0(3)11(3)
feature dhcp

service dhcp
ip dhcp relay
ip dhcp relay information option
I4ip dhcp relay information option vpn
ipv6 dhcp relay

interface Vlanl00l
  ip dhcp relay address 192.0.2.42
```

```
  ip dhcp relay source—interface loopback1
```

- 9372-2

```
9372—2# sh run dhcp

!Command: show running-config dhcp
!Time: Mon Aug 24 08:26:16 2018

version 7.0(3) 11(3)
feature dhcp

service dhcp
ip dhcp relay
ip dhcp relay information option
ip dhcp relay information option vpn
ipv6 dhcp relay

interface Vlan100l
  ip dhcp relay address 192.0.2.42
  ip dhcp relay source—interface loopback1
```

Debug Output

- The following is a packet dump for DHCP interact sequences.

```
9372-1# ethanalyzer local interface inband display-filter
"udp.srcport==67 or udp.dstport==67" limit-captured frames 0

Capturing on inband
20180824 09:31:38.129393 0.0.0.0 -> 255.255.255.0 DHCP DHCP Discover - Transaction ID
0x860cd13
20180824 09:31:38.129952 10.11.11.11 -> 192.0.2.42 DHCP DHCP Discover - Transaction ID
 0x860cd13
20180824 09:31:40.130134 192.0.2.42 -> 10.11.11.11 DHCP DHCP Offer - Transaction ID
0x860cd13
20180824 09:31:40.130552 172.16.16.1 -> 172.16.16.11 DHCP DHCP Offer - Transaction ID
0x860cd13
20180824 09:31:40.130990 0.0.0.0 -> 255.255.255.0 DHCP DHCP Request - Transaction ID
0x860cd13
20180824 09:31:40.131457 10.11.11.11 -> 192.0.2.42 DHCP DHCP Request - Transaction ID
0x860cd13
20180824 09:31:40.132009 192.0.2.42 -> 10.11.11.11 DHCP DHCP ACK - Transaction ID
0x860cd13
20180824 09:31:40.132268 172.16.16.1 -> 172.16.16.11 DHCP DHCP ACK - TransactionID
0x860cd13
```



Note Ethanalyzer might not capture all DHCP packets because of inband interpretation issues when you use the filter. You can avoid this by using SPAN.

- DHCP Discover packet 9372-1 sent to DHCP server.

  giaddr is set to 10.11.11.11(loopback1) and suboptions 5/11/151 are set accordingly.

```
Bootstrap Protocol
  Message type: Boot Request (1)
  Hardware type: Ethernet (0x01)
  Hardware address length: 6
  Hops: 1
  Transaction ID: 0x0860cd13
  Seconds elapsed: O
  Bootp flags: 0x0000 (unicast)
  Client IP address: 0.0.0.0 (0.0.0.0)
  Your (client) IP address: 0.0.0.0 (0.0.0.0)
  Next server IP address: 0.0.0.0 (0.0.0.0)
  Relay agent iP address: 10.11.11.11 (10.11.11.11)
  Client MAC address: Hughes_01:51:51 (00:00:10:01:51:51)
  Client hardware address padding: 00000000000000000000
  Server host name not given
  Boot file name not given
  Magic cookie: DHCP
  Option: (53) DHCP Message Type
    Length: 1
    DHCP: Discover (1)
  Option: (55) Parameter Request List
  Option: (61) Client Identifier
  Option: (82) Agent Information Option
    Length: 47
  Option 82 suboption: (1) Aqent Circuit ID
  Option 82 suboption: (151) Agent Remote ID
  Option 82 suboption: (11) Server ID Override
    Length: 4
    Server ID override: 172.16.16.1 (172.16.16.1)
  Option 82 suboption: (5) Link selection
    Length: 4
    Link selection: 172.16.16.0 (172.16.16.0)
```

```
ASR1K-DHCP# sh ip dhcp bin
Bindings from all pools not associated with VRF:
IP address ClientID/Lease expiration Type State Interface
         Hardware address/
         User name

Bindings from VRF pool vxlan-900001:
IP address ClientID/Lease expiration Type State Interface
         Hardware address/
         User name

172.16.16.10 0100.0010.0175.75 Aug 25 2018 10:02 AM Automatic Active GigabitEthernet2/1/0
172.16.16.11 0100.0010.0151.51 Aug 25 2018 09:50 AM Automatic Active GigabitEthernet2/1/0

9372-1# sh ip route vrf vxlan-900001
IP Route Table for VRF "vxlan-900001"
'*' denotes best ucast nexthop
'**' denotes best mcast nexthop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

10.11.11.11/8, ubest/mbest: 2/0, attached
  *via 10.11.11.11, Lo1, [0/0], 19:13:56, local
  *via 10.11.11.11, Lo1, [0/0], 19:13:56, direct
10.22.22.22/8, ubest/mbest: 1/0
  *via 2.2.2.2%default, [200/0], 19:13:56, bgp65535,internal, tag 65535 (evpn)segid:
900001 tunnelid: 0x2020202
```

```
encap: VXLAN
172.16.16.0/20, ubest/mbest: 1/0, attached
  *via 172.16.16.1, Vlan1001, [0/0], 19:13:56, direct
172.16.16.1/32, ubest/mbest: 1/0, attached
  *via 172.16.16.1, Vlan1001, [0/0], 19:13:56, local
172.16.16.10/32, ubest/mbest: 1/0
  *via 2.2.2.2%default, [200/0], 00:01:27, bgp65535,
internal, tag 65535 (evpn)segid: 900001 tunnelid: 0x2020202
encap: VXLAN
172.16.16.11/32, ubest/mbest: 1/0, attached
  *via 172.16.16.11, Vlan1001, [190/0], 00:13:56, hmm
192.0.2.20/24, ubest/mbest: 1/0, attached
  *via 192.0.2.25, Vlan10, [0/0], 00:36:08, direct
192.0.2.25/24, ubest/mbest: 1/0, attached
  *via 192.0.2.25, Vlan10, [0/0], 00:36:08, local
9372-1# ping 172.16.16.10 vrf vxlan-900001 cou 1
PING 172.16.16.10 (172.16.16.10): 56 data bytes
64 bytes from 172.16.16.10: icmp_seq=0 ttl=62 time=0.808 ms
- 172.16.16.10 ping statistics -
1 packets transmitted, 1 packets received, 0.00% packet loss
round-trip min/avg/max = 0.808/0.808/0.808 ms

9372-1# ping 172.16.16.11 vrf vxlan-900001 cou 1
PING 172.16.16.11 (172.16.16.11): 56 data bytes
64 bytes from 172.16.16.11: icmp_seq=0 ttl=63 time=0.872 ms
- 172.16.16.11 ping statistics -
1 packets transmitted, 1 packets received, 0.00% packet loss
round-trip min/avg/max = 0.872/0.871/0.872 ms
```

Client on Tenant VRF (VRF X) and Server on Different Tenant VRF (VRF Y)

The DHCP server is placed into another tenant VRF vxlan-900002 so that DHCP response packets can access the original relay agent. We use loopback2 to avoid any anycast ip address that is used as the source address for the DHCP relay packets.

```
9372-1# sh run int vl 10
!Command: show runningconfig interface Vlan10
!Time: Tue Aug 25 08:48:22 2018

version 7.0(3)I1(3)
interface Vlan10
  no shutdown
  vrf member vxlan900002
  ip address 192.0.2.40/24

9372-1# sh run int lo2
!Command: show runningconfig interface loopback2
!Time: Tue Aug 25 08:48:57 2018
version 7.0(3)I1(3)
interface loopback2
  vrf member vxlan900002
  ip address 10.33.33.33/8

9372-2# sh run int lo2
!Command: show runningconfig interface loopback2
```

```
!Time: Tue Aug 25 08:48:44 2018
version 7.0(3)I1(3)
interface loopback2
  vrf member vxlan900002
  ip address 10.44.44.44/8

9372-1# ping 192.0.2.42 vrf vxlan-900002 source 10.33.33.33 cou 1
PING 192.0.2.42 (192.0.2.42) from 10.33.33.33: 56 data bytes
64 bytes from 192.0.2.42: icmp_seq=0 ttl=254 time=0.544 ms
- 192.0.2.42 ping statistics -
1 packets transmitted, 1 packets received, 0.00% packet loss
round-trip min/avg/max = 0.544/0.544/0.544 ms

9372-2# ping 192.0.2.42 vrf vxlan-900002 source 10.44.44.44 count 1
PING 192.0.2.42 (192.0.2.42) from 10.44.44.44: 56 data bytes
64 bytes from 192.0.2.42: icmp_seq=0 ttl=253 time=0.678 ms
- 192.0.2.42 ping statistics -
1 packets transmitted, 1 packets received, 0.00% packet loss
round-trip min/avg/max = 0.678/0.678/0.678 ms
```

DHCP Relay Configuration

- 9372-1

```
9372—1# sh run dhcp

!Command: show running—config dhcp
!Time: Mon Aug 24 08:26:00 2018

version 7.0(3) Ii (3)
feature dhcp

service dhcp
ip dhcp relay
ip dhcp relay information option
ip dhcp relay information option vpn
ipv6 dhcp relay

interface VlanlOOl
  ip dhcp relay address 192.0.2.42 use—vrf vxlan—900002
  ip dhcp relay source—interface loopback2
```

- 9372-2

```
!Command: show running-config dhcp
!Time: Mon Aug 24 08:26:16 2018

version 7.0(3)11(3)
feature dhcp

service dhcp
ip dhcp relay
ip dhcp relay information option
ip dhcp relay information option vpn
ipv6 dhcp relay

interface VlaniOOl
  ip dhcp relay address 192.0.2.42 use-vrf vxlan—900002
```

```
ip dhcp relay source—interface loopback2
```

Debug Output

- The following is a packet dump for DHCP interact sequences.

```
9372-1# ethanalyzer local interface inband display-filter "udp.srcport==67 or
udp.dstport==67" limit-captured-frames 0
Capturing on inband
20180825 08:59:35.758314 0.0.0.0 -> 255.255.255.0 DHCP DHCP Discover - Transaction ID
0x3eebccae
20180825 08:59:35.758878 10.33.33.33 -> 192.0.2.42 DHCP DHCP Discover - Transaction ID
 0x3eebccae
20180825 08:59:37.759560 192.0.2.42 -> 10.33.33.33 DHCP DHCP Offer - Transaction ID
0x3eebccae
20180825 08:59:37.759905 172.16.16.1 -> 172.16.16.11 DHCP DHCP Offer - Transaction ID
0x3eebccae
20180825 08:59:37.760313 0.0.0.0 -> 255.255.255.0 DHCP DHCP Request - Transaction ID
0x3eebccae
20180825 08:59:37.760733 10.33.33.33 -> 192.0.2.42 DHCP DHCP Request - Transaction ID
0x3eebccae
20180825 08:59:37.761297 192.0.2.42 -> 10.33.33.33 DHCP DHCP ACK - Transaction ID
0x3eebccae
20180825 08:59:37.761554 172.16.16.1 -> 172.16.16.11 DHCP DHCP ACK - Transaction ID
0x3eebccae
```

- DHCP Discover packet 9372-1 sent to DHCP server.

  giaddr is set to 10.33.33.33 (loopback2) and suboptions 5/11/151 are set accordingly.

```
Bootstrap Protocol
  Message type: Boot Request (1)
  Hardware type: Ethernet (0x01)
  Hardware address length: 6
  Hops: 1
  Transaction ID: Ox3eebccae
  Seconds elapsed: O
  Bootp flags: 0x0000 (unicast)
  Client IP address: 0.0.0.0 (0.0.0.0)
  Your (client) IP address: 0.0.0.0 (0.0.0.0)
  Next server IP address: 0.0.0.0 (0.0.0.0)
  Relay agent IP address: 10.33.33.33 (10.33.33.33)
  Client MAC address: i-iughes_01:51:51 (00:00:10:01:51:51)
  Client hardware address padding: 00000000000000000000
  Server host name not given
  Boot file name not given
  Magic cookie: DHCP
  Option: (53) DHCP Message Type
    Length: 1
    DHCP: Discover (1)
  Option: (55) Parameter Request List
  Option: (61) client identifier
  Option: (82) Agent Informatìon option
    Length: 47
  Option 82 Suboption: (1) Agent circuit W
  Option 82 suboption: (2) Agent Remote 10
  Option 82 suboption: (151) VRF name/VPN ID
  Option 82 Suboption: (11) Server ID Override
    Length: 4
```

```
    Server ID Override: 172.16.16.1 (172.16.16.1)
  Option 82 Suboption: (5) Link selection
    Length: 4
    Link selection: 172.16.16.0 (172.16.16.0)
```

Client on Tenant VRF and Server on Non-Default Non-VXLAN VRF

The DHCP server is placed into the management VRF and is reachable the through M0 interface. The IP address changes to 10.122.164.147 accordingly.

```
9372-1# sh run int m0
!Command: show running-config interface mgmt0
!Time: Tue Aug 25 09:17:04 2018
version 7.0(3)I1(3)
interface mgmt0
  vrf member management
  ip address 10.122.165.134/8

9372-1# ping 10.122.164.147 vrf management cou 1
PING 10.122.164.147 (10.122.164.147): 56 data bytes
64 bytes from 10.122.164.147: icmp_seq=0 ttl=251 time=1.024 ms
- 10.122.164.147 ping statistics -
1 packets transmitted, 1 packets received, 0.00% packet loss
round-trip min/avg/max = 1.024/1.024/1.024 ms

9372-2# sh run int m0
!Command: show running-config interface mgmt0
!Time: Tue Aug 25 09:17:47 2018
version 7.0(3)I1(3)
interface mgmt0
  vrf member management
  ip address 10.122.165.148/8

9372-2# ping 10.122.164.147 vrf management cou 1
PING 10.122.164.147 (10.122.164.147): 56 data bytes
64 bytes from 10.122.164.147: icmp_seq=0 ttl=251 time=1.03 ms
- 10.122.164.147 ping statistics -
1 packets transmitted, 1 packets received, 0.00% packet loss
round-trip min/avg/max = 1.03/1.03/1.03 ms
```

DHCP Relay Configuration

- 9372-1

```
9372—1# sh run dhcp 9372—2# sh run dhcp

!Command: show running-config dhcp
!Time: Mon Aug 24 08:26:00 2018

version 7.0(3)11(3)
feature dhcp

service dhcp
```

```
ip dhcp relay
ip dhcp relay information option
ip dhcp relay information option vpn
ipv6 dhcp relay

interface VlanlOOl
  ip dhcp relay address 10.122.164.147 use—vrf management
```

- 9372-2

```
9372-2# sh run dhcp
!Command: show running-config dhcp
!Time: Tue Aug 25 09:17:47 2018

version 7.0(3)11(3)
feature dhcp

service dhcp
ip dhcp relay
ip dhcp relay information option
ip dhcp relay information option vpn
ipv6 dhcp relay

interface VlanlOOl
ip dhcp relay address 10.122.164.147 use—vrf management
```

Debug Output

- The following is a packet dump for DHCP interact sequences.

```
9372-1# ethanalyzer local interface inband display-filter "udp.srcport==67 or
udp.dstport==67" limit-captured-frames 0
Capturing on inband
20180825 09:30:54.214998 0.0.0.0 -> 255.255.255.0 DHCP DHCP Discover - Transaction ID
0x28a8606d
20180825 09:30:56.216491 172.16.16.1 -> 172.16.16.11 DHCP DHCP Offer - Transaction ID
0x28a8606d
20180825 09:30:56.216931 0.0.0.0 -> 255.255.255.0 DHCP DHCP Request - Transaction ID
0x28a8606d
20180825 09:30:56.218426 172.16.16.1 -> 172.16.16.11 DHCP DHCP ACK - Transaction ID
0x28a8606d

9372-1# ethanalyzer local interface mgmt display-filter "ip.src==10.122.164.147 or
ip.dst==10.122.164.147" limit-captured-frames 0
Capturing on mgmt0
20180825 09:30:54.215499 10.122.165.134 -> 10.122.164.147 DHCP DHCP Discover - Transaction
 ID 0x28a8606d
20180825 09:30:56.216137 10.122.164.147 -> 10.122.165.134 DHCP DHCP Offer - Transaction
 ID 0x28a8606d
20180825 09:30:56.217444 10.122.165.134 -> 10.122.164.147 DHCP DHCP Request - Transaction
 ID 0x28a8606d
20180825 09:30:56.218207 10.122.164.147 -> 10.122.165.134 DHCP DHCP ACK - Transaction
ID 0x28a8606d
```

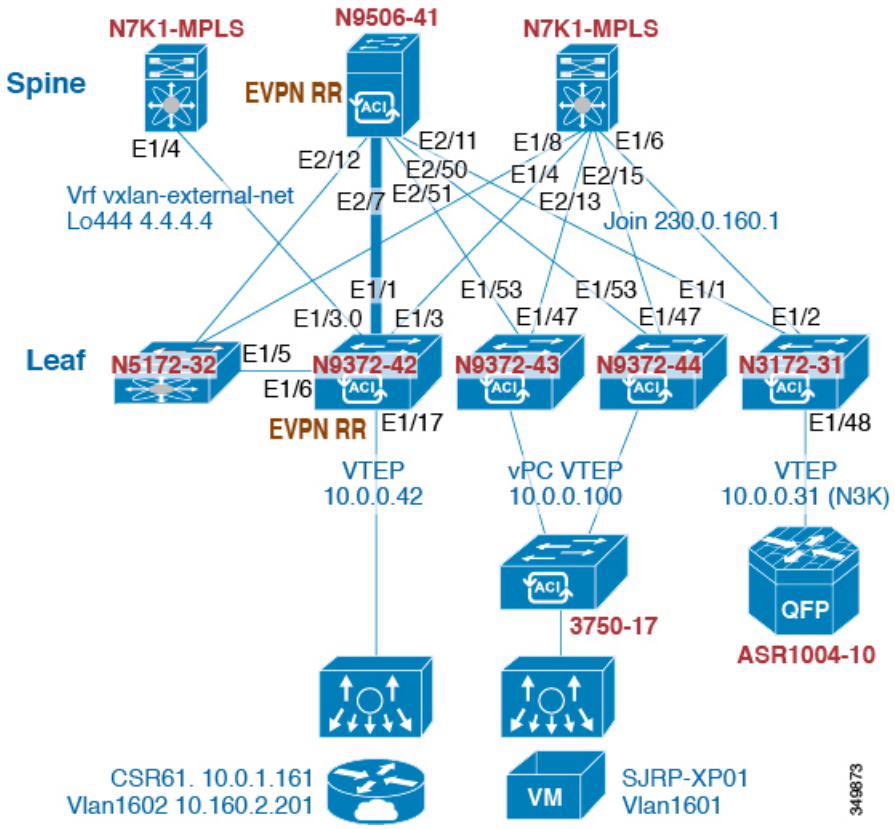
- DHCP Discover packet 9372-1 sent to DHCP server.

  giaddr is set to 10.122.165.134 (mgmt0) and suboptions 5/11/151 are set accordingly.

```
Bootstrap Protocol
  Message type: Boot Request (1)
  Hardware type: Ethernet (0x01)
  Hardware address length: 6
  Hops: 1
  Transaction ID: 0x28a8606d
  Seconds elapsed: O
  Bootp flags: 0x0000 (Unicast)
  Client IP address: 0.0.0.0 (0.0.0.0)
  Your (client) IP address: 0.0.0.0 (0.0.0.0)
  Next server IP address: 0.0.0.0 (0.0.0.0)
  Relay agent IP address: 10.122.165.134 (10.122.165.134)
  Client MAC address: Hughes_01:51:51 (00:00:10:01:51:51)
  Client hardware address padding: 00000000000000000000
  Server host name not given
  Boot file name not given
  Magic cookie: DHCP
  Option: (53) DHCP Message Type
    Length: 1
    DHCP: Discover (1)
  Option: (55) Parameter Request List
  Option: (61) Client identifier
  Option: (82) Agent Information Option
    Length: 47
    Option 82 Suboption: (1) Agent Circuit ID
    Option 82 Suboption: (2) Agent Remote ID
    Option 82 Suboption: (151) VRF name/VPN ID
    Option 82 Suboption: (11) Server ID Override
      Length: 4
      Server ID Override: 172.16.16.1 (172.16.16.1)
    Option 82 Suboption: (5) Link selection
      Length: 4
      Link selection: 172.16.16.0 (172.16.16.0)
```

Configuring vPC Peers Example

The following is an example of how to configure routing between vPC peers in the overlay VLAN for a DHCP relay configuration.



- Enable DHCP service.

```
service dhcp
```

- Configure DHCP relay.

```
ip dhcp relay
ip dhcp relay information option
ip dhcp relay sub-option type cisco
ip dhcp relay information option vpn
```

- Create loopback under VRF where you need DHCP relay service.

```
interface loopback601
  vrf member evpn-tenant-kk1
  ip address 192.0.2.36/24
  ip router ospf 1 area 0     /* Only required for vPC VTEP. */
```

- Advertise LoX into the Layer 3 VRF BGP.

```
Router bgp 2
vrf X
   network 10.1.1.42/8
```

- Configure DHCP relay on the SVI under the VRF.

```
interface Vlan1601
  vrf member evpn-tenant-kk1
  ip address 10.160.1.254/8
  fabric forwarding mode anycast-gateway
  ip dhcp relay address 10.160.2.201
  ip dhcp relay source-interface loopback601
```

- Configure Layer 3 VNI SVI with **ip forward**.

```
interface Vlan1600
  vrf member evpn-tenant-kk1
    ip forward
```

- Create the routing VLAN/SVI for the vPC VRF.



Note Only required for vPC VTEP

```
Vlan 1605
interface Vlan1605
  vrf member evpn-tenant-kk1
  ip address 10.160.5.43/8
  ip router ospf 1 area 10.10.10.41
```

- Create the VRF routing.



Note Only required for vPC VTEP.

```
router ospf 1
vrf evpn-tenant-kk1
    router-id 10.160.5.43
```

vPC VTEP DHCP Relay Configuration Example

To address a need to configure a VLAN that is allowed across the MCT/peer-link, such as a vPC VLAN, an SVI can be associated to the VLAN and is created within the tenant VRF. This becomes an underlay peering, with the underlay protocol, such as OSPF, that needs the tenant VRF instantiated under the routing process.

Alternatively, instead of placing the SVI within the routing protocol and instantiate the Tenant-VRF under the routing process, you can use the static routes between the vPC peers across the MCT. This approach ensures that the reply from the server returns to the correct place and each VTEP uses a different loopback interface for the GiAddr.

The following are examples of these configurations:

- Configuration of SVI within underlay routing:

```
/* vPC Peer-1 */

router ospf UNDERLAY
vrf tenant-vrf

interface Vlan2000
  no shutdown
  mtu 9216
  vrf member tenant-vrf
  ip address 192.168.1.1/16
  ip router ospf UNDERLAY area 0.0.0.0


/* vPC Peer-2 */

router ospf UNDERLAY
vrf tenant-vrf

interface Vlan2000
  no shutdown
  mtu 9216
  vrf member tenant-vrf
  ip address 192.168.1.2/16
  ip router ospf UNDERLAY area 0.0.0.0
```

- Configuration of SVI using static routes between vPC peers across the MCT:

```
/* vPC Peer-1 */

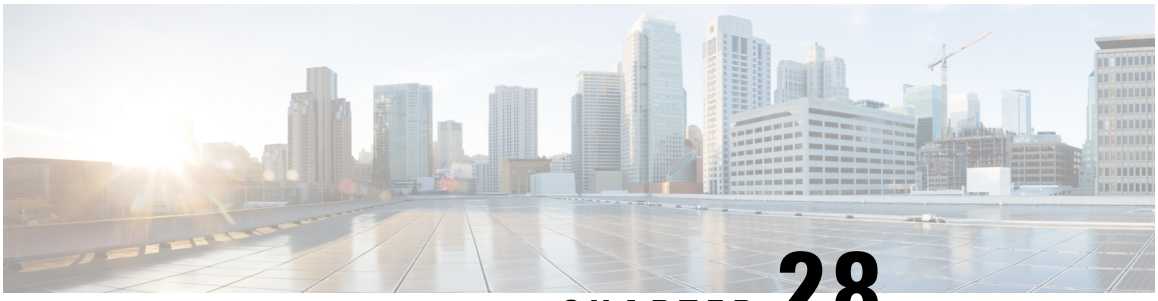
interface Vlan2000
  no shutdown
  mtu 9216
  vrf member tenant-vrf
  ip address 192.168.1.1/16

vrf context tenant-vrf
ip route 192.168.1.2/16 192.168.1.1


/* vPC Peer-2 */

interface Vlan2000
  no shutdown
  mtu 9216
  vrf member tenant-vrf
  ip address 192.168.1.2/16

vrf context tenant-vrf
ip route 192.168.1.1/16 192.168.1.2
```



CHAPTER 28

Configure VXLAN Cross Connect

- VXLAN cross connect, on page 673
- Topology of VXLAN Cross Connect, on page 674
- Guidelines and Limitations of VXLAN Cross Connect, on page 674
- Prerequisites, on page 676
- Configure the VXLAN Cross Connect, on page 676
- Verification of VXLAN Cross Connect Configuration, on page 678
- Remove a Cross Connect VNI, on page 680

VXLAN cross connect

VXLAN cross connect is a tunneling mechanism used on VXLAN fabric which

- provides point-to-point data tunneling,
- control packets such as CDP, LLDP, LACP, STP, BFD, and PAGP from one VTEP to another, and
- often referred to as **xConnect**.

BGP EVPN signaling identifies endpoints based on how the provider VNI is stretched in the fabric. Every attachment circuit will be part of a unique provider VNI. All inner customer tags will be preserved, as is, and packets are encapsulated in the provider VNI using **dot1q-tunnel**. On the decapsulation endpoint, the provider VNI will forward the packet to its attachment circuit while preserving all customer dot1q tags in the packets.

Supported xConnect tunnel combinations

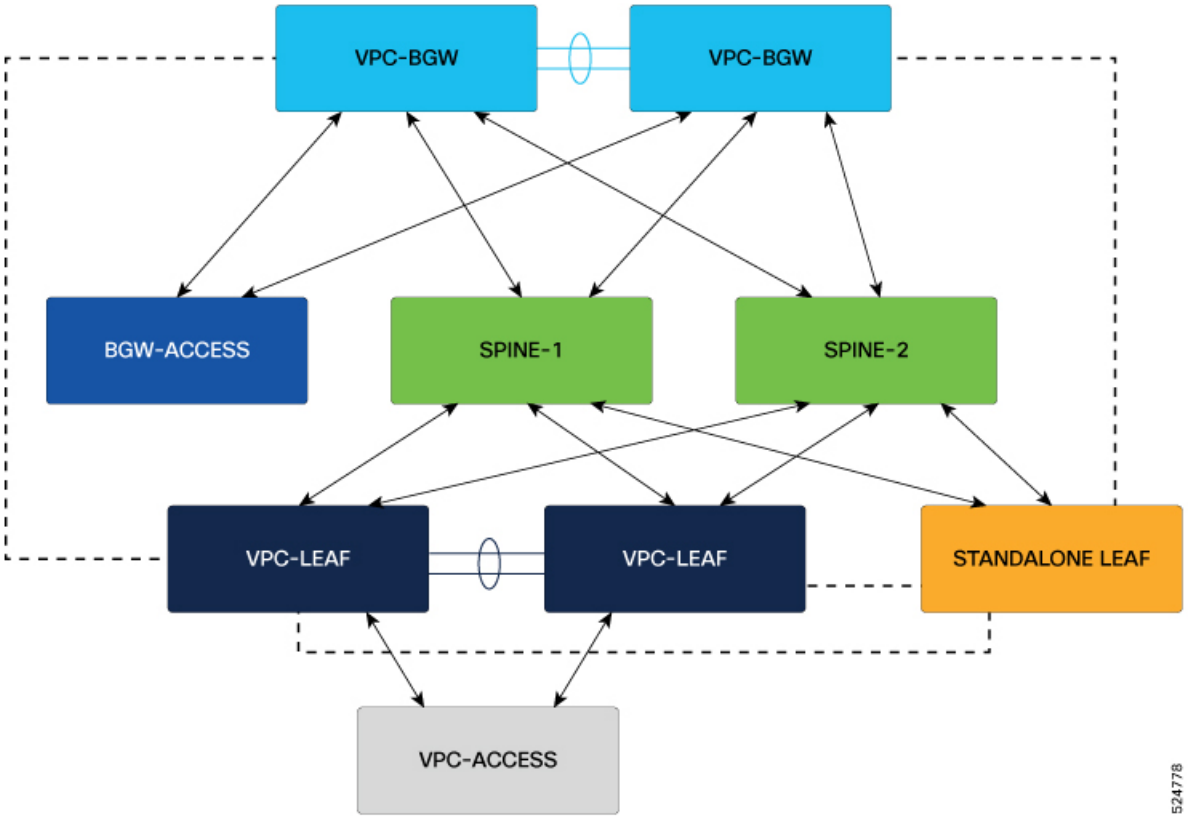
The supported tunnel combinations for VXLAN cross connect are

- physical interface to physical interface,
- port-channel to port-channel,
- mixed combinations of physical interface and port-channel, and
- additionally, this is supported over both underlay types (ingress replication and multicast) or a combination of both.

Topology of VXLAN Cross Connect

The following figure represents the **NGOAM xConnect** configuration on a VXLAN fabric. NGOAM xConnect is used to establish **pseudo-wire tunnels between VTEPs (VXLAN Tunnel Endpoints)**, which helps in

- proactive failure detection across the network fabric, and
- monitoring the enabled tunnels end-to-end connectivity across the fabric.



The dashed line in the diagram represents the NGOAM xConnect pseudo-wire, which is implemented between specific nodes to detect and react to failures.

Guidelines and Limitations of VXLAN Cross Connect

Generic guidelines and limitations of VXLAN Cross Connect

- MAC learning will be disabled on the xConnect VNIs and none of the host MAC will be learned on the tunnel access ports.
- Only one attachment circuit can be configured for a provider VNI on a given VTEP.
- A VNI can only be stretched in a point-to-point fashion. Point-to-multipoint is not supported.

- The scale of xConnect VLANs depends on the number of ports available on the switch. Every xConnect VLAN can tunnel all 4k customer VLANs.
- On vPC VTEP, cross connect needs backup SVI as native VLAN on the vPC peer-link.
- On vPC VTEP, spanning tree must be disabled on both vPC peers for xConnect VLANs.

ISSU and patch guidelines and limitations of VXLAN Cross Connect

- When an upgrade is performed non-disruptively from Cisco NX-OS Release 7.0(3)I7(4) to Cisco NX-OS Release 9.2(x) code, and if a VLAN is created and configured as xConnect, you must enter the **copy running-config startup-config** command and reload the switch. If the box was upgraded disruptively to Cisco NX-OS Release 9.2(x) code, a reload is not needed on configuring a VLAN as xConnect.
- Make sure that the **ngoam xconnect hb-interval** is set to 5000 milliseconds on all VTEPs before attempting ISSU/patch activation to avoid link flaps.
- Before activating the patch for the cfs process, you must move the NGOAM xConnect hb-interval to the maximum value of 5000 milliseconds. This prevents interface flaps during the patch activation.

Configuration and port guidelines and limitations of VXLAN Cross Connect

- If ARP Suppression is enabled on a VLAN, and you enable xConnect on the VLAN, the xConnect feature takes precedence.
- xConnect enabled VLANs are accessible only through switch port access mode.
- Configuring a static MAC on xConnect tunnel interfaces is not supported.
- The vPC orphan tunneled port per VNI should be either on the vPC primary switch or secondary switch, but not both.
- xConnect is not supported on FEX ports.
- xConnect access ports must be flapped after disabling NGOAM on all the VTEPs.
- After deleting and adding a VLAN, or removing xConnect from a VLAN, physical ports must be flapped with NGOAM.

Supported Features of VXLAN Cross Connect

- Supported only on BGP EVPN topology.
- Supported on vPC fabric peering.
- Does not support LACP bundling of attachment circuits.
- Does not support SVI on an xconnect VLAN.
- Does not support multi-site solution.

Supported platform and release of VXLAN cross connect

SupportedRelease	SupportedPlatform
9.3(3) and later	CiscoNexus 9300-FX/FX2/GX Series switches
9.3(5) and later	CiscoNexus 9300-FX3 Series switches
10.2(3)F and later	CiscoNexus 9300-GX2 Series switches
10.4(1)F and later	CiscoNexus 9332D-H2R switches
10.4(2)F and later	CiscoNexus 93400LD-H1 switches
10.4(3)F and later	CiscoNexus 9364C-H1 switches

Prerequisites

Result

- Before you begin, verify that the **feature ngoam** is enabled. For more information on how to enable, see .

```
switch# configure terminal
switch(config)# feature ngoam
```

- (Optional) Use the **ngoam xconnect hb-interval** *interval* command in global configuration mode, only if user wants to configure non-default heartbeat interval. Range:150 to 5000. Default: 190.

```
switch(config)# ngoam xconnect hb-interval 200
```

Configure the VXLAN Cross Connect

Use this procedure to configure the VXLAN Cross Connect feature on Cisco Nexus 9000 devices.

Procedure

Step 1 Follow these steps to create and configure xConnect VLANs:

a) Enter the **vlan** *session-num* command in global configuration mode, to create a VLAN.

Example:

```
switch(config)# vlan 550
switch(config-vlan)#
```

b) Enter the **vn-segment** *segment-num* command in VLAN configuration mode, to specify VXLAN VNID (Virtual Network Identifier).

Example:

```
switch(config-vlan)# vn-segment 5555
```

c) Enter the **xconnect** command in VLAN configuration mode, to define the provider VLAN with the attached VNI to be in cross connect mode.

Example:

```
switch(config-vlan)# xconnect
```

d) Enter the **exit** command in VLAN configuration mode, to exit the config mode and apply the configuration to the switch.

Example:

```
switch(config-vlan)# exit
switch(config)#
```

Step 2 Follow these steps to configure xConnect on a interface access ports:

a) Enter the **interface** [*physical-interface* | *port-channel*] command in global configuration mode, to create a VLAN.

Example:

For physical interface:

```
switch(config)# interface Ethernet1/30/3
```

For port-channel:

```
switch(config)# interface port-channel 550
```

b) Enter the **switchport** command in interface configuration mode, to enable the switchport mode.

Example:

```
switch(config-if)# switchport
```

c) Enter the **switchport mode dot1q-tunnel** command in interface configuration mode, to create a 802.1Q tunnel on the port.

Example:

```
switch(config-if)# switchport mode dot1q-tunnel
```

d) Enter the **switchport access vlan** *vlan-id* command in interface configuration mode, to specify the port assigned to a xConnect VLAN.

Example:

```
switch(config-if)# switchport access vlan 660
```

e) Enter the **exit** command to exit the config mode and apply the configuration to the switch.

Example:

```
switch(config-if)# exit
switch(config)#
```

Step 3 Run the **show running-config vlan** *session-num* command to verify the VLAN information.

Example:

```
switch# show running-config vlan 550

!Command: show running-config vlan 550
!Running configuration last done at: Sun Feb 23 16:34:03 2025
!Time: Mon Feb 24 13:53:01 2025

version 10.5(2) Bios:version 05.51
vlan 550
```

```
vlan 550
  vn-segment 5555
  xconnect
```

Step 4 Run the **show running-config interface port-channel** *pc-num* command to verify interface port channel information.

Example:

```
switch# show running-config interface port-channel 550

!Command: show running-config interface port-channel550
!Running configuration last done at: Sun Feb 23 16:34:03 2025
!Time: Mon Feb 24 13:53:30 2025

version 10.5(2) Bios:version 05.51

interface port-channel550
  switchport
  switchport mode dot1q-tunnel
  switchport access vlan 550
  spanning-tree bpdufilter enable
switch#
```

Step 5 Run the **show running-config interface Ethernet** *port-num* command to verify interface access ports information.

Example:

```
switch# show running-config interface Ethernet1/30/3

!Command: show running-config interface port-channel550
!Running configuration last done at: Sun Feb 23 16:34:03 2025
!Time: Mon Feb 24 13:53:30 2025

version 10.5(2) Bios:version 05.51

interface port-channel550
  switchport
  switchport mode dot1q-tunnel
  switchport access vlan 550
  spanning-tree bpdufilter enable
switch#
```

Verification of VXLAN Cross Connect Configuration

To display the status for the VXLAN Cross Connect configuration, enter one of the following commands:

Command	Purpose
show nve vni	Displays list of all VXLAN VNI and its status.
show nve vni *vlan-id*	Displays VNI status for specified VNI.
show ngoam xconnect session all	Displays all the xConnect session and its status.
show ngoam xconnect session *vlan-id*	Displays the specified xConnect session with details.

- Example of **show nve vni** command: Once the xConnect enabled VLANs are tagged on the NVE interface, you can see the **Conn** flag for the xConnect enabled VLANs.

```
switch# show nve vni
Codes: CP - Control Plane        DP - Data Plane
       UC - Unconfigured         SA - Suppress ARP
       S-ND - Suppress ND
       SU - Suppress Unknown Unicast
       Xconn - Crossconnect
       MS-IR - Multisite Ingress Replication
       HYB - Hybrid IRB mode

Interface VNI      Multicast-group   State Mode Type [BD/VRF]      Flags
--------- -------- ----------------- ----- ---- ------------------ -----
nve1      5555     225.5.5.5         Up    CP   L2 [550]           Xconn
nve1      6666     225.6.6.6         Up    CP   L2 [660]           Xconn
nve1      7777     225.7.7.7         Up    CP   L2 [770]           Xconn
nve1      10101    225.1.1.1         Up    CP   L2 [101]           SA
nve1      10102    225.1.1.2         Up    CP   L2 [102]           SA
nve1      10103    225.1.1.3         Up    CP   L2 [103]           SA
nve1      10104    225.1.1.4         Up    CP   L2 [104]           SA
nve1      10105    225.1.1.5         Up    CP   L2 [105]           SA
nve1      10106    225.1.1.6         Up    CP   L2 [106]           SA
nve1      10107    225.1.1.7         Up    CP   L2 [107]           SA
nve1      10108    225.1.1.8         Up    CP   L2 [108]           SA
nve1      10109    225.1.1.9         Up    CP   L2 [109]           SA
nve1      10110    225.1.1.10        Up    CP   L2 [110]           SA
nve1      20201    225.1.1.1         Up    CP   L2 [201]           SA
nve1      20202    225.1.1.2         Up    CP   L2 [202]           SA
nve1      20203    225.1.1.3         Up    CP   L2 [203]           SA
nve1      20204    225.1.1.4         Up    CP   L2 [204]           SA
nve1      20205    225.1.1.5         Up    CP   L2 [205]           SA
nve1      20206    225.1.1.6         Up    CP   L2 [206]           SA
nve1      20207    225.1.1.7         Up    CP   L2 [207]           SA
nve1      20208    225.1.1.8         Up    CP   L2 [208]           SA
nve1      20209    225.1.1.9         Up    CP   L2 [209]           SA
nve1      20210    225.1.1.10        Up    CP   L2 [210]           SA
nve1      202020   UnicastBGP        Up    CP   L2 [20]            Xconn
nve1      1001001  225.1.1.1         Up    CP   L2 [1001]          SA
nve1      1001002  225.1.1.2         Up    CP   L2 [1002]          SA
nve1      1001003  225.1.1.3         Up    CP   L2 [1003]          SA
nve1      1001004  225.1.1.4         Up    CP   L2 [1004]          SA
nve1      1001005  225.1.1.5         Up    CP   L2 [1005]          SA
nve1      1001006  225.1.1.6         Up    CP   L2 [1006]          SA
nve1      1001007  225.1.1.7         Up    CP   L2 [1007]          SA
nve1      1001008  225.1.1.8         Up    CP   L2 [1008]          SA
nve1      1001009  225.1.1.9         Up    CP   L2 [1009]          SA
nve1      1001010  225.1.1.10        Up    CP   L2 [1010]          SA
nve1      2002001  225.1.1.1         Up    CP   L2 [2001]          SA
nve1      2002002  225.1.1.2         Up    CP   L2 [2002]          SA
nve1      2002003  225.1.1.3         Up    CP   L2 [2003]          SA
nve1      2002004  225.1.1.4         Up    CP   L2 [2004]          SA
nve1      2002005  225.1.1.5         Up    CP   L2 [2005]          SA
nve1      2002006  225.1.1.6         Up    CP   L2 [2006]          SA
nve1      2002007  225.1.1.7         Up    CP   L2 [2007]          SA
nve1      2002008  225.1.1.8         Up    CP   L2 [2008]          SA
nve1      2002009  225.1.1.9         Up    CP   L2 [2009]          SA
nve1      2002010  225.1.1.10        Up    CP   L2 [2010]          SA
nve1      5005001  n/a               Up    CP   L3 [vxlan-5001]
nve1      5005002  n/a               Up    CP   L3 [vxlan-5002]
nve1      5005003  n/a               Up    CP   L3 [vxlan-5003]
nve1      5005004  n/a               Up    CP   L3 [vxlan-5004]
nve1      5005005  n/a               Up    CP   L3 [vxlan-5005]
nve1      5005006  n/a               Up    CP   L3 [vxlan-5006]
nve1      5005007  n/a               Up    CP   L3 [vxlan-5007]
nve1      5005008  n/a               Up    CP   L3 [vxlan-5008]
nve1      5005009  n/a               Up    CP   L3 [vxlan-5009]
```

```
nve1      5005010  n/a               Up    CP   L3 [vxlan-5010]
switch#
```

- Example of **show ngoam xconnect session all** command: use the command to check all the NGOAM xConnect session on the switch.

```
switch# show ngoam xconnect session all

States: LD = Local interface down, RD = Remote interface Down
          HB = Heartbeat lost, DB = Database/Routes not present
          * - Showing Vpc-peer interface info
Vlan           Peer-ip/vni        XC-State      Local-if/State        Rmt-if/State
================================================================================
20      100.100.100.8 / 202020      Active      Eth1/7/1 / UP       Eth1/50 / UP
550   200.200.200.200 / 5555       Active        Po550 / UP     Eth1/63/4 / UP
660     100.100.100.8 / 6666       Active     Eth1/30/3 / UP       Eth1/48 / UP
770   200.200.200.200 / 7777       Active        Po770 / UP         Po770 / UP
switch#
```

- Use the following commands to check the details of the specific NGOAM xConnect session on the switch.
  - The below example shows the output of vPC pairs as local interface and remote interface.

```
switch# sh ngoam xconnect session 550
Vlan ID: 550
Peer IP: 200.200.200.200  VNI : 5555
State: Active
Last state update: 02/23/2025 22:21:01.958
Local interface: Po550  State:  UP
Local vpc interface: Po550  State:  UP
Remote interface: Eth1/63/4  State:  UP
Remote vpc interface: Eth1/63/2  State:  UP
```

  - The below example shows the local interface as vPC and the remote interface as non-vPC pairs.

```
switch# sh ngoam xconnect session 660
Vlan ID: 660
Peer IP: 100.100.100.8  VNI : 6666
State: Active
Last state update: 02/23/2025 15:58:14.735
Local interface: Eth1/30/3  State:  UP
Local vpc interface: Eth1/30/3  State:  UP
Remote interface: Eth1/48  State:  UP
Remote vpc interface: Unknown  State:  DOWN
switch#
```

Remove a Cross Connect VNI

Use this procedure to remove the cross connect tag on the VLAN of Cisco Nexus switches.

Procedure

Step 1 Enter the **vlan** *session-num* command in global configurtion mode, to remove the VNI under NVE.

Example:

```
switch# configure terminal
switch(config)# vlan 550
switch(config-vlan)#
```

Step 2 Enter the **no xconnect** command in VLAN configurtion mode, to remove the xconnect tag from the VLAN configurations.

Example:

```
switch(config-vlan)# no xconnect
```

Note
Removing the xconnect tag is mandatory before removing the vn-segment for the xconnect enabled VLANs. Otherwise, you can see the below syslog that indicates the procedure.

```
xconnect is enabled on vlan 550, please disable xconnect before removing vn-segment
Cannot run commands in the mode at this moment. Please try again.
```

Step 3 Enter the **no vn-segment** *segment-num* command in VLAN configurtion mode, to remove the VNI.

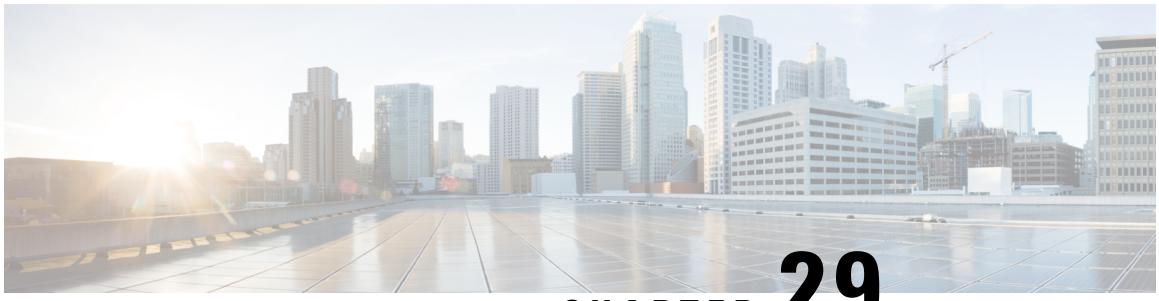
Example:

```
switch(config-vlan)# no vn-segment
```

Step 4 Enter the **no vlan** *session-num* command in VLAN configurtion mode, to remove the VLAN.

Example:

```
switch(config-vlan)# no vlan 550
```



CHAPTER 29

Configuring Q-in-VNI over VXLAN

- Q-in-VNI over VXLAN, on page 683
- Guidelines and Limitations for Q-in-VNI, on page 683
- Configuring Q-in-VNI, on page 686
- Configuring Selective Q-in-VNI, on page 687
- Configuring Q-in-VNI with Layer 2 Protocol Tunneling, on page 689
- Configuring Q-in-VNI with LACP Tunneling, on page 693
- Selective Q-in-VNI with Multiple Provider VLANs, on page 696
- Configuring QinQ-QinVNI, on page 699

Q-in-VNI over VXLAN

In traditional network segmentation, the number of segments that can be created is limited to 4096 VLANs. This can be insufficient in situations, such as cloud computing environments where multiple tenants or customers may require their own virtual networks. VxLAN solves the problem of limited network segmentation and isolation in traditional VLAN-based networks.

Q-in-VNI over VxLAN addresses the requirement of limited network segmentation and isolation by stacking multiple VLANs above each other, allowing for even more virtual network segmentation and isolation. This provides a large number of virtual networks to be created, with the necessary flexibility and scalability for cloud computing environments and other situations where many virtual networks are required.

In summary, Q-in-VNI deployment using VxLAN EVPN as the transport network provides a highly efficient and scalable solution for delivering Layer 2 VPN services. It allows users to maintain their existing VLAN structure while connecting to the service provider's network and provides high availability and redundancy with VxLAN EVPN's robust control plane.

For more information on Q-in-VNI over VXLAN fabric deployment, see **Q-in-VNI over VxLAN Fabric Deployment Guide**.

Guidelines and Limitations for Q-in-VNI

Q-in-VNI has the following guidelines and limitations:

Configuration guidelines and limitations

- The **system dot1q-tunnel transit** [**vlan** *vlan-range*] command is required when running this feature on vPC VTEPs.
- Port VLAN mapping and Q-in-VNI cannot coexist on the same port.
- Port VLAN mapping and Q-in-VNI cannot coexist on a switch if the **system dot1q-tunnel transit** command is enabled. Beginning with Cisco NX-OS Release 9.3(5), port VLAN mapping and Q-in-VNI can coexist on the same switch but on different ports and different provider VLANs, which are configured using the **system dot1q-tunnel transit vlan** *vlan-range* command.
- For proper operation during L3 uplink failure scenarios on vPC VTEPs, configure a backup SVI and enter the **system nve infra-vlans** *backup-svi-vlan* command.
- When configuring access ports and trunk ports for Cisco Nexus 9000 Series switches with a Leaf Spine Engine (LSE), you can have access ports, trunk ports, and dot1q ports on different interfaces on the same switch.
- You cannot have the same VLAN configured for both dot1q and trunk ports/access ports.
- Disable ARP suppression on the provider VNI for ARP traffic originated from a customer VLAN in order to flow.

```
switch(config)# interface nve 1
switch(config-if-nve)# member VNI 10000011
switch(config-if-nve-vni)# no suppress-arp
```

- Q-in-VNI cannot coexist with a VTEP that has Layer 3 subinterfaces configured. Beginning with Cisco NX-OS Release 9.3(5), this limitation no longer applies to Cisco Nexus 9300-FX/FX2, and 9300-GX platform switches.
- When VLAN1 is configured as the native VLAN with selective Q-in-VNI with the multiple provider tag, traffic on the native VLAN gets dropped. Do not configure VLAN1 as the native VLAN when the port is configured with selective Q-in-VNI. When VLAN1 is configured as a customer VLAN, the traffic on VLAN1 gets dropped.
- The base port mode must be a dot1q tunnel port with an access VLAN configured.
- VNI mapping is required for the access VLAN on the port.

Supported platforms and features

- Cisco Nexus 9300 platform switches support single tag. You can enable it by entering the **no overlay-encapsulation vxlan-with-tag** command for the NVE interface:

```
switch(config)# interface nve 1
switch(config-if-nve)# no overlay-encapsulation vxlan-with-tag
switch# show run int nve 1

!Command: show running-config interface nve1
!Time: Wed Jul 20 23:26:25 2016

version 7.0(3u)I4(2u)

interface nve1
  no shutdown
  source-interface loopback0
  host-reachability protocol bgp
  member vni 900001 associate-vrf
```

```
member vni 2000980
  mcast-group 225.4.0.1
```

- Beginning with Cisco NX-OS Release 10.1(1), Selective Q-in-VNI and VXLAN VLAN on Same Port feature is supported on Cisco Nexus 9300-FX3 platform switches.
- Q-in-VNI only supports VXLAN bridging. It does not support VXLAN routing.
- Q-in-VNI and selective Q-in-VNI are supported with VXLAN Flood and Learn with Ingress Replication and VXLAN EVPN with Ingress Replication.
- Beginning with Cisco NX-OS Release 10.2(3)F, the Cisco Nexus 9300-FX3/GX2 platform switches support Q-in-VNI to coexist with a VTEP that has Layer 3 subinterfaces configured.
- Beginning with Cisco NX-OS Release 10.4(1)F, the Cisco Nexus 9332D-H2R switches support Q-in-VNI to coexist with a VTEP that has Layer 3 subinterfaces configured.
- Beginning with Cisco NX-OS Release 10.4(2)F, the Cisco Nexus 93400LD-H1 switches support Q-in-VNI to coexist with a VTEP that has Layer 3 subinterfaces configured.
- Beginning with Cisco NX-OS Release 10.4(3)F, the Cisco Nexus 9364C-H1 switches support Q-in-VNI to coexist with a VTEP that has Layer 3 subinterfaces configured.
- Beginning with Cisco NX-OS Release 9.3(5), Q-in-VNI is supported on Cisco Nexus 9300-GX platform switches.
- Beginning with Cisco NX-OS Release 10.2(3)F, Q-in-VNI is supported on the Cisco Nexus 9300-GX2 platform switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, Q-in-VNI is supported on the Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, Q-in-VNI is supported on the Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, Q-in-VNI is supported on the Cisco Nexus 9364C-H1 switches.
- Beginning with Cisco NX-OS Release 9.3(5), Q-in-VNI supports vPC Fabric Peering.
- Beginning with Cisco NX-OS Release 10.3(3)F, IPv6 underlay is supported on Q-in-VNI, Selective Q-in-VNI and Q-in-Q-Q-in-VNI for VXLAN EVPN on Cisco Nexus 9300-FX/FX2/FX3/GX/GX2 switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, IPv6 underlay is supported on Q-in-VNI, Selective Q-in-VNI and Q-in-Q-Q-in-VNI for VXLAN EVPN on Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, IPv6 underlay is supported on Q-in-VNI, Selective Q-in-VNI and Q-in-Q-Q-in-VNI for VXLAN EVPN on Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, IPv6 underlay is supported on Q-in-VNI, Selective Q-in-VNI and Q-in-Q-Q-in-VNI for VXLAN EVPN on Cisco Nexus 9364C-H1 switches.

Unsupported platforms and features

- The dot1q tunnel mode does not support ALE ports on Cisco Nexus 9300 Series and Cisco Nexus 9500 platform switches.

- Q-in-VNI does not support FEX.
- Q-in-VNI is not supported as part of multi-site solution.
- Q-in-VNI and Selective Q-in-VNI are not supported on Cisco Nexus 9500 Series switches with X97160YC-EX and 9700-FX/FX3/GX line cards.

Configuring Q-in-VNI

Using Q-in-VNI provides a way for you to segregate traffic by mapping to a specific port. In a multi-tenant environment, you can specify a port to a tenant and send/receive packets over the VXLAN overlay.

SUMMARY STEPS

1. **configure terminal**
2. **interface** *type port*
3. **switchport mode dot1q-tunnel**
4. **switchport access vlan** *vlan-id*
5. **spanning-tree bpdufilter enable**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**interface** *type port*	Enters interface configuration mode.
Step 3	**switchport mode dot1q-tunnel**	Creates a 802.1Q tunnel on the port.
Step 4	**switchport access vlan** *vlan-id*	Specifies the port assigned to a VLAN.
Step 5	**spanning-tree bpdufilter enable**	Enables BPDU Filtering for the specified spanning tree edge interface. By default, BPDU Filtering is disabled.

Example

The following is an example of configuring Q-in-VNI:

```
switch# config terminal
switch(config)# interface ethernet 1/4
switch(config-if)# switchport mode dot1q-tunnel
switch(config-if)# switchport access vlan 10
switch(config-if)# spanning-tree bpdufilter enable
switch(config-if)#
```

Configuring Selective Q-in-VNI

Selective Q-in-VNI is a VXLAN tunneling feature that allows a user specific range of customer VLANs on a port to be associated with one specific provider VLAN. Packets that come in with a VLAN tag that matches any of the configured customer VLANs on the port are tunneled across the VXLAN fabric using the properties of the service provider VNI. The VXLAN encapsulated packet carries the customer VLAN tag as part of the L2 header of the inner packet.

The packets that come in with a VLAN tag that is not present in the range of the configured customer VLANs on a selective Q-in-VNI configured port are dropped. This includes the packets that come in with a VLAN tag that matches the native VLAN on the port. Packets coming untagged or with a native VLAN tag are L3 routed using the native VLAN's SVI that is configured on the selective Q-in-VNI port (no VXLAN).

See the following guidelines for selective Q-in-VNI:

- Selective Q-in-VNI is supported on both vPC and non-vPC ports on Cisco Nexus 9300-FX/FXP/FX2/FX3 and 9300-GX platform switches.
- Beginning with Cisco NX-OS Release 9.3(5), selective Q-in-VNI supports vPC Fabric Peering.
- Configuring selective Q-in-VNI on one VTEP and configuring plain Q-in-VNI on the VXLAN peer is supported. Configuring one port with selective Q-in-VNI and the other port with plain Q-in-VNI on the same switch is supported.
- Selective Q-in-VNI is an ingress VLAN tag-policing feature. Only ingress VLAN tag policing is performed with respect to the selective Q-in-VNI configured range.

  For example, selective Q-in-VNI customer VLAN range of 100-200 is configured on VTEP1 and customer VLAN range of 200-300 is configured on VTEP2. When traffic with VLAN tag of 175 is sent from VTEP1 to VTEP2, the traffic is accepted on VTEP1, since the VLAN is in the configured range and it is forwarded to the VTEP2. On VTEP2, even though VLAN tag 175 is not part of the configured range, the packet egresses out of the selective Q-in-VNI port. If a packet is sent with VLAN tag 300 from VTEP1, it is dropped because 300 is not in VTEP1's selective Q-in-VNI configured range.
- Beginning with Cisco NX-OS Release 10.1(1), Selective Q-in-VNI and Advertise PIP on a VTEP feature is supported on Cisco Nexus 9300-FX3 platform switches.
- Beginning with Cisco NX-OS Release 9.3(5), the **advertise-pip** command is supported with selective Q-in-VNI on a VTEP.
- Port VLAN mapping and selective Q-in-VNI cannot coexist on the same port.
- Port VLAN mapping and selective Q-in-VNI cannot coexist on a switch if the **system dot1q-tunnel transit** command is enabled. Beginning with Cisco NX-OS Release 9.3(5), port VLAN mapping and Q-in-VNI can coexist on the same switch but on different ports and different provider VLANs, which are configured using the **system dot1q-tunnel transit vlan** *vlan-range* command.
- Configure the **system dot1q-tunnel transit** [**vlan** *vlan-id*] command on vPC switches with selective Q-in-VNI configurations. This command is required to retain the inner Q-tag as the packet goes over the vPC peer link when one of the vPC peers has an orphan port. With this CLI configuration, the **vlan dot1Q tag native** functionality does not work. Prior to Cisco NX-OS Release 9.3(5), every VLAN created on the switch is a provider VLAN and cannot be used for any other purpose.

  Beginning with Cisco NX-OS Release 9.3(5), selective Q-in-VNI and VXLAN VLANs can be supported on the same port. With the [**vlan** *vlan-range*] option, you can specify the provider VLANs and allow

other VLANs to be used for regular VXLAN traffic. In the following example, the VXLAN VLAN is 50, the provider VLAN is 501, the customer VLANs are 31-40, and the native VLAN is 2400.

```
system dot1q-tunnel transit vlan 501
interface Ethernet1/1/2
  switchport
  switchport mode trunk
  switchport trunk native vlan 2400
  switchport vlan mapping 31-40 dot1q-tunnel 501
  switchport trunk allowed vlan 50,501,2400
  spanning-tree port type edge trunk
  mtu 9216
  no shutdown
```

- The native VLAN configured on the selective Q-in-VNI port cannot be a part of the customer VLAN range. If the native VLAN is part of the customer VLAN range, the configuration is rejected.

  The provider VLAN can overlap with the customer VLAN range. For example, **switchport vlan mapping 100-1000 dot1q-tunnel 200**.

- By default, the native VLAN on any port is VLAN 1. If VLAN 1 is configured as part of the customer VLAN range using the **switchport vlan mapping** <*range*>**dot1q-tunnel** <*sp-vlan*> CLI command, the traffic with customer VLAN 1 is not carried over as VLAN 1 is the native VLAN on the port. If customer wants VLAN 1 traffic to be carried over the VXLAN cloud, they should configure a dummy native VLAN on the port whose value is outside the customer VLAN range.

- To remove some VLANs or a range of VLANs from the configured switchport VLAN mapping range on the selective Q-in-VNI port, use the **no** form of the **switchport vlan mapping** <*range*>**dot1q-tunnel** <*sp-vlan*> command.

  For example, VLAN 100-1000 is configured on the port. To remove VLAN 200-300 from the configured range, use the **no switchport vlan mapping** <*200-300*> **dot1q-tunnel** <*sp-vlan*> command.

```
interface Ethernet1/32
  switchport
  switchport mode trunk
  switchport trunk native vlan 4049
  switchport vlan mapping 100-1000 dot1q-tunnel 21
  switchport trunk allowed vlan 21,4049
  spanning-tree bpdufilter enable
  no shutdown

switch(config-if)# no sw vlan mapp 200-300 dot1q-tunnel 21
switch(config-if)# sh run int e 1/32

version 7.0(3)I5(2)

interface Ethernet1/32
  switchport
  switchport mode trunk
  switchport trunk native vlan 4049
  switchport vlan mapping 100-199,301-1000 dot1q-tunnel 21
  switchport trunk allowed vlan 21,4049
  spanning-tree bpdufilter enable
  no shutdown
```

See the following configuration examples.

- See the following example for the provider VLAN configuration:

```
vlan 50
```

```
 vn-segment 10050
```

- See the following example for configuring VXLAN Flood and Learn with Ingress Replication:

```
member vni 10050
    ingress-replication protocol static
      peer-ip 100.1.1.3
      peer-ip 100.1.1.5
      peer-ip 100.1.1.10
```

- See the following example for the interface nve configuration:

```
interface nve1
 no shutdown
 source-interface loopback0  member vni 10050
mcast-group 230.1.1.1
```

- See the following example for configuring an SVI in the native VLAN to routed traffic.

```
vlan 150
interface vlan150
 no shutdown
 ip address 150.1.150.6/24
 ip pim sparse-mode
```

- See the following example for configuring selective Q-in-VNI on a port. In this example, native VLAN 150 is used for routing the untagged packets. Customer VLANs 200-700 are carried across the dot1q tunnel. The native VLAN 150 and the provider VLAN 50 are the only VLANs allowed.

```
switch# config terminal
switch(config)#interface Ethernet 1/31
switch(config-if)#switchport
switch(config-if)#switchport mode trunk
switch(config-if)#switchport trunk native vlan 150
switch(config-if)#switchport vlan mapping 200-700 dot1q-tunnel 50
switch(config-if)#switchport trunk allowed vlan 50,150
switch(config-if)#no shutdown
```

- Disable ARP suppression on the provider VNI for ARP traffic originated from a customer VLAN in order to flow.

```
switch(config)# interface nve 1
switch(config-if-nve)# member VNI 10000011
 switch(config-if-nve-vni)# no suppress-arp
```

Configuring Q-in-VNI with Layer 2 Protocol Tunneling

Q-in-VNI with L2PT Overview

Q-in-VNI with Layer 2 Protocol Tunneling (L2PT) is used to transport control and data packets across a VXLAN EVPN fabric for multi-tagged traffic.

To enable Q-in-VNI with L2PT at the VLAN level, use the **l2protocol tunnel vxlan vlan** <*vlan-range*> command which marks the VLANs for tunneling all packets including L2 protocol packets. The **switchport trunk allow-multi-tag** command is also required for the VXLAN fabric to tunnel packets with multiple tags.

For more information on Q-in-VNI with L2PT configuration, refer to Configuring Q-in-VNI with L2PT, on page 690.

Guidelines and Limitations for Q-in-VNI with L2PT

Q-in-VNI with L2PT has the following guidelines and limitations:

- Beginning with Cisco NX-OS Release 10.3(2)F, Q-in-VNI with L2PT is supported on Cisco Nexus 9300-FX/FX2/FX3/GX/GX2 ToR switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, Q-in-VNI with L2PT is supported on Cisco Nexus 9332D-H2R ToR switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, Q-in-VNI with L2PT is supported on Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, Q-in-VNI with L2PT is supported on Cisco Nexus 9364C-H1 switches.
- Once the **l2protocol tunnel vxlan** command is run on an interface, all VLANs in the command become tunneling VLANs and cannot be used on any other port for any other purpose.
- Only two interfaces in the network can be member of the tunnel VLAN. For vPC cases, both vPC ports on the vPC switches and MCT will also be part of the tunnel VLAN.
- Same VLAN must not be tunneled on multiple interfaces.
- The **l2protocol tunnel vxlan** command is allowed only on trunk ports. It also requires “multi-tag” configuration to preserve the multiple tags across the vxlan fabric.
- Cross Connect feature and **l2protocol tunnel vxlan** command can not be used together on a switch.
- Existing L2PT command options like "STP" can not be used along with the **l2protocol tunnel vxlan** command.
- Beginning with Cisco NX-OS Release 10.3(3)F, Ethertype support for Q-in-VNI with L2PT is provided on Cisco Nexus 9300-FX2/FX3/GX/GX2 ToR switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, Ethertype support for Q-in-VNI with L2PT is provided on Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, Ethertype support for Q-in-VNI with L2PT is provided on Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, Ethertype support for Q-in-VNI with L2PT is provided on Cisco Nexus 9364C-H1 switches.

Configuring Q-in-VNI with L2PT

Follow this procedure to configure the Q-in-VNI with L2PT on VXLAN VLAN:

SUMMARY STEPS

1. **configure terminal**
2. **interface ethernet** *slot/port*
3. **switchport**
4. **switchport mode trunk**
5. **switchport dot1q ethertype** *ethertype-value*
6. **switchport trunk allow-multi-tag**
7. **switchport trunk allowed vlan** *vlan-list*
8. **l2protocol tunnel vxlan vlan** <*vlan-range*>

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**interface ethernet** *slot/port* **Example:** `switch(config)# interface ethernet1/1`	Specifies the interface that you are configuring.
Step 3	**switchport** **Example:** `switch(config-inf)# switchport`	Configures it as a Layer 2 port.
Step 4	**switchport mode trunk** **Example:** `switch(config-inf)# switchport mode trunk`	Sets the interface as a Layer 2 trunk port.
Step 5	**switchport dot1q ethertype** *ethertype-value* **Example:** `switch(config-inf)# switchport dot1q ethertype 0x88a8`	Sets the Ethertype for the port.
Step 6	**switchport trunk allow-multi-tag** **Example:** `switch(config-inf)# switchport trunk allow-multi-tag`	Sets the allowed VLANs as the provider VLANs excluding the native VLAN. In the config example provided, VLANs 1201 and 1202 are the provider VLANs and can carry multiple inner Q-tags.
Step 7	**switchport trunk allowed vlan** *vlan-list* **Example:** `switch(config-inf)# switchport trunk allowed vlan 1201-1202`	Sets the allowed VLANs for the trunk interface.

	Command or Action	Purpose
Step 8	**l2protocol tunnel vxlan vlan** <*vlan-range*> **Example:** `switch(config-inf)# l2protocol tunnel vxlan vlan 1201-1202`	Sets all VLANs in the command as tunneling VLANs. These VLANs cannot be used on any other port for any other purpose.

Verifying Q-in-VNI with L2PT Configuration

To display the status for the Q-in-VNI with L2PT configuration, enter one of the following commands:

Command	Purpose
show run interface ethernet *slot/port*	Displays L2PT VXLAN VLAN interface information.
show run l2pt	Displays L2PT VXLAN VLAN configuration information.
show l2protocol tunnel interface ethernet *slot/port*	Displays L2PT interface information.
show vpc consistency-parameters interface *slot/port*	Displays the status of the parameters that must be consistent across all vPC interfaces including L2PT VXLAN VLAN.

The following example shows sample output for the **show run interface ethernet** *slot/port* command:

```
switch(config-if)# sh run int e1/1
 interface Ethernet1/1
  switchport
  switchport mode trunk
  switchport trunk allow-multi-tag
  switchport trunk allowed vlan 1201-1202
  l2protocol tunnel vxlan vlan 1201-1202
  no shutdown
```

The following example shows sample output for the **show run l2pt** command:

```
switch# sh run l2pt
interface Ethernet1/1
  switchport mode trunk
  l2protocol tunnel vxlan vlan 1201-1202
  no shutdown
```

The following example shows sample output for the **show l2protocol tunnel interface ethernet** *slot/port* command:

```
switch# show l2protocol tunnel interface e1/1
COS for Encapsulated Packets: 5
Interface: Eth1/1 Vxlan Vlan 1201-1202
```

The following example shows sample output for the **show vpc consistency-parameters interface** *slot/port* command:

```
switch# sh run int po101

interface port-channel101
  switchport
  switchport mode trunk
  switchport trunk native vlan 80
```

```
  switchport trunk allow-multi-tag
  switchport trunk allowed vlan 80,1201-1203,1301
  spanning-tree port type edge trunk
  vpc 101
  l2protocol tunnel vxlan vlan 1201-1203,1301

switch# sh vpc consistency-parameters interface po101

    Legend:
        Type 1 : vPC will be suspended in case of mismatch

Name                        Type  Local Value            Peer Value
-------------               ----  ---------------------- -----------------------
delayed-lacp                1     disabled               disabled
lacp suspend disable        1     enabled                enabled
mode                        1     active                 active
Switchport Isolated         1     0                      0
Interface type              1     port-channel           port-channel
LACP Mode                   1     on                     on
Virtual-ethernet-bridge     1     Disabled               Disabled
Speed                       1     25 Gb/s                25 Gb/s
Duplex                      1     full                   full
MTU                         1     1500                   1500
Port Mode                   1     trunk                  trunk
Native Vlan                 1     80                     80
Admin port mode             1     trunk                  trunk
Port-type External          1     Disabled               Disabled
STP Port Guard              1     Default                Default
STP Port Type               1     Edge Trunk Port        Edge Trunk Port
STP MST Simulate PVST       1     Default                Default
lag-id                      1     [(7f9b,                [(7f9b,
                                  0-23-4-ee-be-4, 8065,  0-23-4-ee-be-4, 8065,
                                  0, 0), (8000,          0, 0), (8000,
                                  a8-9d-21-f8-4b-31, 64, a8-9d-21-f8-4b-31, 64,
                                   0, 0)]                 0, 0)]
Allow-Multi-Tag             1     Enabled                Enabled
Vlan xlt mapping            1     Disabled               Disabled
L2PT Vxlan Vlans            2     1201-1203,1301         1201-1203,1301
vPC card type               1     N9K TOR                N9K TOR
Allowed VLANs               -     80,1201-1203,1301      80,1201-1203,1301
Local suspended VLANs       -     -                      -
```

Configuring Q-in-VNI with LACP Tunneling

Q-in-VNI can be configured to tunnel LACP packets.

SUMMARY STEPS

1. **configure terminal**
2. **interface** *type port*
3. **switchport mode dot1q-tunnel**
4. **switchport access vlan** *vlan-id*
5. **interface nve** *x*
6. **overlay-encapsulation vxlan-with-tag tunnel-control-frames lacp**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**interface** *type port*	Enters interface configuration mode.
Step 3	**switchport mode dot1q-tunnel**	Enables dot1q-tunnel mode.
Step 4	**switchport access vlan** *vlan-id*	Specifies the port assigned to a VLAN.
Step 5	**interface nve** *x*	Creates a VXLAN overlay interface that terminates VXLAN tunnels.
Step 6	**overlay-encapsulation vxlan-with-tag tunnel-control-frames lacp**	Enables Q-in-VNI for LACP tunneling. **Note** Use this form of the command for NX-OS 7.0(3)I3(1) and later releases. For NX-OS 7.0(3)I2(2) and earlier releases, use the **overlay-encapsulation vxlan-with-tag tunnel-control-frames** command.

Example

- The following is an example of configuring a Q-in-VNI for LACP tunneling (NX-OS 7.0(3)I2(2) and earlier releases):

```
switch# config terminal
switch(config)# interface ethernet 1/4
switch(config-if)# switchport mode dot1q-tunnel
switch(config-if)# switchport access vlan 10
switch(config-if)# spanning-tree bpdufilter enable
switch(config-if)# interface nve1
switch(config-if)# overlay-encapsulation vxlan-with-tag tunnel-control-frames
```

Note

- STP is disabled on VNI mapped VLANs.
- No spanning-tree VLAN <> on the VTEP.
- No MAC address-table notification for mac-move.
- As a best practice, configure a fast LACP rate on the interface where the LACP port is configured. Otherwise the convergence time is approximately 90 seconds.

- The following is an example of configuring a Q-in-VNI for LACP tunneling (NX-OS 7.0(3)I3(1) and later releases):

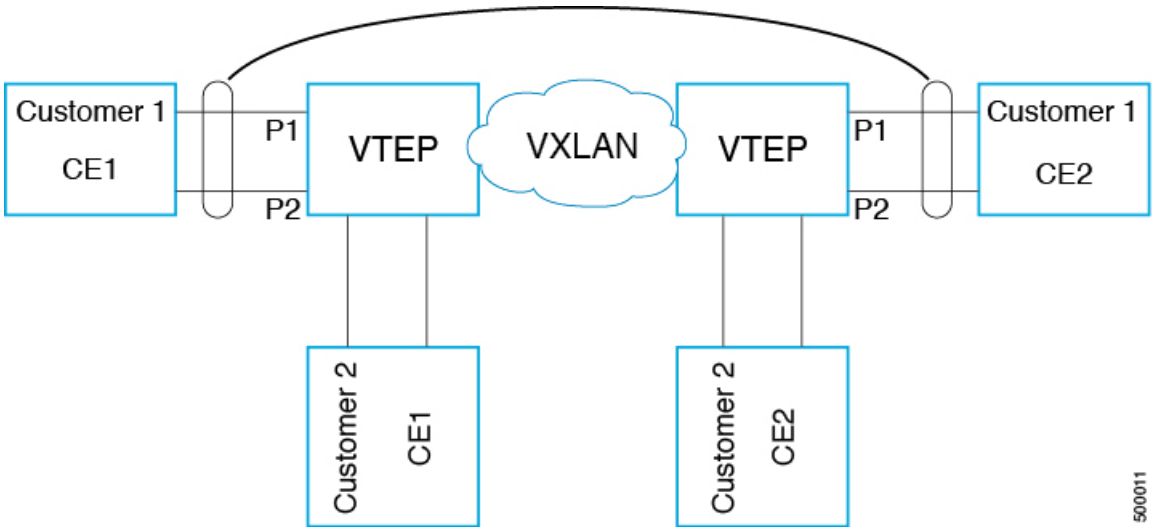
```
switch# config terminal
switch(config)# interface ethernet 1/4
switch(config-if)# switchport mode dot1q-tunnel
switch(config-if)# switchport access vlan 10
switch(config-if)# spanning-tree bpdufilter enable
switch(config-if)# interface nve1
switch(config-if)# overlay-encapsulation vxlan-with-tag tunnel-control-frames lacp
```

Note

- STP is disabled on VNI mapped VLANs.
- No spanning-tree VLAN <> on the VTEP.
- No MAC address-table notification for mac-move.
- As a best practice, configure a fast LACP rate on the interface where the LACP port is configured. Otherwise the convergence time is approximately 90 seconds.

- The following is an example topology that pins each port of a port-channel pair to a unique VM. The port-channel is stretched from the CE perspective. There is no port-channel on VTEP. The traffic on P1 of CE1 transits to P1 of CE2 using Q-in-VNI.

Figure 72: LACP Tunneling Over VXLAN P2P Tunnels



Note

- Q-in-VNI can be configured to tunnel LACP packets. (Able to provide port-channel connectivity across data-centers.)
  - Gives impression of L1 connectivity and co-location across data-centers.
  - Exactly two sites. Traffic coming from P1 of CE1 goes out of P1 of CE2. If P1 of CE1 goes down, LACP provides coverage (over time) to redirect traffic to P2.
- Uses static ingress replication with VXLAN with flood and learn. Each port of the port channel is configured with Q-in-VNI. There are multiple VNIs for each member of a port-channel and each port is pinned to specific VNI.
  - To avoid saturating the MAC, you should turn off/disable learning of VLANS.
- Configuring Q-in-VNI to tunnel LACP packets is not supported for VXLAN EVPN.
- The number of port-channel members supported is the number of ports supported by the VTEP.

Selective Q-in-VNI with Multiple Provider VLANs

About Selective Q-in-VNI with Multiple Provider VLANs

Selective Q-in-VNI with multiple provider VLANs is a VXLAN tunneling feature. This feature allows a user specific range of customer VLANs on a port to be associated with one specific provider VLAN. It also enables you to have multiple customer-VLAN to provider-VLAN mappings on a port. Packets that come in with a VLAN tag which matches any of the configured customer VLANs on the port are tunneled across the VXLAN fabric using the properties of the service provider VNI. The VXLAN encapsulated packet carries the customer VLAN tag as part of the Layer 2 header of the inner packet.

Guidelines and Limitations for Selective Q-in-VNI with Multiple Provider VLANs

Selective Q-in-VNI with multiple provider VLANs has the following guidelines and limitations:

- All the existing guidelines and limitations for Selective Q-in-VNI apply.
- This feature is supported with VXLAN BGP EVPN IR mode only.
- When enabling multiple provider VLANs on a vPC port channel, make sure that the configuration is consistent across the vPC peers.
- Port VLAN mapping and selective Q-in-VNI cannot coexist on the same port.
- Port VLAN mapping and selective Q-in-VNI cannot coexist on a switch if the **system dot1q-tunnel transit** command is enabled. Beginning with Cisco NX-OS Release 9.3(5), port VLAN mapping and selective Q-in-VNI can coexist on the same switch but on different ports and different provider VLANs, which are configured using the **system dot1q-tunnel transit vlan** *vlan-range* command.

- The **system dot1q-tunnel transit** [**vlan** *vlan-range*] command is required when using this feature on vPC VTEPs.
- For proper operation during Layer 3 uplink failure scenarios on vPC VTEPs, configure the backup SVI and enter the **system nve infra-vlans** *backup-svi-vlan* command.
- As a best practice, do not allow provider VLANs on a regular trunk.
- We recommend not creating or allowing customer VLANs on the switch where customer-VLAN to provider-VLAN mapping is configured.
- We do not support specific native VLAN configuration when the **switchport vlan mapping all dot1q-tunnel** command is entered.
- Beginning with Cisco NX-OS Release 9.3(5), selective Q-in-VNI with a multiple provider tag supports vPC Fabric Peering.
- Disable ARP suppression on the provider VNI for ARP traffic originated from a customer VLAN in order to flow.

```
switch(config)# interface nve 1
switch(config-if-nve)# member VNI 10000011
switch(config-if-nve-vni)# no suppress-arp
```

- All incoming traffic should be tagged when the interface is configured with the **switchport vlan mapping all dot1q-tunnel** command.

Configuring Selective Q-in-VNI with Multiple Provider VLANs

You can configure selective Q-in-VNI with multiple provider VLANs.

Before you begin

You must configure provider VLANs and associate the VLAN to a vn-segment.

SUMMARY STEPS

1. Enter global configuration mode.
2. Configure Layer 2 VLANs and associate them to a vn-segment.
3. Enter interface configuration mode where the traffic comes in with a dot1Q VLAN tag.

DETAILED STEPS

Procedure

Step 1 Enter global configuration mode.

```
switch# configure terminal
```

Step 2 Configure Layer 2 VLANs and associate them to a vn-segment.

```
switch(config)# vlan 10
vn-segment 10000010
```

```
switch(config)# vlan 20
 vn-segment 10000020
```

Step 3 Enter interface configuration mode where the traffic comes in with a dot1Q VLAN tag.

```
switch(config)# interf port-channel 10
switch(config-if)# switchport
switch(config-if)# switchport mode trunk
switch(config-if)# switchport trunk native vlan 3962
switch(config-if)# switchport vlan mapping 2-400 dot1q-tunnel 10
switch(config-if)# switchport vlan mapping 401-800 dot1q-tunnel 20
switch(config-if)# switchport vlan mapping 801-1200 dot1q-tunnel 30
switch(config-if)# switchport vlan mapping 1201-1600 dot1q-tunnel 40
switch(config-if)# switchport vlan mapping 1601-2000 dot1q-tunnel 50
switch(config-if)# switchport vlan mapping 2001-2400 dot1q-tunnel 60
switch(config-if)# switchport vlan mapping 2401-2800 dot1q-tunnel 70
switch(config-if)# switchport vlan mapping 2801-3200 dot1q-tunnel 80
switch(config-if)# switchport vlan mapping 3201-3600 dot1q-tunnel 90
switch(config-if)# switchport vlan mapping 3601-3960 dot1q-tunnel 100
switch(config-if)# switchport trunk allowed vlan 10,20,30,40,50,60,70,80,90,100,3961-3967
```

Example

This example shows how to configure Selective Qinvni with multiple provider VLANs:

```
switch# show run vlan 121
vlan 121
vlan 121
  vn-segment 10000021

switch#
switch# sh run interf port-channel 5

interface port-channel5
  description VPC PO
  switchport
  switchport mode trunk
  switchport trunk native vlan 504
  switchport vlan mapping 11 dot1q-tunnel 111
  switchport vlan mapping 12 dot1q-tunnel 112
  switchport vlan mapping 13 dot1q-tunnel 113
  switchport vlan mapping 14 dot1q-tunnel 114
  switchport vlan mapping 15 dot1q-tunnel 115
  switchport vlan mapping 16 dot1q-tunnel 116
  switchport vlan mapping 17 dot1q-tunnel 117
  switchport vlan mapping 18 dot1q-tunnel 118
  switchport vlan mapping 19 dot1q-tunnel 119
  switchport vlan mapping 20 dot1q-tunnel 120
  switchport trunk allowed vlan 111-120,500-505
  vpc 5

switch#

switch# sh spanning-tree vlan 111

VLAN0111
  Spanning tree enabled protocol rstp
  Root ID    Priority    32879
             Address     7079.b3cf.956d
             This bridge is the root
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec
```

```
  Bridge ID  Priority    32879  (priority 32768 sys-id-ext 111)
             Address     7079.b3cf.956d
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
Po1              Desg FWD 1         128.4096 (vPC peer-link) Network P2p
Po5              Desg FWD 1         128.4100 (vPC) P2p
Eth1/7/2         Desg FWD 10        128.26   P2p

switch#


switch# sh vlan internal info mapping | b Po5
  ifindex Po5(0x16000004)
  vlan mapping enabled: TRUE
  vlan translation mapping information (count=10):
    Original Vlan       Translated Vlan
    -------------       ---------------
    11                   111
    12                   112
    13                   113
    14                   114
    15                   115
    16                   116
    17                   117
    18                   118
    19                   119
    20                   120
switch#


switch# sh consistency-checker vxlan selective-qinvni interface port-channel 5
Performing port specific checks for intf port-channel5
Port specific selective QinVNI checks for interface port-channel5 : PASS
Performing port specific checks for intf port-channel5
Port specific selective QinVNI checks for interface port-channel5 : PASS

switch#
```

Configuring QinQ-QinVNI

Overview for QinQ-QinVNI

- QinQ-QinVNI is a VXLAN tunneling feature that allows you to configure a trunk port as a multi-tag port to preserve the customer VLANs that are carried across the network.
- On a port that is configured as multi-tag, packets are expected with multiple-tags or at least one tag. When multi-tag packets ingress on this port, the outer-most or first tag is treated as provider-tag or provider-vlan. The remaining tags are treated as customer-tag or customer-vlan.
- This feature is supported on both vPC and non-vPC ports.
- Ensure that the **switchport trunk allow-multi-tag** command is configured on both of the vPC-peers. It is a type 1 consistency check.
- This feature is supported with VXLAN Flood and Learn and VXLAN EVPN.

Guidelines and Limitations for QinQ-QinVNI

QinQ-QinVNI has the following guidelines and limitations:

- This feature is supported on the Cisco Nexus 9300-FX/FX2/FX3, and 9300-GX platform switches.
- Beginning with Cisco NX-OS Release 10.2(3)F, QinQ-QinVNI is supported on the Cisco Nexus 9300-GX2 platform switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, QinQ-QinVNI is supported on the Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, QinQ-QinVNI is supported on the Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, QinQ-QinVNI is supported on the Cisco Nexus 9364C-H1 switches.
- This feature supports vPC Fabric Peering.
- On a multi-tag port, provider VLANs must be a part of the port. They are used to derive the VNI for that packet.
- Untagged packets are associated with the native VLAN. If the native VLAN is not configured, the packet is associated with the default VLAN (VLAN 1).
- Packets coming in with an outermost VLAN tag (provider-vlan), not present in the range of allowed VLANs on a multi-tag port, are dropped.
- Packets coming in with an outermost VLAN tag (provider-vlan) tag matching the native VLAN are routed or bridged in the native VLAN's domain.
- This feature supports VXLAN bridging but does not support VXLAN routing.
- Multicast data traffic with more than two Q-Tags is not supported when snooping is enabled on the VXLAN VLAN.
- You need at least one multi-tag trunk port allowing the provider VLANs in Up state on both vPC peers. Otherwise, traffic traversing via the peer-link for these provider VLANs will not carry all inner C-Tags.
- The **system dot1q-tunnel transit** [**vlan** *vlan-range*] command is required when running this feature on vPC VTEPs.

Configuring QinQ-QinVNI



Note You can also carry native VLAN (untagged traffic) on the same multi-tag trunk port.

The native VLAN on a multi-tag port cannot be configured as a provider VLAN on another multi-tag port or a dot1q enabled port on the same switch.

The **allow-multi-tag** command is allowed only on a trunk port. It is not available on access or dot1q ports.

The **allow-multi-tag** command is not allowed on Peer Link ports. Port channel with multi-tag enabled must not be configured as a vPC peer-link.

SUMMARY STEPS

1. **configure terminal**
2. **interface ethernet** *slot/port*
3. **switchport**
4. **switchport mode trunk**
5. **switchport trunk native vlan** *vlan-id*
6. **switchport trunk allowed vlan** *vlan-list*
7. **switchport trunk allow-multi-tag**

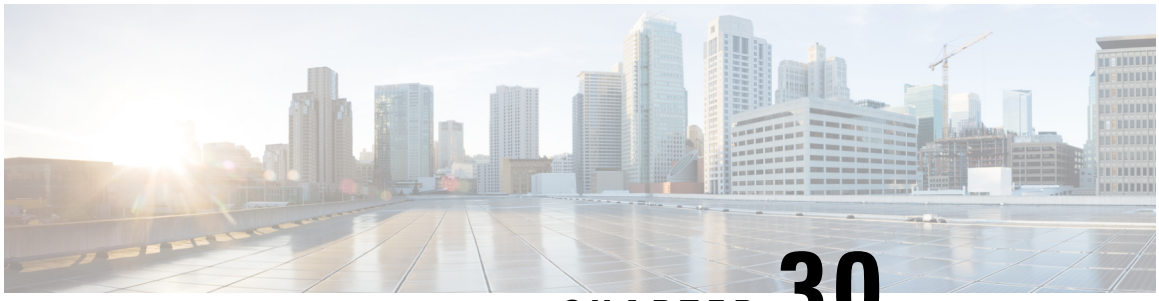
DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**interface ethernet** *slot/port* **Example:** `switch(config)# interface ethernet1/7`	Specifies the interface that you are configuring.
Step 3	**switchport** **Example:** `switch(config-inf)# switchport`	Configures it as a Layer 2 port.
Step 4	**switchport mode trunk** **Example:** `switch(config-inf)# switchport mode trunk`	Sets the interface as a Layer 2 trunk port.
Step 5	**switchport trunk native vlan** *vlan-id* **Example:** `switch(config-inf)# switchport trunk native vlan 30`	Sets the native VLAN for the 802.1Q trunk. Valid values are from 1 to 4094. The default value is VLAN1.
Step 6	**switchport trunk allowed vlan** *vlan-list* **Example:** `switch(config-inf)# switchport trunk allowed vlan 10,20,30`	Sets the allowed VLANs for the trunk interface. The default is to allow all VLANs on the trunk interface: 1 to 3967 and 4048 to 4094. VLANs 3968 to 4047 are the default VLANs reserved for internal use by default.
Step 7	**switchport trunk allow-multi-tag** **Example:** `switch(config-inf)# switchport trunk allow-multi-tag`	Sets the allowed VLANs as the provider VLANs excluding the native VLAN. In the following example, VLANs 10 and 20 are provider VLANs and can carry multiple Inner Q-tags. Native VLAN 30 will not carry inner Q-tags.

Example

```
interface Ethernet1/7
switchport
switchport mode trunk
switchport trunk native vlan 30
switchport trunk allow-multi-tag
switchport trunk allowed vlan 10,20,30
no shutdown
```



CHAPTER 30

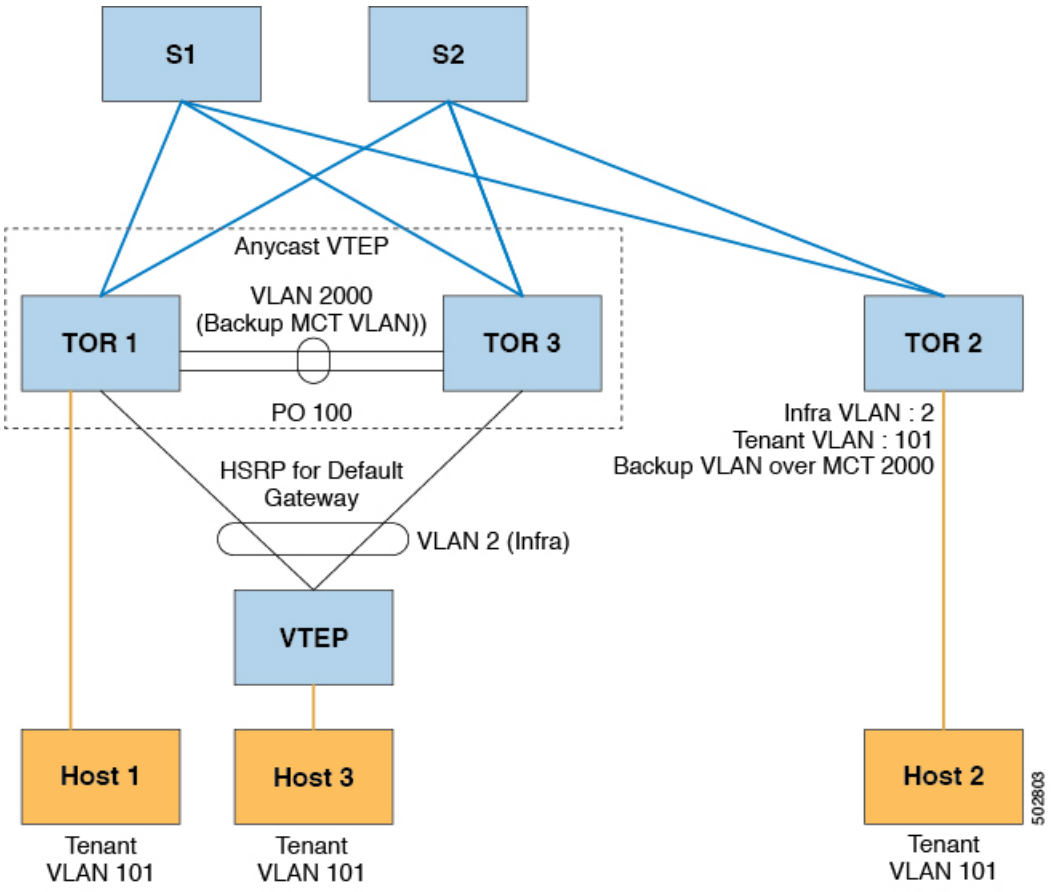
Configuring Bud Node

This chapter contains these sections:

- VXLAN Bud Node Over vPC Overview, on page 703
- VXLAN Bud Node Over vPC Topology Example, on page 704

VXLAN Bud Node Over vPC Overview

Figure 73: Underlay Network Based on PIM-SM and OSPF



Note For bud-node topologies, the source IP of the VTEP behind vPC must be in the same subnet as the infra VLAN. This SVI should have proxy ARP enabled. For example:

```
Interface Vlan2
ip proxy-arp
```

Note The **system nve infra-vlans** command specifies VLANs used for all SVI interfaces, for uplink interfaces with respect to bud-node topologies, and vPC peer-links in VXLAN as infra-VLANs. You must not configure certain combinations of infra-VLANs. For example, 2 and 514, 10 and 522, which are 512 apart.

For Cisco Nexus 9300-FX/FX2/FX3 and 9300-GX platform switches, use the **system nve infra-vlans** command to configure any VLANs that are used as infra-VLANs.

VXLAN Bud Node Over vPC Topology Example

- Enable the required features:

```
feature ospf
feature pim
feature interface-vlan
feature vn-segment-vlan-based
feature hsrp
feature lacp
feature vpc
feature nv overlay
```

- Configuration for PIM anycast RP.

  In this example, 1.1.1.1 is the anycast RP address.

```
ip pim rp-address 1.1.1.1 group-list 225.0.0.0/8
```

- VLAN configuration

  In this example, tenant VLANs 101-103 are mapped to vn-segments.

```
vlan 1-4,101-103,2000
vlan 101
  vn-segment 10001
vlan 102
  vn-segment 10002
vlan 103
  vn-segment 10003
```

- vPC configuration

```
vpc domain 1
  peer-switch
  peer-keepalive destination 172.31.144.213
  delay restore 180
  peer-gateway
  ipv6 nd synchronize
  ip arp synchronize
```

- Infra VLAN SVI configuration

```
interface Vlan2
  no shutdown
  no ip redirects
  ip proxy-arp
  ip address 10.200.1.252/24
  no ipv6 redirects
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  ip igmp static-oif route-map match-mcast-groups
  hsrp version 2
  hsrp 1
    ip 10.200.1.254
```

- Route-maps for matching multicast groups

  Each VXLAN multicast group needs to have a static OIF on the backup SVI MCT.

```
route-map match-mcast-groups permit 1
  match ip multicast group 225.1.1.1/32
```

- Backup SVI over MCT configuration
  - Configuration Option 1:

    ```
    interface Vlan2000
    no shutdown
    ip address 20.20.20.1/24
    ip router ospf 1 area 0.0.0.0
    ip pim sparse-mode
    ```

  - Configuration Option 2:

    ```
    interface Vlan2000
    no shutdown
    ip address 20.20.20.1/24
    ip router ospf 1 area 0.0.0.0
    ip pim sparse-mode
    ```

- vPC interface configuration that carries the infra VLAN

```
interface port-channel1
  switchport mode trunk
  switchport trunk allowed vlan 2
  vpc 1
```

- MCT configuration

```
interface port-channel100
  switchport mode trunk
  spanning-tree port type network
  vpc peer-link
```



Note You can choose either of the following two command procedures for creating the NVE interfaces. Use the first one for a small number of VNIs. Use the second procedure to configure a large number of VNIs.

NVE configuration

Option 1

```
interface nve1
  no shutdown
  source-interface loopback0
  member vni 10001 mcast-group 225.1.1.1
  member vni 10002 mcast-group 225.1.1.1
  member vni 10003 mcast-group 225.1.1.1
```

Option 2

```
interface nve1
  no shutdown
  source-interface loopback0
  global mcast-group 225.1.1.1
  member vni 10001
  member vni 10002
  member vni 10003
```

- Loopback interface configuration

```
interface loopback0
  ip address 101.101.101.101/32
  ip address 99.99.99.99/32 secondary
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
```

- Show commands

```
tor1# sh nve vni
Codes: CP - Control Plane        DP - Data Plane
       UC - Unconfigured         SA - Suppress ARP

Interface VNI      Multicast-group   State Mode Type [BD/VRF]      Flags
--------- -------- ----------------- ----- ---- ------------------ -----
nve1      10001    225.1.1.1         Up    DP   L2 [101]
nve1      10002    225.1.1.1         Up    DP   L2 [102]
nve1      10003    225.1.1.1         Up    DP   L2 [103]

tor1# sh nve peers
Interface Peer-IP          State LearnType Uptime   Router-Mac
--------- ---------------  ----- --------- -------- -----------------
nve1      10.200.1.1       Up    DP        00:07:23 n/a
nve1      10.200.1.2       Up    DP        00:07:18 n/a
nve1      102.102.102.102  Up    DP        00:07:23 n/a


tor1# sh ip mroute 225.1.1.1
IP Multicast Routing Table for VRF "default"

(*, 225.1.1.1/32), uptime: 00:07:41, ip pim nve static igmp
  Incoming interface: Ethernet2/1, RPF nbr: 10.1.5.2
  Outgoing interface list: (count: 3)
    Vlan2, uptime: 00:07:23, igmp
    Vlan2000, uptime: 00:07:31, static
    nve1, uptime: 00:07:41, nve

(10.200.1.1/32, 225.1.1.1/32), uptime: 00:07:40, ip mrib pim nve
  Incoming interface: Vlan2, RPF nbr: 10.200.1.1
  Outgoing interface list: (count: 3)
    Vlan2, uptime: 00:07:23, mrib, (RPF)
    Vlan2000, uptime: 00:07:31, mrib
    nve1, uptime: 00:07:40, nve

(10.200.1.2/32, 225.1.1.1/32), uptime: 00:07:41, ip mrib pim nve
  Incoming interface: Vlan2, RPF nbr: 10.200.1.2
  Outgoing interface list: (count: 3)
    Vlan2, uptime: 00:07:23, mrib, (RPF)
    Vlan2000, uptime: 00:07:31, mrib
    nve1, uptime: 00:07:41, nve

(99.99.99.99/32, 225.1.1.1/32), uptime: 00:07:41, ip mrib pim nve
  Incoming interface: loopback0, RPF nbr: 99.99.99.99
  Outgoing interface list: (count: 3)
    Vlan2, uptime: 00:07:23, mrib
    Vlan2000, uptime: 00:07:31, mrib
    Ethernet2/5, uptime: 00:07:39, pim
(102.102.102.102/32, 225.1.1.1/32), uptime: 00:07:40, ip mrib pim nve
  Incoming interface: Ethernet2/1, RPF nbr: 10.1.5.2
  Outgoing interface list: (count: 1)
    nve1, uptime: 00:07:40, nve


tor1# sh vpc
Legend:
                - local vPC is down, forwarding via vPC peer-link

vPC domain id                     : 1
Peer status                       : peer adjacency formed ok
vPC keep-alive status             : peer is alive
Configuration consistency status  : success
Per-vlan consistency status       : success
Type-2 consistency status         : success
```

```
vPC role                            : secondary, operational primary
Number of vPCs configured           : 4
Peer Gateway                        : Enabled
Dual-active excluded VLANs          : -
Graceful Consistency Check          : Enabled
Auto-recovery status                : Disabled
Delay-restore status                : Timer is off.(timeout = 180s)
Delay-restore SVI status            : Timer is off.(timeout = 10s)

vPC Peer-link status
---------------------------------------------------------------------
id   Port   Status Active vlans
--   ----   ------ --------------------------------------------------
1    Po100  up     1-4,101-103,2000

vPC status
----------------------------------------------------------------------
id   Port   Status Consistency Reason                     Active vlans
--   ----   ------ ----------- ------                     ------------
1    Po1    up     success     success                    2

2    Po2    up     success     success                    2

tor1# sh vpc consistency-parameters global

    Legend:
        Type 1 : vPC will be suspended in case of mismatch

Name                        Type  Local Value            Peer Value
-------------               ----  ---------------------- -----------------------
Vlan to Vn-segment Map      1     3 Relevant Map(s)      3 Relevant Map(s)
STP Mode                    1     Rapid-PVST             Rapid-PVST
STP Disabled                1     None                   None
STP MST Region Name         1     ""                     ""
STP MST Region Revision     1     0                      0
STP MST Region Instance to  1
 VLAN Mapping
STP Loopguard               1     Disabled               Disabled
STP Bridge Assurance        1     Enabled                Enabled
STP Port Type, Edge         1     Normal, Disabled,      Normal, Disabled,
BPDUFilter, Edge BPDUGuard        Disabled               Disabled
STP MST Simulate PVST       1     Enabled                Enabled
Nve Oper State, Secondary   1     Up, 99.99.99.99, DP    Up, 99.99.99.99, DP
IP, Host Reach Mode
Nve Vni Configuration       1     10001-10003            10001-10003
Interface-vlan admin up     2     2,2000                 2,2000
Interface-vlan routing      2     1-4,2000               1-4,2000
capability
Allowed VLANs               -     1-4,101-103,2000       1-4,101-103,2000
Local suspended VLANs       -     -
```



PART I

Configuring VXLAN Security

- Configuring Secure VXLAN EVPN Multi-Site Using CloudSec, on page 711
- Configuring VXLAN ACL, on page 737
- Configuring PVLANs, on page 749
- Configure First Hop Security , on page 753



CHAPTER 31

Configuring Secure VXLAN EVPN Multi-Site Using CloudSec

This chapter contains the following sections:

- About Secure VXLAN EVPN Multi-Site Using CloudSec, on page 711
- Guidelines and Limitations for Secure VXLAN EVPN Multi-Site Using CloudSec, on page 712
- Configuring Secure VXLAN EVPN Multi-Site Using CloudSec, on page 714
- Verifying the Secure VXLAN EVPN Multi-Site Using CloudSec, on page 723
- Displaying Statistics for Secure VXLAN EVPN Multi-Site Using CloudSec, on page 729
- Configuration Examples for Secure VXLAN EVPN Multi-Site Using CloudSec, on page 730
- Migrating from Multi-Site with VIP to Multi-Site with PIP, on page 732
- Migration of Existing vPC BGW, on page 732
- vPC Border Gateway Support for Cloudsec, on page 733
- Enhanced Convergence for vPC BGW CloudSec Deployments, on page 734
- Migration from PSK CloudSec Configuration to Certificate Based Authentication CloudSec Configuration, on page 735

About Secure VXLAN EVPN Multi-Site Using CloudSec

Secure VXLAN EVPN Multi-Site using CloudSec ensures data security and data integrity for VXLAN-based Multi-Site fabrics. Using the cryptographic machinery of IEEE MACsec for UDP packets, this feature provides a secure tunnel between authorized VXLAN EVPN endpoints.

The CloudSec session is point to point over DCI between border gateways (BGWs) on two different sites. All communication between sites uses Multi-Site PIP instead of VIP. For migration information, see Migrating from Multi-Site with VIP to Multi-Site with PIP, on page 732.

Secure VXLAN EVPN Multi-Site using CloudSec is enabled on a per-peer basis. Peers that do not support CloudSec can operate with peers that do support CloudSec, but the traffic is unencrypted. We recommend allowing unencrypted traffic only during migration from non-CloudSec-enabled sites to CloudSec-enabled sites.

CloudSec key exchange uses BGP while MACsec uses the MACsec Key Agreement (MKA). The CloudSec control plane uses the BGP IPv4 address family to exchange the key information. CloudSec keys are carried as part of Tunnel Encapsulation (tunnel type 18) attribute with BGP IPv4 routes using underlay BGP session.

Key Lifetime and Hitless Key Rollover

A CloudSec keychain can have multiple pre-shared keys (PSKs), each configured with a key ID and an optional lifetime. Pre-shared keys are seed keys used to derive further keys for traffic encryption and integrity validation. A list of pre-shared keys can be configured in a keychain with different lifetimes.

A key lifetime specifies when the key expires. CloudSec rolls over to the next configured pre-shared key in the keychain after the lifetime expires. The time zone of the key can be local or UTC. The default time zone is UTC. In the absence of a lifetime configuration, the default lifetime is unlimited.

To configure the CloudSec keychain, see Configuring a CloudSec Keychain and Keys, on page 717.

When the lifetime of the first key expires, it automatically rolls over to the next key in the list. If the same key is configured on both sides of the link at the same time, the key rollover is hitless. That is, the key rolls over without traffic interruption. The lifetime of the keys must be overlapped in order to achieve hitless key rollover.

Certificate Expiration and Replacement

Certificates are used for exchanging Master Session Keys. When certificates expire, no further MSK rekeys will happen. The current secured sessions will continue to stay up and SAK rekeys will happen as configured. The certificate will have to be deleted from under the trustpoint and a new certificate needs to be imported for further MSK rekeys to occur.

Guidelines and Limitations for Secure VXLAN EVPN Multi-Site Using CloudSec

Secure VXLAN EVPN Multi-Site using CloudSec has the following guidelines and limitations:

- Beginning with Cisco NX-OS Release 10.2(2)F, vPC Border Gateway is supported on Cisco Nexus 9300-FX2, -FX3 switches.
- Secure VXLAN EVPN Multi-Site using CloudSec is supported on Cisco Nexus 9300-FX2 platform switches beginning with Cisco NX-OS Release 9.3(5).
- Secure VXLAN EVPN Multi-Site using CloudSec (VXLAN Tunnel Encryption feature) is supported on Cisco Nexus 9300-FX3 platform switches from Cisco NX-OS Release 10.1(1) onwards.
- L3 interfaces and L3 port channels are supported as DCI links.
- CloudSec traffic that is destined for the switch must enter the switch through the DCI uplinks.
- Secure VXLAN EVPN Multi-Site using CloudSec is supported for sites that are connected through a route server or sites that are connected using full mesh (without a route server). For sites that are connected through a route server, upgrade the server to Cisco NX-OS Release 9.3(5) or a later release and follow the instructions in Enabling CloudSec VXLAN EVPN Tunnel Encryption, on page 714.
- VXLAN Tunnel Encryption feature is not supported on Cisco Nexus 9348GC-FX3, 9348GC-FX3PH, and 9332D-H2R, 93400LD-H1, 9364C-H1 switches.
- ICV is disabled by default in Cisco NX-OS Release 9.3(7). ICV should be disabled on the node when forming cloudsec tunnel sessions with node from the previous release (Cisco NX-OS Release 9.3(6)).

- Beginning with Cisco NX-OS Release 10.3.3, VXLAN Tunnel Encryption feature can be configured using Pre Shared Keys (PSK) or certificates using the Public Key Infrastructure(PKI).
- All of the BGWs on the same site should be configured for Secure VXLAN EVPN Multi-Site using CloudSec.
- Secure VXLAN EVPN Multi-Site using CloudSec on DCI links and MACsec on the internal fabric can coexist. However, they can't be enabled simultaneously on the same port or port group (MAC ID).
- Secure VXLAN EVPN Multi-Site using CloudSec peers must have the same keychain configuration in order to decrypt the secure traffic between them.
- A maximum of 60 peers are supported in the BGP IPv4 update of security key distribution in the Cisco Nexus 9300-FX2 family switches.
- Beginning with Cisco NX-OS Release 10.2(3), BGP IPv4 update of security key distribution is supported on Cisco Nexus 9300-FX3 platform switches.
- In order to keep a session alive when all keys with an active timer expire, configure no more than one key per keychain without a lifetime. As a best practice, we recommend configuring a lifetime for each key.
- CloudSec keys are exchanged between BGWs using Tunnel Encapsulation attribute with BGP IPv4 routes using underlay BGP session.

  If this attribute do not get propagated by intermediate nodes, you have to configure direct BGP IPv4 unicast session between the CloudSec end point nodes i.e., BGWs.
- Direct eBGP peering must be established between BGWs in each site if:
  - BGP is used as the IPv4 unicast routing protocol, but the Tunnel Encryption attribute is not propagated through DCI.
  - A routing protocol other than BGP is used for IPv4 unicast routing in the DCI (e.g., OSPF).
- eBGP peering is to be established over a Loopback interface that is different from the following interface:
  - The tunnel-encryption source-interface
  - The nve source-interface
- eBGP peering must filter the loopback IP used as the source of the adjacency. For example, if Loopback10 is used to establish eBGP peering for CloudSec, the IP of Lo10 should not be advertised over this adjacency.
- Secure VXLAN EVPN Multi-Site using CloudSec doesn't support the following:
  - Directly connected L2 hosts on border gateways
  - IP unnumbered configurations on the DCI interface
  - Multicast underlay
  - OAM pathtrace
  - TRM
  - VIP-only model on border gateways

- Beginning with Cisco NX-OS Release 10.5(3)F, Secure VXLAN EVPN Multi-Site using CloudSec supports downstream VNI feature on Cisco Nexus 9300-FX/FX2/FX3/GX/GX2/H2R/H1 Series switches, 9500 Series switches with X97160YC-EX and 9700-FX/GX/FX3 line cards.
- If CloudSec is enabled, non-disruptive ISSU is not supported.
- Different certificate types (SUDI, 3rd party RSA, 3rd party ECC) cannot be mixed in Cloudsec PKI deployments. All nodes should have the same type of certificates
- Nodes with different RSA key sizes are compatible for encryption/decryption.
- PSK and PKI sessions cannot coexist in deployments.
- Size of certificates should not exceed 1.5KB (2048 bit key size).
- MCT-less VPC BGWs is not supported.
- Migration between different certificate types can be done by moving to should-secure, removing trustpoint config from all participating nodes and then, configuring the new trustpoint on all nodes.
- When Cloudsec is initially enabled with the **feature tunnel-encryption** command, the vPC peer-link port-channel and its physical member interfaces will flap.
- Feature **dci-advertise-pip** is not supported on Cisco Nexus 9700-FX/FX3 line cards, and 9500 Series switches with FM-E/FM-E2 fabric modules.

Configuring Secure VXLAN EVPN Multi-Site Using CloudSec

Follow these procedures to configure Secure VXLAN EVPN Multi-Site using CloudSec:

Enabling CloudSec VXLAN EVPN Tunnel Encryption

Follow these steps to enable CloudSec VXLAN EVPN Tunnel Encryption.

Before you begin

Configure BGP peers in the IPv4 unicast address family. Make sure that the IPv4 prefix is propagated with the tunnel community attribute that carries CloudSec keys.

Configure VXLAN EVPN Multi-Site and use the following commands to ensure that peer IP addresses are advertised for CloudSec VXLAN EVPN Tunnel Encryption:

```
evpn multisite border-gateway ms-id
dci-advertise-pip
```



Caution Configuring VXLAN EVPN Multi-Site without **dci-advertise-pip** reverts border gateways to VIP-only mode, which is not supported for CloudSec VXLAN EVPN Tunnel Encryption.

You have two options for sites that are connected through a route server:

- Keep dual RDs enabled – This default behavior ensures that the memory scale remains the same from previous releases in order to handle leaf devices with limited memory. All same-site BGWs use the same RD value for reoriginated routes while advertising EVPN routes to the remote BGW.
- Disable dual RDs – If you don't have memory limitations on leaf devices, you can configure the **no dual rd** command on the BGW. Different RD values are used for reoriginated routes on the same BGWs while advertising EVPN routes to the remote BGW.

Perform one of the following actions, depending on whether dual RDs are enabled on the BGW:

- If dual RDs are configured on the BGWs, follow these steps:

  1. Apply BGP additional paths on the BGW.

```
router bgp as-num
    address-family l2vpn evpn
      maximum-paths number
    additional-paths send
    additional-paths receive
```

  2. Configure multipath for each L3VNI VRF on the BGW.

```
vrf evpn-tenant-00001
  address-family ipv4 unicast
    maximum-paths 64
  address-family ipv6 unicast
    maximum-paths 64
```

  3. Apply BGP additional paths on the route server.

```
router bgp as-num
  address-family l2vpn evpn
    retain route-target all
    additional-paths send
    additional-paths receive
    additional-paths selection route-map name

route-map name permit 10
  set path-selection all advertise
```

- If **no dual rd** is configured on the BGWs or full mesh is configured, follow these steps:

  1. Configure the address family and maximum paths on the BGW.

```
router bgp as-num
  address-family l2vpn evpn
    maximum-paths number
```

  2. Configure multipath for each L3VNI VRF on the BGW.

```
vrf evpn-tenant-00001
    address-family ipv4 unicast
      maximum-paths 64
    address-family ipv6 unicast
      maximum-paths 64
```

Note BGP additional paths are not required on the route server.

SUMMARY STEPS

1. **configure terminal**
2. [**no**] **feature tunnel-encryption**
3. [**no**] **tunnel-encryption source-interface loopback** *number*
4. **tunnel-encryption icv**
5. (Optional) **copy running-config startup-config**

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# configure terminal`
	`switch(config)#`
		Enters global configuration mode.

Step 2	[**no**] **feature tunnel-encryption** **Example:** `switch(config)# feature tunnel-encryption`	Enables CloudSec VXLAN EVPN Tunnel Encryption.
Step 3	[**no**] **tunnel-encryption source-interface loopback** *number* **Example:** `switch(config)# tunnel-encryption source-interface loopback 2`	Specifies the BGP loopback as the tunnel-encryption source interface. The IP address of the configured source interface is used as the prefix to announce CloudSec VXLAN EVPN Tunnel Encryption key routes. **Note** Enter the BGP loopback interface and not the NVE source interface. **Note** Any changes in the MTU should be done before the tunnel-encryption configuration on the interface. This will avoid the CRC drop errors.
Step 4	**tunnel-encryption icv** **Example:** `switch(config)# tunnel-encryption icv`	Enables the Integrity Check Value (ICV). ICV provides integrity check for the frame arriving on the port. If the generated ICV is the same as the ICV in the frame, then the frame is accepted; otherwise it is dropped. This is supported from Cisco NX-OS Release 9.3(7) onwards.

	Command or Action	Purpose
Step 5	(Optional) **copy running-config startup-config** **Example:** `switch(config)# copy running-config startup-config`	Copies the running configuration to the startup configuration.

What to do next

After enabling CloudSec VXLAN EVPN tunnel encryption, you can follow any of the following procedure for authentication.

Configuring a CloudSec Keychain and Keys, on page 717

or

Configuring CloudSec Certificate Based Authentication Using PKI, on page 718

Configuring a CloudSec Keychain and Keys

You can create a CloudSec keychain and keys on the device.

Before you begin

Make sure that Secure VXLAN EVPN Multi-Site using CloudSec is enabled.

SUMMARY STEPS

1. **configure terminal**
2. [**no**] **key chain** *name* **tunnel-encryption**
3. [**no**] **key** *key-id*
4. [**no**] **key-octet-string** *octet-string* **cryptographic-algorithm** {**AES_128_CMAC** | **AES_256_CMAC**}
5. [**no**] **send-lifetime** *start-time* **duration** *duration*
6. (Optional) **show key chain** *name*
7. (Optional) **copy running-config startup-config**

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# configure terminal`
	`switch(config)#`
		Enters global configuration mode.

Step 2	[**no**] **key chain** *name* **tunnel-encryption**
	Example:
	`switch(config)# key chain kc1 tunnel-encryption`
	`switch(config-tunnelencryptkeychain)#`
		Creates a CloudSec keychain to hold a set of CloudSec keys and enters tunnel-encryption keychain configuration mode.

	Command or Action	Purpose

Step 3	[**no**] **key** *key-id*
	Example:
	`switch(config-tunnelencryptkeychain)# key 2000`
	`switch(config-tunnelencryptkeychain-tunnelencryptkey)#`
		Creates a CloudSec key and enters tunnel-encryption key configuration mode. The range is from 1 to 32 octets, and the maximum size is 64.
		Note
		The key must consist of an even number of characters.

Step 4	[**no**] **key-octet-string** *octet-string* **cryptographic-algorithm** {**AES_128_CMAC** | **AES_256_CMAC**} **Example:** `switch(config-tunnelencryptkeychain-tunnelencryptkey)# key-octet-string abcdef0123456789abcdef0123456789 abcdef0123456789abcdef0123456789 cryptographic-algorithm AES_256_CMAC`	Configures the octet string for the key. The *octet-string* argument can contain up to 64 hexadecimal characters. The octet key is encoded internally, so the key in clear text does not appear in the output of the **show running-config tunnel-encryption** command.
Step 5	[**no**] **send-lifetime** *start-time* **duration** *duration* **Example:** `switch(config-tunnelencryptkeychain-tunnelencryptkey)# send-lifetime 00:00:00 May 06 2020 duration 100000`	Configures a send lifetime for the key. By default, the device treats the start time as UTC. The *start-time* argument is the time of day and date that the key becomes active. The *duration* argument is the length of the lifetime in seconds. The range is from 1800 seconds to 2147483646 seconds (approximately 68 years).
Step 6	(Optional) **show key chain** *name* **Example:** `switch(config-tunnelencryptkeychain-tunnelencryptkey)# show key chain kc1`	Displays the keychain configuration.
Step 7	(Optional) **copy running-config startup-config** **Example:** `switch(config-tunnelencryptkeychain-tunnelencryptkey)# copy running-config startup-config`	Copies the running configuration to the startup configuration.

What to do next

Configure a CloudSec VXLAN EVPN Tunnel Encryption policy.

Configuring CloudSec Certificate Based Authentication Using PKI

This chapter contains the following sections:

Attaching a Certificate to CloudSec

You may associate the Cisco NX-OS device with a trust point CA. Cisco NX-OS supports RSA algorithm and ECC (224 and 521 bit) algorithm certificates. Follow the below steps to associate trustpoint or Secure Unique Device Identifier (SUDI) to cloudsec. User need to execute any one of the following commands.

Before you begin

See Configuring PKI to know how to configure a trustpoint and install or import a valid certificate.

SUMMARY STEPS

1. **tunnel-encryption pki trustpoint** *name*
2. **tunnel-encryption pki sudi** *name*

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**tunnel-encryption pki trustpoint** *name*
	Example:
	`switch# tunnel-encryption pki trustpoint myCA_2K`
	`switch(config)#`
		Associate trustpoint to cloud security. Or execute the command in Step 2. Dynamic change of trustpoint labels cannot be done because it will disrupt data traffic.

Step 2	**tunnel-encryption pki sudi** *name*
	Example:
	`switch(config)# tunnel-encryption pki sudi`
	`switch(config-trustpoint)#`
		Associate SUDI to cloud security.
		Note
		Cisco devices have a unique identifier known as the Secure Unique Device Identifier (SUDI) Certificate. This hardware Certificate may be leveraged in lieu of Step 1.

Separate Loopback

Execute any one of the following steps to configure PKI loopback.

SUMMARY STEPS

1. **tunnel-encryption pki source-interface** *loopback*
2. **tunnel-encryption pki source-interface cloudsec-loopback**

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**tunnel-encryption pki source-interface** *loopback*
	Example:
	`switch# tunnel-encryption pki source-interface loopback0`
	`switch(config)#`
		Configures a separate loopback. Or execute the command in Step 2.

	Command or Action	Purpose
Step 2	**tunnel-encryption pki source-interface cloudsec-loopback** **Example:** `switch(config)# tunnel-encryption pki source-interface cloudsec-loopback`	Uses the same loopback as cloudsec source interface loopback.

Configuring a CloudSec Policy

You can create multiple CloudSec policies with different parameters. However, only one policy can be active on an interface.

Before you begin

Make sure that Secure VXLAN EVPN Multi-Site using CloudSec is enabled.

SUMMARY STEPS

1. **configure terminal**
2. (Optional) [**no**] **tunnel-encryption must-secure-policy**
3. [**no**] **tunnel-encryption policy** *name*
4. (Optional) [**no**] **cipher-suite** *name*
5. (Optional) [**no**] **window-size** *number*
6. (Optional) [**no**] **sak-rekey-time** *time*
7. (Optional) **show tunnel-encryption policy**
8. (Optional) **copy running-config startup-config**

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# configure terminal`
	`switch(config)#`
		Enters global configuration mode.

Step 2	(Optional) [**no**] **tunnel-encryption must-secure-policy** **Example:** `switch(config)# tunnel-encryption must-secure-policy`	Ensures that no unencrypted packets are sent over the wire for the session. Packets that are not carrying CloudSec headers are dropped. The **no** form of this command allows unencrypted traffic. We recommend allowing unencrypted traffic only during migration from non-CloudSec-enabled sites to CloudSec-enabled sites. By default, Secure VXLAN EVPN Multi-Site using CloudSec operates in "should secure" mode.

	Command or Action	Purpose

Step 3	[**no**] **tunnel-encryption policy** *name*
	Example:
	`switch(config)# tunnel-encryption policy p1`
	`switch(config-tunenc-policy)#`
		Creates a CloudSec policy.

Step 4	(Optional) [**no**] **cipher-suite** *name* **Example:** `switch(config-tunenc-policy)# cipher-suite GCM-AES-XPN-256`	Configures one of the following ciphers: GCM-AES-XPN-128 or GCM-AES-XPN-256. The default value is GCM-AES-XPN-256.
Step 5	(Optional) [**no**] **window-size** *number* **Example:** `switch(config-tunenc-policy)# window-size 134217728`	Configures the replay protection window such that the interface will not accept any packet that is less than the configured window size. The range is from 134217728 to 1073741823 IP packets. The default value is 268435456.
Step 6	(Optional) [**no**] **sak-rekey-time** *time* **Example:** `switch(config-tunenc-policy)# sak-rekey-time 1800`	Configures the time in seconds to force an SAK rekey. This command can be used to change the session key to a predictable time interval. The range is from 1800 to 2592000 seconds. There is not a default value. We recommend using the same rekey value for all the peers.
Step 7	(Optional) **show tunnel-encryption policy** **Example:** `switch(config-tunenc-policy)# show tunnel-encryption policy`	Displays the CloudSec policy configuration.
Step 8	(Optional) **copy running-config startup-config** **Example:** `switch(config-tunenc-policy)# copy running-config startup-config`	Copies the running configuration to the startup configuration.

What to do next

Configure CloudSec VXLAN EVPN Tunnel Encryption peers.

Configuring CloudSec Peers

This chapter contains the following sections.

Configuring CloudSec Peers

You can configure the CloudSec peers.

Before you begin

Enable Secure VXLAN EVPN Multi-Site using CloudSec.

SUMMARY STEPS

1. **configure terminal**
2. [**no**] **tunnel-encryption peer-ip** *peer-ip-address*
3. [**no**] **keychain** *name* **policy** *name*
4. **pki policy** *policy name*

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# configure terminal`
	`switch(config)#`
		Enters global configuration mode.

Step 2	[**no**] **tunnel-encryption peer-ip** *peer-ip-address* **Example:** `switch(config)# tunnel-encryption peer-ip 33.1.33.33`	Specifies the IP address of the NVE source interface on the peer.
Step 3	[**no**] **keychain** *name* **policy** *name* **Example:** `switch(config)# keychain kc1 policy p1`	Attaches a policy to a CloudSec peer. Step 4 is an alternative to this step.
Step 4	**pki policy** *policy name* **Example:** `switch(config)# pki policy p1`	Attaching cloudsec policy to peer with PKI.

What to do next

Configure CloudSec VXLAN EVPN Tunnel Encryption on an uplink interface.

Enabling Secure VXLAN EVPN Multi-Site Using CloudSec on DCI Uplinks

Follow these steps to enable Secure VXLAN EVPN Multi-Site using CloudSec on all DCI uplinks.

Note This configuration cannot be applied on Layer 2 ports.

Note When CloudSec is applied or removed from an operational DCI uplink, the link will flap. The flap may not be instantaneous as the link may remain down for several seconds.

Before you begin

Make sure that Secure VXLAN EVPN Multi-Site using CloudSec is enabled.

SUMMARY STEPS

1. **configure terminal**
2. [**no**] **interface ethernet** *port/slot*
3. [**no**] **tunnel-encryption**

DETAILED STEPS

Procedure

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# configure terminal`
	`switch(config)#`
		Enters global configuration mode.

Step 2	[**no**] **interface ethernet** *port/slot*
	Example:
	`switch(config)# interface ethernet 1/1`
	`switch(config-if)#`
		Enters interface configuration mode.

Step 3	[**no**] **tunnel-encryption** **Example:** `switch(config-if)# tunnel-encryption`	Enables Secure VXLAN EVPN Multi-Site using CloudSec on the specified interface.

Verifying the Secure VXLAN EVPN Multi-Site Using CloudSec

To display Secure VXLAN EVPN Multi-Site using CloudSec configuration information, perform one of the following tasks:

Command	Purpose
show tunnel-encryption info global	Displays configuration information for Secure VXLAN EVPN Multi-Site using CloudSec.
show tunnel-encryption policy [*policy-name*]	Displays the configuration for a specific CloudSec policy or for all CloudSec policies.
show tunnel-encryption session [**peer-ip** *peer-ip-address*] [**detail**]	Displays information about CloudSec sessions, including whether sessions are secure between endpoints.
show running-config tunnel-encryption	Displays the running configuration information for Secure VXLAN EVPN Multi-Site using CloudSec.

Command	Purpose
show bgp ipv4 unicast *ip-address*	Displays the tunnel encryption information for BGP routes.
show bgp l2vpn evpn	Displays the Layer 2 VPN EVPN address family and routing table information.
show ip route *ip-address* **vrf** *vrf*	Displays the VRF routes.
show l2route evpn mac evi *evi*	Displays Layer 2 route information.
show nve interface *interface* **detail**	Displays the NVE interface detail.
show running-config rpm	Displays the key text in the running configuration. **Note** If you enter the **key-chain tunnelencrypt-psk no-show** command prior to running this command, the key text is hidden (with asterisks) in the running configuration. If you enter the **reload ascii** command, the key text is omitted from the running configuration.
show running-config cert-enroll	Shows the trustpoint and keypair configuration.
show crypto ca certificates <*trustpoint_label*>	Shows the certificate contents under a trustpoint.

The following example displays configuration information for Secure VXLAN EVPN Multi-Site using CloudSec:

```
switch# show tunnel-encryption info global
Global Policy Mode: Must-Secure
 SCI list: 0000.0000.0001.0002 0000.0000.0001.0004
 No. of Active Peers        : 1
```

The following example displays all configured CloudSec policies. The output shows the cipher, window size, and SAK retry time for each policy.

```
switch# show tunnel-encryption policy
Tunnel-Encryption Policy      Cipher            Window     SAK Rekey time
----------------------------  ----------------  ---------  --------------
cloudsec                      GCM-AES-XPN-256   134217728  1800
p1                            GCM-AES-XPN-256   1073741823
system-default-tunenc-policy  GCM-AES-XPN-256   268435456
```

The following example displays information about CloudSec sessions. The output shows the peer IP address and policy, the keychain available, and whether the sessions are secure.

```
switch# show tunnel-encryption session
Tunnel-Encryption  Peer Policy  Keychain  RxStatus        TxStatus
-----------------  -----------  --------  -----------     --------------
33.1.33.33         p1           kc1       Secure (AN: 0)  Secure (AN: 2)
33.2.33.33         p1           kc1       Secure (AN: 0)  Secure (AN: 2)
33.3.33.33         p1           kc1       Secure (AN: 0)  Secure (AN: 2)
44.1.44.44         p1           kc1       Secure (AN: 0)  Secure (AN: 0)
```

```
44.2.44.44         p1           kc1       Secure (AN: 0)  Secure (AN: 0)
```

The following example displays information about Cloudsec sessions based on PKI Certificate Trustpoint.

```
switch# sh tunnel-encryption session
Tunnel-Encryption Peer   Policy                                    Keychain
RxStatus         TxStatus
------------------------ ----------------------------------------
---------------------------------------- ----------------- -----------------
20.20.20.2               p1                                        PKI: myCA (RSA)
Secure (AN: 0)    Secure (AN: 0)
32.11.11.4               p1                                        PKI: myCA (RSA)
Secure (AN: 0)    Secure (AN: 0)
```

The following example shows the tunnel encryption information for BGP routes:

```
switch# show bgp ipv4 unicast 199.199.199.199  Source-loopback configured on peer BGW for
 CloudSec
BGP routing table information for VRF default, address family IPv4 Unicast
BGP routing table entry for 199.199.199.199/32, version 109
Paths: (1 available, best #1)
Flags: (0x8008001a) (high32 0x000200) on xmit-list, is in urib, is best urib route, is in
HW
Multipath: eBGP

  Advertised path-id 1
  Path type: external, path is valid, is best path, no labeled nexthop, in rib
  AS-Path: 1000 200 , path sourced external to AS
    89.89.89.89 (metric 0) from 89.89.89.89 (89.89.89.89)
      Origin IGP, MED not set, localpref 100, weight 0
      Tunnel Encapsulation attribute: Length 120

  Path-id 1 advertised to peers:
    2.2.2.2
```

The following example shows if the MAC is attached with the virtual ESI:

```
switch(config)# show bgp l2vpn evpn 0012.0100.000a
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 110.110.110.110:32876
BGP routing table entry for [2]:[0]:[0]:[48]:[0012.0100.000a]:[0]:[0.0.0.0]/216, version
13198
Paths: (1 available, best #1)
Flags: (0x000202) (high32 00000000) on xmit-list, is not in l2rib/evpn, is not in HW
Multipath: eBGP

  Advertised path-id 1
  Path type: external, path is valid, is best path, no labeled nexthop
             Imported to 1 destination(s)
             Imported paths list: l2-10109
  AS-Path: 1000 200 , path sourced external to AS
    10.10.10.10 (metric 0) from 89.89.89.89 (89.89.89.89)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 10109
      Extcommunity: RT:100:10109 ENCAP:8
      ESI: 0300.0000.0000.0200.0309

  Path-id 1 not advertised to any peer

Route Distinguisher: 199.199.199.199:32876
BGP routing table entry for [2]:[0]:[0]:[48]:[0012.0100.000a]:[0]:[0.0.0.0]/216, version
24823
Paths: (1 available, best #1)
Flags: (0x000202) (high32 00000000) on xmit-list, is not in l2rib/evpn, is not in HW
```

```
Multipath: eBGP

  Advertised path-id 1
  Path type: external, path is valid, is best path, no labeled nexthop
             Imported to 1 destination(s)
             Imported paths list: l2-10109
  AS-Path: 1000 200 , path sourced external to AS
    9.9.9.9 (metric 0) from 89.89.89.89 (89.89.89.89)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 10109
      Extcommunity: RT:100:10109 ENCAP:8
      ESI: 0300.0000.0000.0200.0309

  Path-id 1 not advertised to any peer
```

The following example shows the ECMP created for EVPN type-5 routes received from the remote site:

```
switch(config)# show ip route 205.205.205.9 vrf vrf903
IP Route Table for VRF "vrf903"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

205.205.205.9/32, ubest/mbest: 2/0
    *via 9.9.9.9%default, [20/0], 11:06:32, bgp-100, external, tag 1000, segid: 900003
tunnelid: 0x9090909 encap: VXLAN

    *via 10.10.10.10%default, [20/0], 3d05h, bgp-100, external, tag 1000, segid: 900003
tunnelid: 0xa0a0a0a encap: VXLAN
```

The following example shows if ESI-based MAC multipath is configured for MACs received from the remote site:

Beginning with Cisco NX-OS Release 10.5(3)F, the labeled nexthop and an asymmetric VNI flag is added as shown. For symmetric VNI, labels and flags will not be shown as part of nexthops for EAD and PL.

```
switch# show l2route evpn mac evi 109 mac 0012.0100.000a detail

Flags -(Rmac):Router MAC (Stt):Static (L):Local (R):Remote (V):vPC link
(Dup):Duplicate (Spl):Split (Rcv):Recv (AD):Auto-Delete (D):Del Pending
(S):Stale (C):Clear, (Ps):Peer Sync (O):Re-Originated (Nho):NH-Override
(Pf):Permanently-Frozen, (Orp): Orphan

Topology    Mac Address    Prod   Flags         Seq No     Next-Hops

----------- -------------- ------ ------------- ----------
--------------------------------------
109         0012.0100.000a BGP    Rcv              0         5.5.5.5 (Label: 20000)(Flags:
 Asy)
                                                             6.6.6.6 (Label: 10000)(Flags:
 Asy)
            Route Resolution Type: ESI
            Forwarding State: Resolved (PL)
            Resultant PL: 5.5.5.5 (Label: 20000)(Flags: Asy)
                          6.6.6.6 (Label: 10000)(Flags: Asy)
            Sent To: L2FM
          ESI : 0300.0000.0000.0200.0309
          Encap: 1
```

The following example shows that VXLAN EVPN Multi-Site with PIP is configured:

```
switch(config)# show nve interface nve1 detail
Interface: nve1, State: Up, encapsulation: VXLAN
 VPC Capability: VPC-VIP-Only [not-notified]
 Local Router MAC: 700f.6a15.c791
 Host Learning Mode: Control-Plane
 Source-Interface: loopback0 (primary: 14.14.14.14, secondary: 0.0.0.0)
 Source Interface State: Up
 Virtual RMAC Advertisement: No
 NVE Flags:
 Interface Handle: 0x49000001
 Source Interface hold-down-time: 180
 Source Interface hold-up-time: 30
 Remaining hold-down time: 0 seconds
 Virtual Router MAC: N/A
 Virtual Router MAC Re-origination: 0200.2e2e.2e2e
 Interface state: nve-intf-add-complete
 Multisite delay-restore time: 180 seconds
 Multisite delay-restore time left: 0 seconds
 Multisite dci-advertise-pip configured: True
 Multisite bgw-if: loopback1 (ip: 46.46.46.46, admin: Up, oper: Up)
 Multisite bgw-if oper down reason:
```

The following example shows the key text in the running configuration. If you enter the **key-chain tunnelencrypt-psk no-show** command, the key text is hidden.

```
switch# show running-config rpm
!Command: show running-config rpm
!Running configuration last done at: Mon Jun 15 14:41:40 2020
!Time: Mon Jun 15 15:10:27 2020

version 9.3(5) Bios:version 05.40
key chain inter tunnel-encryption
  key 3301
    key-octet-string 7 075f79696a58405441412e2a577f0f077d6461003652302552040a0b76015a504e370c
7972700604755f0e22230c03254323277d2f5359741a6b5d3a5744315f2f cryptographic-algorithm
AES_256_CMAC
key chain kc1 tunnel-encryption
  key 3537
    key-octet-string 7
072c746f172c3d274e33592e22727e7409106d00372532575803780077755621 3d4e0c7c00770576772
d08515e0804553124577f5a522e046d6a5f485c35425f59 cryptographic-algorithm AES_256_CMAC
    send-lifetime local 09:09:40 Apr 15 2020 duration 1800
  key 2001
    key-octet-string 7
075f79696a58405441412e2a577f0f077d6461003652302552040a0b76015a504e370c7972700604755
f0e22230c03254323277d2f5359741a6b5d3a5744315f2f cryptographic-algorithm AES_256_CMAC
  key 2065
    key-octet-string 7
0729791f6f5e3d213347292d517308730c156c7737223554270f787c07722a513e450a0a0703070c062
e0256210d0e204120510d29222a051f1e594c2135375359 cryptographic-algorithm AES_256_CMAC
  key 2129
    key-octet-string 7
075c796f6f2a4c2642302f5c56790e767063657a4b564f2156777c0a020228564a32780e0472007005530
c5e560f04204056577f2a222d056d1f5c4c533241525d cryptographic-algorithm AES_256_CMAC
  key 2193
    key-octet-string 7
07577014195b402336345a5f260f797d7d6264044b50415755047a7976755a574d350b7e720a0202715d7
a50530d715346205d0c2d525c001f6b5b385046365a29 cryptographic-algorithm AES_256_CMAC

switch# configure terminal
switch(config)# key-chain tunnelencrypt-psk no-show
switch(config)# show running-config rpm
```

```
!Command: show running-config rpm
!Running configuration last done at: Mon Jun 15 15:10:44 2020
!Time: Mon Jun 15 15:10:47 2020

version 9.3(5) Bios:version 05.40
key-chain tunnelencrypt-psk no-show
key chain inter tunnel-encryption
  key 3301
    key-octet-string 7 ****** cryptographic-algorithm AES_256_CMAC
key chain kc1 tunnel-encryption
  key 3537
    key-octet-string 7 ****** cryptographic-algorithm AES_256_CMAC
    send-lifetime local 09:09:40 Apr 15 2020 duration 1800
  key 2001
    key-octet-string 7 ****** cryptographic-algorithm AES_256_CMAC
  key 2065
    key-octet-string 7 ****** cryptographic-algorithm AES_256_CMAC
  key 2129
    key-octet-string 7 ****** cryptographic-algorithm AES_256_CMAC
  key 2193
    key-octet-string 7 ****** cryptographic-algorithm AES_256_CMAC
```

The following example shows the trustpoint and keypair configuration.

```
switch# show running-config cert-enroll
!Command: show running-config cert-enroll
!Running configuration last done at: Fri Apr 21 10:53:30 2023
!Time: Fri Apr 21 12:07:31 2023

version 10.3(3) Bios:version 05.47
crypto key generate rsa label myRSA exportable modulus 1024
crypto key generate rsa label myKey exportable modulus 1024
crypto key generate rsa label tmpCA exportable modulus 2048
crypto key generate ecc label src15_ECC_key exportable modulus 224
crypto ca trustpoint src15_ECC_CA
     ecckeypair switch_ECC_key and so on
     revocation-check  crl
crypto ca trustpoint myRSA
     rsakeypair myRSA
     revocation-check  crl
crypto ca trustpoint tmpCA
     rsakeypair tmpCA
     revocation-check  crl
crypto ca trustpoint myCA
     rsakeypair myKey
     revocation-check  crl
```

The following example shows the certificate contents under a trustpoint.

```
switch(config)# show crypto ca certificates myCA
Trustpoint: myCA
certificate:
subject=CN = switch, serialNumber = FBO22411ABC
issuer=C = US, ST = CA, L = San Jose, O = Org, OU = EN, CN = PKI, emailAddress = abc@xyz.com
serial=2F24FCE6823FCBE5A8AC72C82D0E8E24EB327B0C
notBefore=Apr 19 19:43:48 2023 GMT
notAfter=Aug 31 19:43:48 2024 GMT
SHA1 Fingerprint=D0:F8:1E:32:6E:6D:44:21:6B:AE:92:69:69:AD:88:73:69:76:B9:18
purposes: sslserver sslclient

CA certificate 0:
subject=C = US, ST = CA, L = San Jose, O = Org, OU = EN, CN = PKI, emailAddress = abc@xyz.com
issuer=C = US, ST = CA, L = San Jose, O = Cisco, OU = EN, CN = PKI, emailAddress = ca@ca.com
serial=1142A22DDDE63A047DE08294133593620 42CCC31
```

```
notBefore=Jul 12 13:25:59 2022 GMT
notAfter=Jul 12 13:25:59 2023 GMT
SHA1 Fingerprint=33:37:C6:D5:F1:B3:E1:79:D9:5A:71:30:FD:50:E4:28:7D:E1:2D:A3
purposes: sslserver sslclient
```

Displaying Statistics for Secure VXLAN EVPN Multi-Site Using CloudSec

You can display or clear Secure VXLAN EVPN Multi-Site using CloudSec statistics using the following commands:

Command	Purpose
show tunnel-encryption statistics [**peer-ip** *peer-ip-address*]	Displays statistics for Secure VXLAN EVPN Multi-Site using CloudSec.
clear tunnel-encryption statistics [**peer-ip** *peer-ip-address*]	Clears statistics for Secure VXLAN EVPN Multi-Site using CloudSec.

The following example shows sample statistics for Secure VXLAN EVPN Multi-Site using CloudSec:

```
switch# show tunnel-encryption statistics
Peer 16.16.16.16 SecY Statistics:

SAK Rx Statistics for AN [0]:
Unchecked Pkts: 0
Delayed Pkts: 0
Late Pkts: 0
OK Pkts: 8170598
Invalid Pkts: 0
Not Valid Pkts: 0
Not-Using-SA Pkts: 0
Unused-SA Pkts: 0
Decrypted In-Pkts: 8170598
Decrypted In-Octets: 4137958460 bytes
Validated In-Octets: 0 bytes

SAK Rx Statistics for AN [3]:
Unchecked Pkts: 0
Delayed Pkts: 0
Late Pkts: 0
OK Pkts: 0
Invalid Pkts: 0
Not Valid Pkts: 0
Not-Using-SA Pkts: 0
Unused-SA Pkts: 0
Decrypted In-Pkts: 0
Decrypted In-Octets: 0 bytes
Validated In-Octets: 0 bytes

SAK Tx Statistics for AN [0]:
Encrypted Protected Pkts: 30868929
Too Long Pkts: 0
Untagged Pkts: 0
Encrypted Protected Out-Octets: 15758962530 bytes
```

Note In tunnel encryption statistics, if you observe a traffic drop coinciding with an increase in late packets, it could be due to any of the following reasons:

- The packets are being discarded because they are received outside the replay window.
- The tunnel encryption peers are out of sync.
- There is a valid security risk.

In these situations, you should reset the peer session by removing and then reconfiguring the tunnel-encryption peer on the corresponding remote peer, in order to synchronize them again.

Configuration Examples for Secure VXLAN EVPN Multi-Site Using CloudSec

The following example shows how to configure Secure VXLAN EVPN Multi-Site using keychain:

```
key chain kc1 tunnel-encryption
key 2006
key-octet-string 7 075f79696a58405441412e2a577f0f077d6461003652302552040
a0b76015a504e370c7972700604755f0e22230c03254323277d2f5359741a6b5d3a5744315f2f
cryptographic-algorithm AES_256_CMAC

feature tunnel-encryption
tunnel-encryption source-interface loopback4
tunnel-encryption must-secure-policy

tunnel-encryption policy p1
  window-size 1073741823

tunnel-encryption peer-ip 11.1.11.11
  keychain kc1 policy p1
tunnel-encryption peer-ip 11.2.11.11
  keychain kc1 policy p1
tunnel-encryption peer-ip 44.1.44.44
  keychain kc1 policy p1
tunnel-encryption peer-ip 44.2.44.44
  keychain kc1 policy p1

interface Ethernet1/1
  tunnel-encryption

interface Ethernet1/7
  tunnel-encryption

interface Ethernet1/55
  tunnel-encryption

interface Ethernet1/59
  tunnel-encryption

evpn multisite border-gateway 111
dci-advertise-pip

router bgp 1000
```

```
router-id 12.12.12.12
no rd dual
address-family ipv4 unicast
  maximum-paths 10
address-family l2vpn evpn
  maximum-paths 10
vrf vxlan-900101
address-family ipv4 unicast
  maximum-paths 10
address-family ipv6 unicast
  maximum-paths 10

show tunnel-encryption session
Tunnel-Encryption Peer  Policy  Keychain  RxStatus        TxStatus
----------------------  ------  --------  --------------  --------------
11.1.11.11               p1     kc1       Secure (AN: 0)  Secure (AN: 2)
11.2.11.11               p1     kc1       Secure (AN: 0)  Secure (AN: 2)
44.1.44.44               p1     kc1       Secure (AN: 0)  Secure (AN: 2)
44.2.44.44               p1     kc1       Secure (AN: 0)  Secure (AN: 2)
```

The following example shows how to configure Certificate based Secure VXLAN EVPN Multi-site using Clousec Certificate based Authentication.

```
feature tunnel-encryption

tunnel-encryption must-secure-policy
tunnel-encryption pki trustpoint myCA
tunnel-encryption pki source-interface loopback3
tunnel-encryption source-interface loopback2
tunnel-encryption policy with-rekey
  sak-rekey-time 1800
tunnel-encryption peer-ip 7.7.7.7
  pki policy system-default-tunenc-policy


interface Ethernet1/20
  tunnel-encryption

interface Ethernet1/21
  tunnel-encryption

interface Ethernet1/25/1
  tunnel-encryption
```

The following example shows how to configure outbound route-map to make BGW's path as the best path. This configuration is done when vPC BGW learns peer vPC BGW's PIP address in BGP.

```
ip prefix-list pip_ip seq 5 permit 44.44.44.44/32 <<PIP2 address>>
route-map pip_ip permit 5
    match ip address prefix-list pip_ip
    set as-path prepend last-as 1
neighbor 45.10.45.10 <<R1 neighbor - Same route-map required for every DCI side underlay
BGP peer>>
    inherit peer EBGP-PEERS
    remote-as 12000
    address-family ipv4 unicast
      route-map pip_ip out
```

Migrating from Multi-Site with VIP to Multi-Site with PIP

Follow these steps for a smooth migration from Multi-Site with VIP to Multi-Site with PIP. The migration needs to be done one site at a time. You can expect minimal traffic loss during the migration.

1. Upgrade all BGWs on all sites to Cisco NX-OS Release 9.3(5) or a later release.
2. Configure BGP maximum paths on all BGWs. Doing so is required for ESI-based MAC multipath and BGP to download all of the next-hops for EVPN Type-2 and Type-5 routes.
3. Pick one site at a time for the migration.
4. Shut down the same-site BGWs except for one BGW. You can use the NVE **shutdown** command to shut down the BGWs.
5. To avoid traffic loss, wait a few minutes before enabling Multi-Site with PIP on the active BGW. Doing so allows the same-site shutdown BGWs to withdraw EVPN routes so remote BGWs send traffic to only the active BGW.
6. Enable Multi-Site with PIP on the active BGW by configuring the **dci-advertise-pip** command.

   The Multi-Site with PIP-enabled BGW advertises the EVPN EAD-per-ES route for the virtual ESI.

   The Multi-Site with PIP-enabled BGW advertises EVPN Type-2 routes with a virtual ESI, using the PIP address as the next-hop and the PIP interface MAC as the RMAC (if applicable) towards the DCI. For Type-5 routes, no virtual ESI is carried. There is no change in the advertisement of EVPN Type-2 and Type-5 routes towards the fabric.

   The remote BGW performs ESI-based MAC multipathing as MAC routes are received with ESI.
7. Unshut the same-site BGWs one at a time and enable Multi-Site with PIP by entering the **dci-advertise-pip** command.

   The remote BGW performs ESI-based MAC multipathing for MAC routes as ESI is the same from all same-site BGWs.

   On the remote BGW, BGP selects paths as multipath and downloads all next-hops for EVPN Type-5 routes.

Migration of Existing vPC BGW

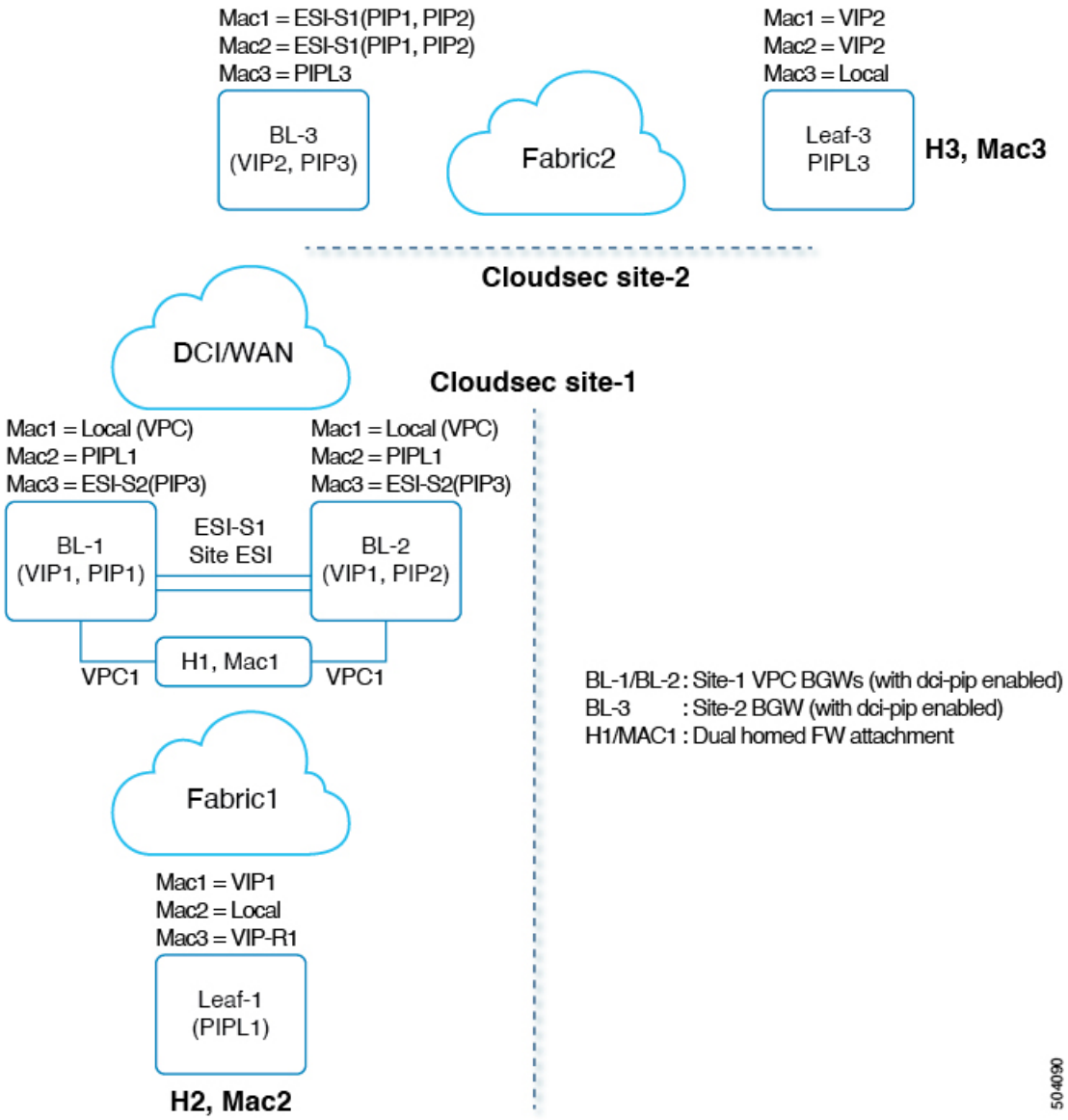
Follow these steps for a smooth migration of the existing vPC BGWs so that they can use Cloudsec. The migration needs to be done one site at a time. You can expect minimal traffic loss during the migration.

1. Upgrade both vPC BGWs to the latest image which has the vPC Cloudsec updates.
2. Shutdown interface nve1 on the vPC secondary.
3. Enable **dci-advertise-pip** on vPC primary.
4. With interface nve1 still in shut mode on vPC secondary, configure **dci-advertise-pip** on vPC secondary.
5. Unshut interface nve1 on vPC secondary.

vPC Border Gateway Support for Cloudsec

The following topology illustrates the vPC Border Gateway (BGW) support for Cloudsec.

Figure 74: vPC BGW Support for Cloudsec



vPC is a dual-homed attachment/connection to the BGW. BGWs act virtually as a single VXLAN end point for redundancy and both switches function in active mode by sharing a common emulated/virtual ip-address (VIP). The VXLAN encapsulation over DCI is based on primary IP addresses of the BGW VTEPs.

In the above topology, Host H1/MAC1 is dually homed to Cloudsec enabled vPC BGWs BL-1/BL-2. H1 continues to be advertised with secondary loopback IP address of the vPC BGWs (VIP1) towards the fabric. However, towards the DCI, both BL-1/BL-2 advertise H1 with next-hop as PIP and site-ESI is also added to the Type-2 NLRI.

For Cloudsec feature on Anycast and vPC BGWs, dci-advertise-pip is configured to change the BGP procedures of how the Type-2/Type-5 routes are advertised to the DCI. All Type-2/Type-5 routes received from the site-internal network are advertised to the DCI with next-hop as PIP of the vPC BGW.

Both vPC BGWs advertise the routes with their primary IP address respectively. Site-ESI attribute is added to the Type-2 NLRIs. All dual attached hosts on the vPC BGWs are advertised with next-hop as PIP and site-ESI attribute is attached over DCI. All orphan hosts are advertised with next-hop as PIP towards DCI and the site-ESI attribute is not attached.

If vPC BGW learns peer vPC BGWs PIP address and advertises on DCI side, BGP path attributes from both vPC BGW will be same. Hence the DCI intermediate nodes may end up choosing the path from vPC BGW which does not own the PIP address. In this scenario MCT link is used for encrypted traffic coming from the remote site. The vPC BGW BGP the learns the peer vPC BGW's PIP address when:

- iBGP is configured between vPC BGWs.
- BGP is used as underlay routing protocol on fabric side.
- IGP used as underlay routing protocol, and IGP routes are redistributed into BGP.

When vPC BGW learns peer vPC BGW's PIP address in BGP, you need to configure the outbound route-map to make BGW's path as the best path.

On a remote site BGW, directly connected L3 host is learnt from both vPC BGWs. The path from directly connected BGW is usually preferred due to lower AS-path. If L3 host or L3 network is dually connected to vPC pair BGW, the local path is selected in both vPC pair.

Enhanced Convergence for vPC BGW CloudSec Deployments

Traditionally, single loopback interface is configured as NVE source interface, where both PIP and VIP of vPC complex are configured. Beginning with Cisco NX-OS Release 10.3(2)F, you can configure a separate loop back for CloudSec enabled vPC BGW. It is recommended to use separate loopback interfaces for source and anycast IP addresses under NVE for better convergence in vPC deployments. The IP address configured on the source-interface is the PIP of the vPC node, and the IP address configured on the anycast interface is the VIP of that vPC complex. Note that the secondary IP configured on the NVE source-interface will have no effect if the NVE anycast interface is also configured.

With separate loopbacks, the convergence for dual-attached EVPN Type-2 and Type-5 routes traffic destined for DCI side will be improved.

Migration to Anycast Interface

If a user wants to specify an anycast interface, the user needs to unconfigure the existing source-interface and reconfigure with both source and anycast interfaces. This will lead to temporary traffic loss. For all green field deployments, it is recommended to configure both the source and anycast interface to avoid the convergence problem specified.

NVE Interface Configuration with Enhanced Convergence for vPC BGW CloudSec Deployments

The user needs to specify anycast interface along with NVE source-interface on vPC BGW. In today's VxLANv6 deployments, the provision to specify both source-interface and anycast interface is already present. In order to improve vPC convergence for VxLANv4, the anycast option is mandatory.

Configuration Example:

```
interface nve <number>
        source-interface <interface> [anycast <anycast-intf>]
```

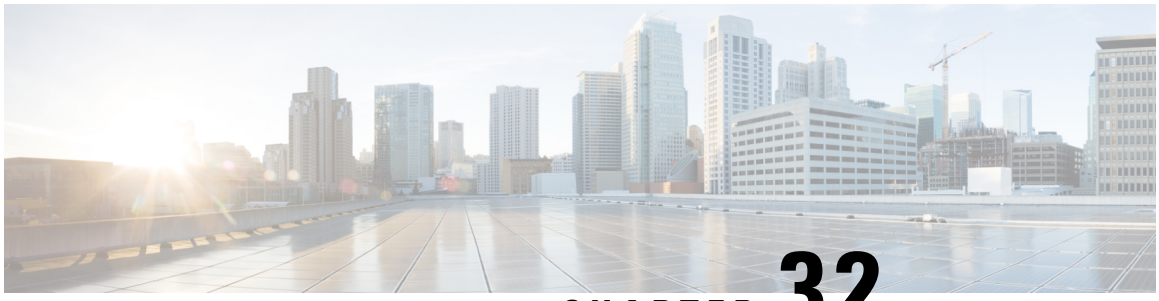
iBGP Session Requirement

Underlay IPv4/IPv6 unicast iBGP session must be configured between vPC BGW peer nodes. This is to accommodate key propagation during the DCI isolation on any vPC BGW.

Migration from PSK CloudSec Configuration to Certificate Based Authentication CloudSec Configuration

During migration to Auto keying, it is expected to send or receive clear traffic on a VTEP-to-VTEP session while the sites are still migrating to new configuration or functionalists. During this time, policy should be configured as **should-secure** to make sure unencrypted traffic is not dropped for the session.

1. Change tunnel-encryption config to **should-secure** on all nodes.
2. Perform migration one node at a time.
3. Remove the keychain and cloudsec policy from peer.
4. Configure trust point and certificate using a valid CA if using SSL certificates OR configure for SUDI certificates.
5. Attach the trust point to Cloudsec.
6. Apply the cloudsec policy back to the peer.
7. After all the nodes have been changed to autokeying, change the configuration to **must-secure** if needed.



CHAPTER 32

Configuring VXLAN ACL

This chapter contains these sections:

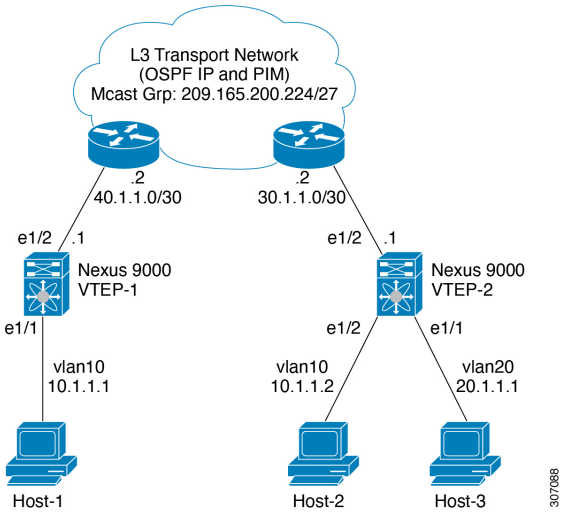
- About Access Control Lists, on page 737
- Guidelines and Limitations for VXLAN ACLs, on page 738
- VXLAN Tunnel Encapsulation Switch, on page 738
- VXLAN Tunnel Decapsulation Switch, on page 743

About Access Control Lists

ACL implementation for VXLAN is the same as regular IP traffic. The host traffic is not encapsulated in the ingress direction at the encapsulation switch. The implementation is a bit different for the VXLAN encapsulated traffic at the decapsulation switch as the ACL classification is based on the inner payload. The supported ACL scenarios for VXLAN are explained in the following topics and the unsupported cases are also covered for both encapsulation and decapsulation switches.

All scenarios that are mentioned in the previous table are explained with the following host details:

Figure 75: Port ACL on VXLAN Encap Switch



- Host-1: 10.1.1.1/24 VLAN-10

- Host-2: 10.1.1.2/24 VLAN-10
- Host-3: 20.1.1.1/24 VLAN-20
- Case 1: Layer 2 traffic/L2 VNI that flows between Host-1 and Host-2 on VLAN-10.
- Case 2: Layer 3 traffic/L3 VNI that flows between Host-1 and Host-3 on VLAN-10 and VLAN-20.

Guidelines and Limitations for VXLAN ACLs

VXLAN ACLs have the following guidelines and limitations:

- A router ACL (RACL) on an SVI of the incoming VLAN-10 and the uplink port (eth1/2) does not support filtering the encapsulated VXLAN traffic with outer or inner headers in an egress direction. The limitation also applies to the Layer 3 port-channel uplink interfaces.
- A router ACL (RACL) on an SVI and the Layer 3 uplink ports is not supported to filter the encapsulated VXLAN traffic with outer or inner headers in an ingress direction. This limitation also applies to the Layer 3 port-channel uplink interfaces.
- A port ACL (PACL) cannot be applied on the Layer 2 port to which a host is connected. Cisco NX-OS does not support a PACL in the egress direction.
- On Cisco Nexus 9324C-SE1U platforms, configuring a Port ACL (PACL) on an interface that belongs to a Service VRF is not supported. The CLI may allow this configuration, but PACLs will not function on ports assigned to a Service VRF. To avoid unexpected behavior, do not apply PACLs to Service VRF interfaces.

VXLAN Tunnel Encapsulation Switch

Port ACL on the Access Port on Ingress

You can apply a port ACL (PACL) on the Layer 2 trunk or access port that a host is connected on the encapsulating switch. As the incoming traffic from access to the network is normal IP traffic. The ACL that is being applied on the Layer 2 port can filter it as it does for any IP traffic in the non-VXLAN environment.

The **ing-racl** TCAM region must be carved as follows:

SUMMARY STEPS

1. **configure terminal**
2. **hardware access-list tcam region ing-racl 256**
3. **ip access-list** *name*
4. *sequence-number* **permit ip** *source-address destination-address*
5. **exit**
6. **interface ethernet** *slot/port*
7. **ip port access-group** *pacl-name* **in**
8. **switchport**
9. **switchport mode trunk**

10. **switchport trunk allowed vlan** *vlan-list*
11. **no shutdown**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**hardware access-list tcam region ing-racl 256** **Example:** `switch(config)# hardware access-list tcam region ing-racl 256`	Attaches the UDFs to the **ing-racl** TCAM region, which applies to IPv4 or IPv6 port ACLs.
Step 3	**ip access-list** *name* **Example:** `switch(config)# ip access list PACL_On_Host_Port`	Creates an IPv4 ACL and enters IP ACL configuration mode. The name arguments can be up to 64 characters.
Step 4	*sequence-number* **permit ip** *source-address destination-address* **Example:** `switch(config-acl)# 10 permit ip 10.1.1.1/32 10.1.1.2/32`	Creates an ACL rule that permits or denies IPv4 traffic matching its condition. The *source-address destination-address* arguments can be the IP address with a network wildcard, the IP address and variable-length subnet mask, the host address, and **any** to designate any address.
Step 5	**exit** **Example:** `switch(config-acl)# exit`	Exits IP ACL configuration mode.
Step 6	**interface ethernet** *slot/port* **Example:** `switch(config)# interface ethernet1/1`	Enters interface configuration mode.
Step 7	**ip port access-group** *pacl-name* **in** **Example:** `switch(config-if)# ip port access-group PACL_On_Host_Port in`	Applies a Layer 2 PACL to the interface. Only inbound filtering is supported with port ACLs. You can apply one port ACL to an interface.
Step 8	**switchport** **Example:** `switch(config-if)# switchport`	Configures the interface as a Layer 2 interface.

	Command or Action	Purpose
Step 9	**switchport mode trunk** **Example:** `switch(config-if)# switchport mode trunk`	Configures the interface as a Layer 2 trunk port.
Step 10	**switchport trunk allowed vlan** *vlan-list* **Example:** `switch(config-if)# switchport trunk allowed vlan 10,20`	Sets the allowed VLANs for the trunk interface. The default is to allow all VLANs on the trunk interface, 1 through 3967 and 4048 through 4094. VLANs 3968 through 4047 are the default VLANs reserved for internal use.
Step 11	**no shutdown** **Example:** `switch(config-if)# no shutdown`	Negates the **shutdown** command.

VLAN ACL on the Server VLAN

A VLAN ACL (VACL) can be applied on the incoming VLAN-10 that the host is connected to on the encap switch. As the incoming traffic from access to network is normal IP traffic, the ACL that is being applied to VLAN-10 can filter it as it does for any IP traffic in the non-VXLAN environment. For more information on VACL, see About Access Control Lists, on page 737.

SUMMARY STEPS

1. **configure terminal**
2. **ip access-list** *name*
3. *sequence-number* **permit ip** *source-address destination-address*
4. **vlan access-map** *map-name* **[***sequence-number***]**
5. **match ip address** *ip-access-list*
6. **action forward**
7. **vlan access-map** *name*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**ip access-list** *name* **Example:** `switch(config)# ip access list Vacl_On_Source_VLAN`	Creates an IPv4 ACL and enters IP ACL configuration mode. The name arguments can be up to 64 characters.

	Command or Action	Purpose
Step 3	*sequence-number* **permit ip** *source-address destination-address* **Example:** `switch(config-acl)# 10 permit ip 10.1.1.1 10.1.1.2`	Creates an ACL rule that permits or denies IPv4 traffic matching its condition. The *source-address destination-address* arguments can be the IP address with a network wildcard, the IP address and variable-length subnet mask, the host address, and **any** to designate any address.
Step 4	**vlan access-map** *map-name* **[***sequence-number***]** **Example:** `switch(config-acl)# vlan access-map Vacl_on_Source_Vlan 10`	Enters VLAN access-map configuration mode for the VLAN access map specified. If the VLAN access map does not exist, the device creates it. If you do no specify a sequence number, the device creates a new entry whose sequence number is 10 greater than the last sequence number in the access map.
Step 5	**match ip address** *ip-access-list* **Example:** `switch(config-acl)# match ip address Vacl_on_Source_Vlan`	Specifies an ACL for the access-map entry.
Step 6	**action forward** **Example:** `switch(config-acl)# action forward`	Specifies the action that the device applies to traffic that matches the ACL.
Step 7	**vlan access-map** *name* **Example:** `switch(config-acl)# vlan access map Vacl_on_Source_Vlan`	Enters VLAN access-map configuration mode for the VLAN access map specified.

Routed ACL on an SVI on Ingress

A router ACL (RACL) in the ingress direction can be applied on an SVI of the incoming VLAN-10 that the host that connects to the encapsulating switch. As the incoming traffic from access to network is normal IP traffic, the ACL that is being applied on SVI 10 can filter it as it does for any IP traffic in the non-VXLAN environment.

The **ing-racl** TCAM region must be carved as follows:

SUMMARY STEPS

1. **configure terminal**
2. **hardware access-list tcam region ing-racl 256**
3. **ip access-list** *name*
4. *sequence-number* **permit ip** *source-address destination-address*
5. **exit**
6. **interface ethernet** *slot/port*
7. **no shutdown**

8. **ip access-group** *racl-name* **in**
9. **vrf member** *vxlan-number*
10. **no ip redirects**
11. **ip address** *ip-address*
12. **no ipv6 redirects**
13. **fabric forwarding mode anycast-gateway**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**hardware access-list tcam region ing-racl 256** **Example:** `switch(config)# hardware access-list tcam region ing-racl 256`	Attaches the UDFs to the **ing-racl** TCAM region, which applies to IPv4 or IPv6 port ACLs.
Step 3	**ip access-list** *name* **Example:** `switch(config)# ip access list PACL_On_Host_Port`	Creates an IPv4 ACL and enters IP ACL configuration mode. The name arguments can be up to 64 characters.
Step 4	*sequence-number* **permit ip** *source-address destination-address* **Example:** `switch(config-acl)# 10 permit ip 10.1.1.1/32 10.1.1.2/32`	Creates an ACL rule that permits or denies IPv4 traffic matching its condition. The *source-address destination-address* arguments can be the IP address with a network wildcard, the IP address and variable-length subnet mask, the host address, and **any** to designate any address.
Step 5	**exit** **Example:** `switch(config-acl)# exit`	Exits IP ACL configuration mode.
Step 6	**interface ethernet** *slot/port* **Example:** `switch(config)# interface ethernet1/1`	Enters interface configuration mode.
Step 7	**no shutdown** **Example:** `switch(config-if)# no shutdown`	Negates **shutdown** command.

	Command or Action	Purpose
Step 8	**ip access-group** *racl-name* **in** **Example:** `switch(config-if)# ip port access-group Racl_On_Source_Vlan_SVI in`	Applies a Layer 2 PACL to the interface. Only inbound filtering is supported with port ACLs. You can apply one port ACL to an interface.
Step 9	**vrf member** *vxlan-number* **Example:** `switch(config-if)# vrf member Cust-A`	Configure SVI for host.
Step 10	**no ip redirects** **Example:** `switch(config-if)# no ip redirects`	Prevents the device from sending redirects.
Step 11	**ip address** *ip-address* **Example:** `switch(config-if)# ip address 10.1.1.10`	Configures an IP address for this interface.
Step 12	**no ipv6 redirects** **Example:** `switch(config-if)# no ipv6 redirects`	Disables the ICMP redirect messages on BFD-enabled interfaces.
Step 13	**fabric forwarding mode anycast-gateway** **Example:** `switch(config-if)# fabric forwarding mode anycast-gateway`	Configure Anycast gateway forwarding mode.

Routed ACL on the Uplink on Egress

A RACL on an SVI of the incoming VLAN-10 and the uplink port (eth1/2) is not supported to filter the encapsulated VXLAN traffic with an outer or inner header in an egress direction. This limitation also applies to the Layer 3 port-channel uplink interfaces.

VXLAN Tunnel Decapsulation Switch

Routed ACL on the Uplink on Ingress

A RACL on a SVI and the Layer 3 uplink ports is not supported to filter the encapsulated VXLAN traffic with outer or inner header in an ingress direction. This limitation also applies to the Layer 3 port-channel uplink interfaces.

Port ACL on the Access Port on Egress

Do not apply a PACL on the Layer 2 port to which a host is connected. Cisco Nexus 9000 Series switches do not support a PACL in the egress direction.

VLAN ACL for the Layer 2 VNI Traffic

A VLAN ACL (VACL) can be applied on VLAN-10 to filter with the inner header when the Layer 2 VNI traffic is flowing from Host-1 to Host-2. For more information on VACL, see About Access Control Lists, on page 737.

The VACL TCAM region must be carved as follows:

SUMMARY STEPS

1. **configure terminal**
2. **hardware access-list tcam region vacl 256**
3. **ip access-list** *name*
4. **statistics per-entry**
5. *sequence-number* **permit ip** *source-address destination-address*
6. *sequence-number* **permit** *protocol source-address destination-address*
7. **exit**
8. **vlan access-map** *map-name* **[***sequence-number***]**
9. **match ip address** *list-name*

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**hardware access-list tcam region vacl 256** **Example:** `switch(config)# hardware access-list tcam region vacl 256`	Changes the ACL TCAM region size.
Step 3	**ip access-list** *name* **Example:** `switch(config)# ip access list VXLAN-L2-VNI`	Creates an IPv4 ACL and enters IP ACL configuration mode. The name arguments can be up to 64 characters.
Step 4	**statistics per-entry** **Example:** `switch(config-acl)# statistics per-entry`	Specifies that the device maintains global statistics for packets that match the rules in the VACL.

	Command or Action	Purpose
Step 5	*sequence-number* **permit ip** *source-address destination-address* **Example:** `switch(config-acl)# 10 permit ip 10.1.1.1/32 10.1.1.2/32`	Creates an ACL rule that permits or denies IPv4 traffic matching its condition. The *source-address destination-address* arguments can be the IP address with a network wildcard, the IP address and variable-length subnet mask, the host address, and **any** to designate any address.
Step 6	*sequence-number* **permit** *protocol source-address destination-address* **Example:** `switch(config-acl)# 20 permit tcp 10.1.1.2/32 10.1.1.1/32`	Creates an ACL rule that permits or denies IPv4 traffic matching its condition. The *source-address destination-address* arguments can be the IP address with a network wildcard, the IP address and variable-length subnet mask, the host address, and **any** to designate any address.
Step 7	**exit** **Example:** `switch(config-acl)# exit`	Exit ACL configuration mode.
Step 8	**vlan access-map** *map-name* **[***sequence-number***]** **Example:** `switch(config)# vlan access-map VXLAN-L2-VNI 10`	Enters VLAN access-map configuration mode for the VLAN access map specified. If the VLAN access map does not exist, the device creates it. If you do no specify a sequence number, the device creates a new entry whose sequence number is 10 greater than the last sequence number in the access map.
Step 9	**match ip address** *list-name* **Example:** `switch(config-access-map)# match ip VXLAN-L2-VNI`	Configure the IP list name.

VLAN ACL for the Layer 3 VNI Traffic

A VLAN ACL (VACL) can be applied on the destination VLAN-20 to filter with the inner header when the Layer 3 VNI traffic is flowing from Host-1 to Host-3. It slightly differs from the previous case as the VACL for the Layer 3 traffic is accounted on the egress on the system. The keyword **output** must be used while dumping the VACL entries for the Layer 3 VNI traffic. For more information on VACL, see About Access Control Lists, on page 737.

The VACL TCAM region must be carved as follows.

SUMMARY STEPS

1. **configure terminal**
2. **hardware access-list tcam region vacl 256**
3. **ip access-list** *name*
4. **statistics per-entry**
5. *sequence-number* **permit ip** *source-address destination-address*

6. *sequence-number* **permit** *protocol source-address destination-address*
7. **vlan access-map** *map-name* **[***sequence-number***]**
8. **action forward**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**hardware access-list tcam region vacl 256** **Example:** `switch(config)# hardware access-list tcam region vacl 256`	Changes the ACL TCAM region size.
Step 3	**ip access-list** *name* **Example:** `switch(config)# ip access list VXLAN-L3-VNI`	Creates an IPv4 ACL and enters IP ACL configuration mode. The name arguments can be up to 64 characters.
Step 4	**statistics per-entry** **Example:** `switch(config)# statistics per-entry`	Specifies that the device maintains global statistics for packets that match the rules in the VACL.
Step 5	*sequence-number* **permit ip** *source-address destination-address* **Example:** `switch(config-acl)# 10 permit ip 10.1.1.1/32 20.1.1.1/32`	Creates an ACL rule that permits or denies IPv4 traffic matching its condition. The *source-address destination-address* arguments can be the IP address with a network wildcard, the IP address and variable-length subnet mask, the host address, and **any** to designate any address.
Step 6	*sequence-number* **permit** *protocol source-address destination-address* **Example:** `switch(config-acl)# 20 permit tcp 20.1.1.1/32 10.1.1.1/32`	Configures the ACL to redirect-specific HTTP methods to a server.
Step 7	**vlan access-map** *map-name* **[***sequence-number***]** **Example:** `switch(config-acl)# vlan access-map VXLAN-L3-VNI 10`	Enters VLAN access-map configuration mode for the VLAN access map specified. If the VLAN access map does not exist, the device creates it. If you do no specify a sequence number, the device creates a new entry whose sequence number is 10 greater than the last sequence number in the access map.

	Command or Action	Purpose
Step 8	**action forward** **Example:** `switch(config-acl)# action forward`	Specifies the action that the device applies to traffic that matches the ACL.

Routed ACL on an SVI on Egress

A router ACL (RACL) on the egress direction can be applied on an SVI of the destination VLAN-20 that Host-3 is connected to on the decap switch to filter with the inner header for traffic flows from the network to access which is normal post-decapsulated IP traffic post. The ACL that is being applied on SVI 20 can filter it as it does for any IP traffic in the non-VXLAN environment. For more information on ACL, see About Access Control Lists, on page 737.

The egr-racl TCAM region must be carved as follows:

SUMMARY STEPS

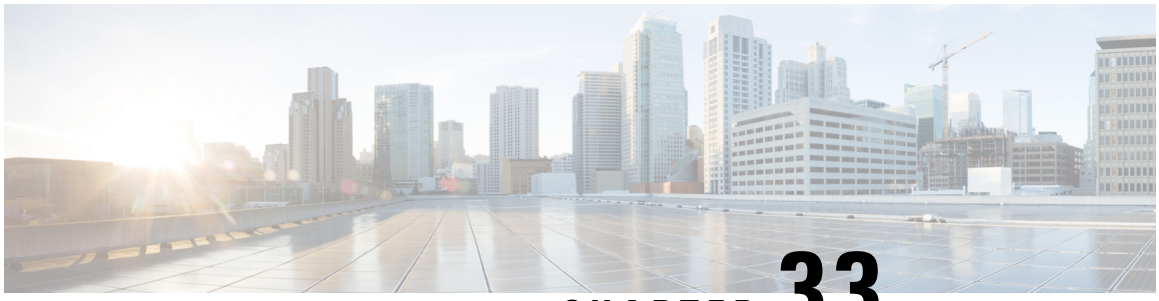
1. **configure terminal**
2. **hardware access-list tcam region egr-racl 256**
3. **ip access-list** *name*
4. *sequence-number* **permit ip** *source-address destination-address*
5. **interface vlan** *vlan-id*
6. **no shutdown**
7. **ip access-group** *access-list* **out**
8. **vrf member** *vxlan-number*
9. **no ip redirects**
10. **ip address** *ip-address/length*
11. **no ipv6 redirects**
12. **fabric forwarding mode anycast-gateway**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**hardware access-list tcam region egr-racl 256** **Example:** `switch(config)# hardware access-list tcam region egr-racl 256`	Changes the ACL TCAM region size.

	Command or Action	Purpose
Step 3	**ip access-list** *name* **Example:** `switch(config)# ip access-list Racl_on_Source_Vlan_SVI`	Creates an IPv4 ACL and enters IP ACL configuration mode. The name arguments can be up to 64 characters.
Step 4	*sequence-number* **permit ip** *source-address destination-address* **Example:** `switch(config-acl)# 10 permit ip 10.1.1.1/32 20.1.1.1/32`	Creates an ACL rule that permits or denies IPv4 traffic matching its condition. The *source-address destination-address* arguments can be the IP address with a network wildcard, the IP address and variable-length subnet mask, the host address, and **any** to designate any address.
Step 5	**interface vlan** *vlan-id* **Example:** `switch(config-acl)# interface vlan vlan20`	Enters interface configuration mode, where *vlan-id* is the ID of the VLAN that you want to configure with a DHCP server IP address.
Step 6	**no shutdown** **Example:** `switch(config-if)# no shutdown`	Negate the shutdown command.
Step 7	**ip access-group** *access-list* **out** **Example:** `switch(config-if)# ip access-group Racl_On_Detination_Vlan_SVI out`	Applies an IPv4 or IPv6 ACL to the Layer 3 interfaces for traffic flowing in the direction specified. You can apply one router ACL per direction.
Step 8	**vrf member** *vxlan-number* **Example:** `switch(config-if)# vrf member Cust-A`	Configure SVI for host.
Step 9	**no ip redirects** **Example:** `switch(config-if)# no ip redirects`	Prevents the device from sending redirects.
Step 10	**ip address** *ip-address/length* **Example:** `switch(config-if)# ip address 20.1.1.10/24`	Configures an IP address for this interface.
Step 11	**no ipv6 redirects** **Example:** `switch(config-if)# no ipv6 redirects`	Disables the ICMP redirect messages on BFD-enabled interfaces.
Step 12	**fabric forwarding mode anycast-gateway** **Example:** `switch(config-if)# fabric forwarding mode anycast-gateway`	Configure Anycast gateway forwarding mode.



CHAPTER 33

Configuring PVLANs

This chapter contains these sections:

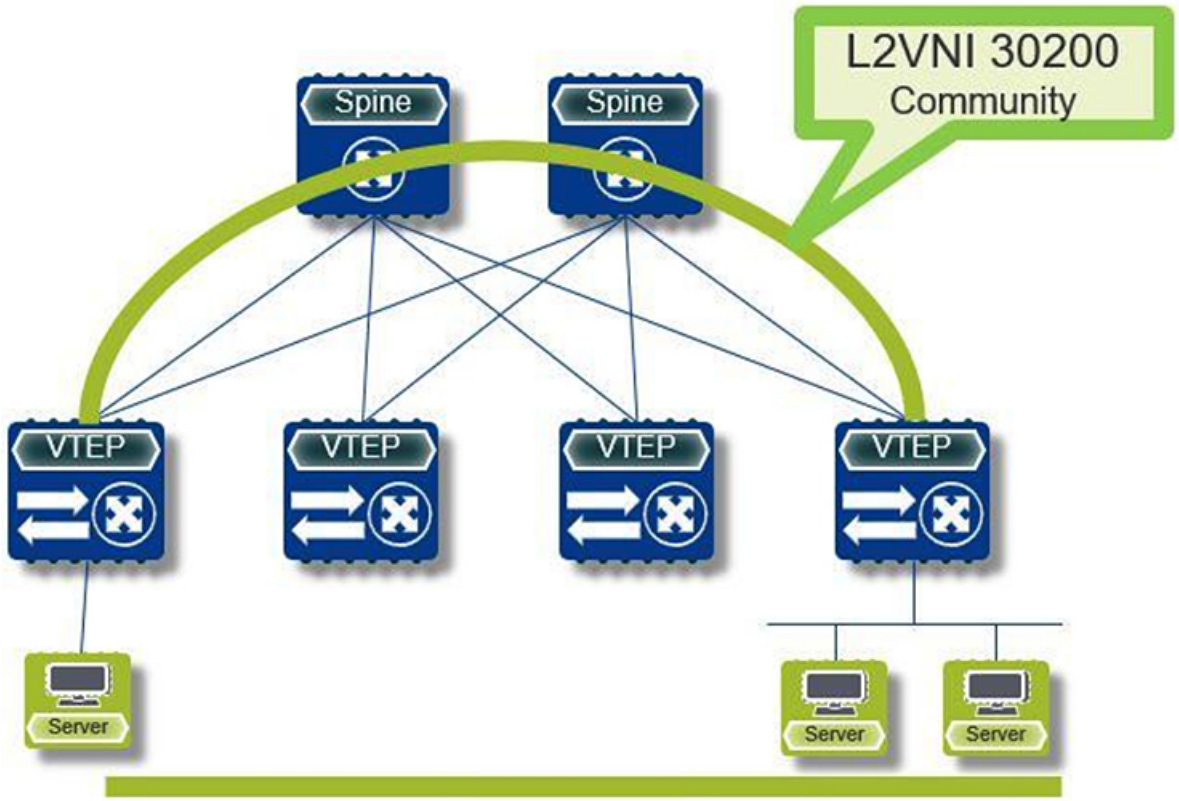
- About Private VLANs over VXLAN, on page 749
- Guidelines and Limitations for Private VLANs over VXLAN, on page 750
- Configuration Example for Private VLANs, on page 751

About Private VLANs over VXLAN

The private VLAN feature allows segmenting the Layer 2 broadcast domain of a VLAN into subdomains. A subdomain is represented by a pair of private VLANs: a primary VLAN and a secondary VLAN. A private VLAN domain can have multiple private VLAN pairs, one pair for each subdomain. All VLAN pairs in a private VLAN domain share the same primary VLAN. The secondary VLAN ID differentiates one subdomain from another.

Private VLANs over VXLAN extends private VLAN across VXLAN. The secondary VLAN can exist on multiple VTEPs across VXLAN. MAC address learning happens over the primary VLAN and advertises via BGP EVPN. When traffic is encapsulated, the VNI used is that of the secondary VLAN. The feature also supports Anycast Gateway. Anycast Gateway must be defined using the primary VLAN.

Figure 76: L2VNI 30200 Community



Guidelines and Limitations for Private VLANs over VXLAN

Private VLANs over VXLAN has the following configuration guidelines and limitations:

- The following platforms support private VLANs over VXLAN:
  - Cisco Nexus 9300-FX/FX2 platform switches
  - Cisco Nexus 9300-GX platform switches
- Beginning with Cisco NX-OS Release 9.3(9), PVLAN configuration is not allowed on vPC Peer-link interfaces.
- Beginning with Cisco NX-OS Release 10.2(3)F, the private VLANs over VXLAN is supported on the Cisco Nexus 9300-FX3/GX2 platform switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, the private VLANs over VXLAN is supported on the Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, the private VLANs over VXLAN is supported on the Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, the private VLANs over VXLAN is supported on the Cisco Nexus 9364C-H1 switches.

- Flood and learn underlay is not supported.
- Fabric Extenders (FEX) VLAN cannot be mapped to a private VLAN.
- vPC Fabric Peering supports private VLANs.
- From Cisco NX-OS Release 10.4(1)F, Private VLAN is supported on Cisco Nexus C9348GC-FX3 and C9348GC-FX3PH switches.
- Beginning with Cisco NX-OS Release 10.5(2)F, Private VLAN is supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.

Configuration Example for Private VLANs

The following is a private VLAN configuration example:

```
vlan 500
  private-vlan primary
  private-vlan association 501-503
  vn-segment 5000
vlan 501
  private-vlan isolated
  vn-segment 5001
vlan 502
  private-vlan community
  vn-segment 5002
vlan 503
  private-vlan community
  vn-segment 5003

vlan 1001
  !L3 VNI for tenant VRF
  vn-segment 900001

interface Vlan500
  no shutdown
  private-vlan mapping 501-503
  vrf member vxlan-900001
  no ip redirects
  ip address 50.1.1.1/8
  ipv6 address 50::1:1:1/64
  no ipv6 redirects
  fabric forwarding mode anycast-gateway

interface Vlan1001
  no shutdown
  vrf member vxlan-900001
  no ip redirects
  ip forward
  ipv6 forward
  ipv6 address use-link-local-only
  no ipv6 redirects

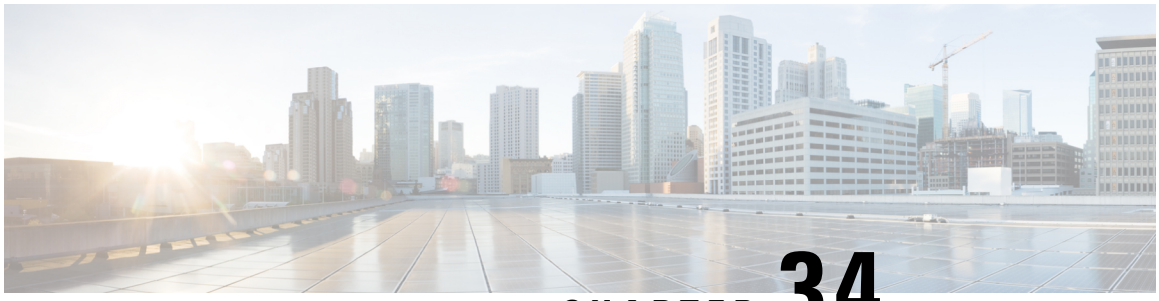
interface nve 1
  no shutdown
  host-reachability protocol bgp
  source-interface loopback0
  member vni 5000
    mcast-group 225.5.0.1
  member vni 5001
```

```
    mcast-group 225.5.0.2
  member vni 5002
    ingress-replication protocol bgp
  member vni 5003
    mcast-group 225.5.0.4
member vni 900001 associate-vrf
```



Note If you use an external gateway, the interface towards the external router must be configured as a PVLAN promiscuous port

```
interface ethernet 2/1
switchport
switchport mode private-vlan trunk promiscuous
switchport private-vlan mapping trunk 500 199,200,201
exit
```



CHAPTER 34

Configure First Hop Security

This chapter contains these sections:

- DHCP Snooping in VXLAN BGP EVPN , on page 753
- How DHCP Snooping on VXLAN works, on page 754
- Guidelines and limitations for DHCP Snooping on VXLAN, on page 756
- Prerequisites for DHCP Snooping, on page 758
- Enable DHCP Snooping on VXLAN, on page 758
- Clear the duplicate host after permanent freeze, on page 759
- Verification of DHCP Snooping configuration , on page 760

DHCP Snooping in VXLAN BGP EVPN

DHCP Snooping in VXLAN BGP EVPN is a process that

- validates ARP/GARP packets sent from a host, preventing ARP spoofing and malicious ARP storms,
- validates data-plane traffic from the host using IPSG, preventing malicious hosts from sending data traffic, and
- replicates DHCP Snooping entries across the VXLAN fabric, enabling DAI and IPSG to function across the fabric even after a host move.

First Hop Security

First Hop Security (FHS) is an access security feature that

- provides security to the network at the access point where the host connects to the first switch,
- authorizes and authenticates hosts, and
- protects the network by ensuring only authorized hosts are allowed network access.

The Dot1x, port-security, and DHCP Snooping are examples of access security features.

DHCP Snooping database

DHCP Snooping database (DB) is a database that

- contains the MAC address of the host, the IP address assigned to the host by the DHCP server, VLAN, and other details like the lease time,
- can contain local or remote snooping DB entries, and
- can be configured using the **ip source binding** *ip address* **vlan** *vlan-id* **interface** *interface* interface command.



Note Snooping entries added through this command are referred to as static entries, and they are distributed across all VTEPs.

Distributed DHCP Snooping database

Distributed DHCP Snooping DB is a database that

- validates ARPs/GARPs sent from the host using DAI,



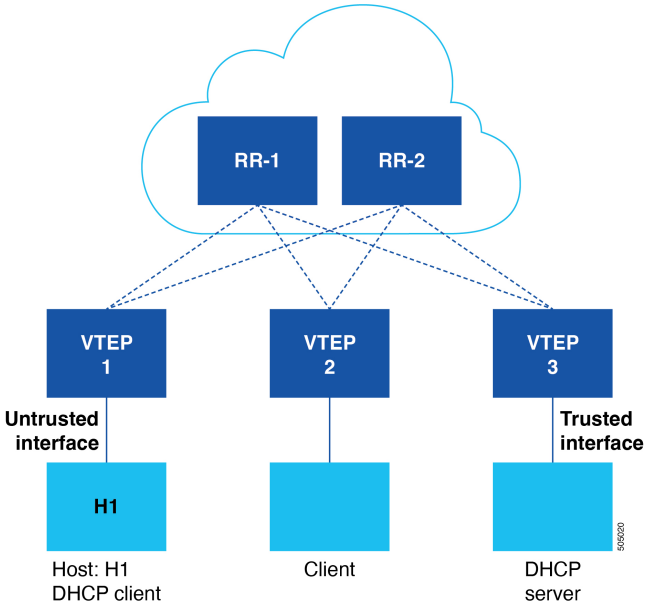
Note The ARP/GARP will be dropped if there is no matching entry in the DB.

- validates data-plane traffic from the host using IPSG, and
- is replicated across the fabric.

How DHCP Snooping on VXLAN works

The DHCP Snooping on VXLAN process involves exchanges between your host and a DHCP server across a VXLAN fabric. It creates and distributes a DHCP Snooping database, which validates ARP/GARP packets and data-plane traffic.

Figure 77: DHCP Snooping on VXLAN



The key components involved in the process are:

- **Host**: Requests an IP address. Host H1 is attached to VTEP1
- **VTEP**: Connects hosts to the VXLAN fabric and enforces security policies.
- **DHCP Server**: Assigns IP addresses and configuration parameters to hosts. DHCP server is attached to VTEP3.
- **VXLAN Fabric**: The network infrastructure that connects VTEPs and enables communication between hosts and servers.
- **DHCP Snooping Database**: Stores IP-to-MAC address bindings and other host information.

The process includes these stages:

- The host sends a DHCP Discover message to find a DHCP server.
- The DHCP server sends a DHCP Offer message to the host, proposing an IP address.
- The host sends a DHCP Request message to accept the offered IP address.
- The DHCP server sends a DHCP Ack message to confirm the IP address assignment.

Note

- The host and the DHCP server exchange a set of messages as part of this host IP assignment procedure. These are known as Discover-Offer-Request-Ack (DORA) exchange messages.
- The DORA exchange, for a particular host (H1), must now be sent over the VXLAN fabric to reach remote DHCP servers (VTEP3).
- VTEP3 checks whether the 'Offer' and 'Ack' messages (part of a DORA sequence) coming from the DHCP server are received via a Trusted Interface.

- Upon completion of the DORA exchange, VTEP1 creates a local DHCP Snooping database entry for the host.
- The local DHCP Snooping database entry is propagated to remote VTEPs using BGP-EVPN.
- Remote VTEPs store the snooping database entry as a remote entry.
- DAI validates ARP/GARP packets against the DHCP Snooping database to prevent spoofing.
- IPSG validates data-plane traffic against the DHCP Snooping database to prevent malicious traffic.

  In the IPSG, only local DHCP clients for that VTEP are programmed. The local DHCP clients are identified with anchor flag set to true in the DHCP Snooping table. If a host moves to a different VTEP and settles down, IPSG has to reprogram the client behind the new VTEP to validate the data-traffic. On the old VTEP, IPSG has to remove this DHCP client. The anchor flag will change accordingly. The host move is triggered by the receipt of an ARP request from the host which is received on the new VTEP that the host moved to.
- The DHCP Snooping database updates to reflect the new location when a host moves to a different VTEP.

  This DHCP Snooping database will be seen as a 'Distributed Database' across the VTEPs, and the snooping entries will be in sync with all VTEPs.

The DHCP Snooping on VXLAN process ensures secure IP address assignment and prevents malicious activities by validating traffic against a distributed database of IP-to-MAC address bindings.

Guidelines and limitations for DHCP Snooping on VXLAN

DHCP Snooping on VXLAN feature has the following guidelines and limitations:

Configuration guidelines and limitations

- Ensure that the DHCP Snooping, DAI and IPSG together are enabled on all VTEPs.

Note DAI and IPSG depend on DHCP Snooping. DHCP Snooping creates the snooping DB and this DB is used by DAI and IPSG.

- The host-move is indicated by ARP/GARP/RARP receipt. In case of RARP (which contains MAC info alone), VTEPs start ARP Refreshes for the IPs learned against MAC. ARP-GARP is essentially the trigger for host-move and not any other data packet.
- In the ingress SUP region, the TCAM must be carved out to 768 entries instead of the default 512 entries to set up the ingress ACLs using the **hardware access-list tcam region ing-sup** command. Reload of a switch is required for the TCAM carving changes to reflect.
- If an I/O module lacks the resources required for an atomic update, you can disable atomic updates by using the **no hardware access-list update atomic** command; however, during the brief time required for the device to remove the preexisting ACL and implement the updated ACL, traffic that the ACL applies to is dropped by default.
- If you want to permit all traffic that an ACL applies to while it receives a nonatomic update, use the **hardware access-list update default-result permit** command.
- In case of multisite and with vPC BGW, if DHCP Snooping is enabled on the vPC BGW, ensure that DHCP clients and DHCP servers are on the same sites.



Note

- DHCP Snooping needs to be enabled (on a VTEP) for the VLAN belonging to the DHCP host that must avail the DHCP service.
- All the VLANs serviced by the DHCP server in the fabric should be enabled with DHCP Snooping on all the VTEPs of the fabric.

Supported features and platforms

- Beginning with Cisco NX-OS Release 10.4(1)F, DHCP Snooping and associated features such as Dynamic ARP Inspection (DAI) and IP Source Guard (IPSG) support is extended to VXLAN fabric on Cisco Nexus 9300-FX/FX2/FX3/GX/GX2 platform switches and Cisco Nexus 9500 switches with X97160YC-EX and 9700-FX/GX line cards.

  Beginning with Cisco NX-OS Release 10.4(2)F, First Hop Security feature is supported on Cisco Nexus 9332D-H2R, and 93400LD-H1 switches.

  Beginning with Cisco NX-OS Release 10.4(3)F, First Hop Security feature is supported on Cisco Nexus 9364C-H1 switches.
- Beginning with Cisco NX-OS Release 10.5(2)F, First Hop Security feature is supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.
- Only IPv4 multicast underlay is supported. However, IPv4 ingress replication underlay, IPv6 ingress replication underlay and IPv6 multicast underlay are not supported.
- Only IPv4 DHCP hosts is supported.
- For vPC VTEPs, only physical MCT is supported.
- On vPC nodes, static DHCP Snooping is supported only with vPC port-channel ports and not with orphan ports.

Unsupported features and platforms

- First Hop Security feature is not supported on EoR.
- The DHCP server cannot be deployed behind the EoR.
- This feature cannot coexist with FabricPath to VXLAN migration feature, and the counter ACL (CNT ACL) feature.

Prerequisites for DHCP Snooping

DHCP Snooping has these prerequisites:

- You should be familiar with DHCP before you configure DHCP Snooping or the DHCP relay agent.
- Make sure that the DHCP Snooping, DAI and IPSG features are enabled together on a leaf VTEP.

Enable DHCP Snooping on VXLAN

You can enable or disable DHCP Snooping on a single-box feature or enable this feature for a VLAN for the entire fabric. By default, DHCP Snooping is disabled on all VLANs.

Follow these steps to enable DHCP Snooping on VXLAN

Before you begin

- Ensure you enable the DHCP feature.
- Ensure you configure the **nv overlay evpn** command.
- Ensure you enable the DHCP Snooping, DAI, and IPSG features. For more information see the Prerequisites for DHCP Snooping, on page 758 section.
- Ensure you enable the DHCP Snooping and DAI features on all the VXLAN nodes. For more information on configuration, see **Configuring DHCP Snooping** section of Cisco Nexus 9000 Series NX-OS Security Configuration Guide.
- Ensure you enable the DHCP Snooping trust and ARP inspection trust on interfaces connected to the DHCP server nodes. For more information on configuration, see **Configuring DHCP Snooping** section of Cisco Nexus 9000 Series NX-OS Security Configuration Guide.
- Ensure you enable the IP Source Guard feature on the interfaces connected to the DHCP client nodes. For more information on configuration, see **Configuring DHCP Snooping** section of Cisco Nexus 9000 Series NX-OS Security Configuration Guide.

Procedure

Step 1 Run the **configure terminal** command to enter global configuration mode.

Example:

```
switch# configure terminal
switch(config)#
```

Step 2 Run the [**no**] **ip dhcp snooping vlan** *vlan-list* **evpn** command to enable DHCP Snooping on the VLANs specified by *vlan-list* .

Example:

```
switch(config)# ip dhcp snooping vlan 100,200,250-252 evpn
```

Beginning with Cisco NX-OS Release 10.4(1)F, the **evpn** option is provided to support host move to other interfaces on the same VTEP or other VTEPs.

Note

- When this feature is enabled with the **evpn** option, the **nve** will be implicitly added as a trusted interface.
- It is possible to have one VLAN list with **evpn** keyword and another VLAN list with **no evpn** keyword.

The **no** form of this command disables DHCP Snooping on the VLANs specified.

Step 3 (Optional) Run the **show running-config dhcp** command to validate the DHCP configuration.

Example:

```
switch(config)# show running-config dhcp
```

Step 4 (Optional) Run the **copy running-config startup-config** command to copy the running configuration to the startup configuration.

Example:

```
switch(config)# copy running-config startup-config
```

Clear the duplicate host after permanent freeze

If the MAC or MAC-IP address becomes permanently frozen, you cannot automatically recover to restart mobility or check for duplicates.

The mobility and duplicate detection logic for DHCP clients in FHS-enabled VTEPs is the same as the BGP EVPN's mobility and duplicate detection logic. However, duplicate detection may happen in any of the VTEPs in non-FHS deployments. In FHS deployments, the host duplicate will always be detected on a VTEP where the DHCP binding entry is remote.

For more information on mobility and duplicate detection, see the Duplicate Detection for IP and MAC Addresses, on page 143 section.

Once the MAC or MAC-IP is permanently frozen, there is no auto recovery mechanism to re-initiate mobility or duplicate check sequences. To clear MAC and MAC-IP permanent freeze state, use these commands:

- For MAC:

```
clear l2route evpn mac [mac-address] [topo] permanently-frozen-list
```

- For MAC-IP:

```
clear fabric forwarding dup-host [{ ip|ipv6 address }] [vrf {vrf-name | vrf-known-name
 | all}]
```

Verification of DHCP Snooping configuration

To validate the DHCP Snooping configuration, enter these commands:

Command	Purpose
show ip dhcp snooping binding evpn	Displays all entries from the DHCP snooping binding database.
show l2route fhs [**topology** *topology id* | **all**]	Displays all entries from the L2RIB database.

The following example shows sample output for the **show ip dhcp snooping binding evpn** command:

```
switch(config)# show ip dhcp snooping binding evpn
MacAddress          IpAddress        Lease(Sec)  Type      BD    Interface       anchor
  Freeze
------------------ -------------- ----------- ---------- ---- ----------------- ----------
---------
00:10:00:10:00:10  10.10.10.10      infinite  static      2001  Ethernet1/48         YES
     NONE
00:15:06:00:00:01  100.1.150.156    86282     dhcp-snoop  2001  Ethernet1/31         YES
     NONE
00:17:06:00:00:01  100.1.150.155    86265     dhcp-snoop  2001  nve1(peer-id: 1)     NO
     NONE
```

The following example shows sample output for the **show l2route fhs** command:

```
switch(config)# show l2route fhs all
Flags - (Stt):Static (Dyn):Dynamic (R):Remote
Topo ID  Mac Address     Host IP             Prod          Flags      Seq No     Next-Hops
---- -- ---- ----------- ------------------ ---------- ----------- ------------ -------------
2001     0015.0600.0001  100.1.150.156      DHCP_DYNAMIC  Dyn,       0          Eth1/31
2001     0017.0600.0001  100.1.150.155      BGP           Dyn,R,     0          1.13.13.13
 (Label: 0)
switch(config)#
```

The following example shows DHCP configurations for a VTEP with DHCP clients:

```
feature dhcp
service dhcp
ip dhcp snooping
ip dhcp snooping vlan 2001-2002 evpn
ip arp inspection vlan 2001-2002

interface Ethernet1/31
ip verify source dhcp-snooping-vlan
```

The following example shows DHCP configurations for a VTEP with DHCP server:

```
feature dhcp
service dhcp
ip dhcp snooping
ip dhcp snooping vlan 2001-2002 evpn
ip arp inspection vlan 2001-2002

interface Ethernet1/47
ip dhcp snooping trust
ip arp inspection trust
```



PART II

Configure Multi-Homing

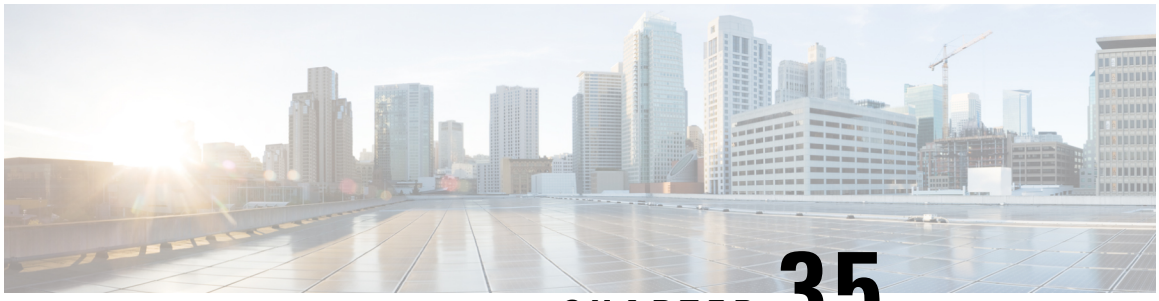
- Cisco Nexus supported Multi-homing methods, on page 761
- Configuring vPC Multi-Homing, on page 763
- Configuring vPC Fabric Peering, on page 765
- Interoperability with EVPN Multi-Homing Using ESI, on page 777
- EVPN Ethernet Segment Identifier Multi-Homing, on page 783

Cisco Nexus supported Multi-homing methods

Cisco Nexus switches support three primary methods for host multi-homing:

1. **vPC with Dedicated Physical Peer-Link (PMCT):** This traditional vPC implementation uses a dedicated physical link between two vPC peer switches for redundancy and state synchronization via Cisco Forwarding Service (CFS).
2. **vPC with Fabric Peering (vMCT):** This method also uses vPC but establishes a virtual peer-link over the VXLAN fabric, eliminating the need for a separate physical connection while still using CFS for state synchronization.
3. **ESI-Multihoming (Based on EVPN):** This modern approach leverages the industry-standard BGP Ethernet VPN (EVPN) control plane over the VXLAN fabric. It does not require a dedicated peer-link and can support redundancy with more than two switches (up to four-way).

These chapters provide more detailed information about these methods.



CHAPTER 35

Configuring vPC Multi-Homing

This chapter contains these sections:

- Advertising Primary IP Address, on page 763
- Border PE Switches in a vPC Setup, on page 764
- DHCP Configuration in a vPC Setup, on page 764
- IP Prefix Advertisement in vPC Setup, on page 764

Advertising Primary IP Address

On a vPC enabled leaf or border leaf switch, by default all Layer-3 routes are advertised with the secondary IP address (VIP) of the leaf switch VTEP as the BGP next-hop IP address. Prefix routes and leaf switch generated routes are not synced between vPC leaf switches. Using the VIP as the BGP next-hop for these types of routes can cause traffic to be forwarded to the wrong vPC leaf or border leaf switch and black-holed. The provision to use the primary IP address (PIP) as the next-hop when advertising prefix routes or loopback interface routes in BGP on vPC enabled leaf or border leaf switches allows users to select the PIP as BGP next-hop when advertising these types of routes, so that traffic will always be forwarded to the right vPC enabled leaf or border leaf switch.

The configuration command for advertising the PIP is **advertise-pip**.

Note On the Cisco Nexus 9300-FX2 switch, the **advertise-pip** command was not supported prior to Cisco NX-OS Release 7.0(3)I7(4). For more information, see CSCvi42831.

The following is a sample configuration:

```
switch(config)# router bgp 65536
  address-family 12vpn evpn
    advertise-pip
interface nve 1
    advertise virtual-rmac
```

The **advertise-pip** command lets BGP use the PIP as next-hop when advertising externally learned routes or for the redistributed direct routes if vPC is enabled.

VMAC (virtual-mac) is used with VIP and system MAC is used with PIP when the VIP/PIP feature is enabled.

With the **advertise-pip** and **advertise virtual-rmac** commands enabled, type 5 routes are advertised with PIP and type 2 routes are still advertised with VIP. In addition, VMAC will be used with VIP and system MAC will be used with PIP.



Note The **advertise-pip** and **advertise-virtual-rmac** commands must be enabled and disabled together for this feature to work properly. If you enable or disable one and not the other, it is considered an invalid configuration.

Border PE Switches in a vPC Setup

The two border PE switches are configured as a vPC. In a VXLAN vPC deployment, a common, virtual VTEP IP address (secondary loopback IP address) is used for communication. The common, virtual VTEP uses a system specific router MAC address. The Layer-3 prefixes or default route from the border PE switch is advertised with this common virtual VTEP IP (secondary IP) plus the system specific router MAC address as the next hop.

Entering the **advertise-pip** and **advertise virtual-rmac** commands cause the Layer 3 prefixes or default to be advertised with the primary IP and system-specific router MAC address, the MAC addresses to be advertised with the secondary IP, and a router MAC address derived from the secondary IP address.

DHCP Configuration in a vPC Setup

When DHCP or DHCPv6 relay function is configured on leaf switches in a vPC setup, and the DHCP server is in the non default, non management VRF, then configure the **advertise-pip** command on the vPC leaf switches. This allows BGP EVPN to advertise Route-type 5 routes with the next-hop using the primary IP address of the VTEP interface.

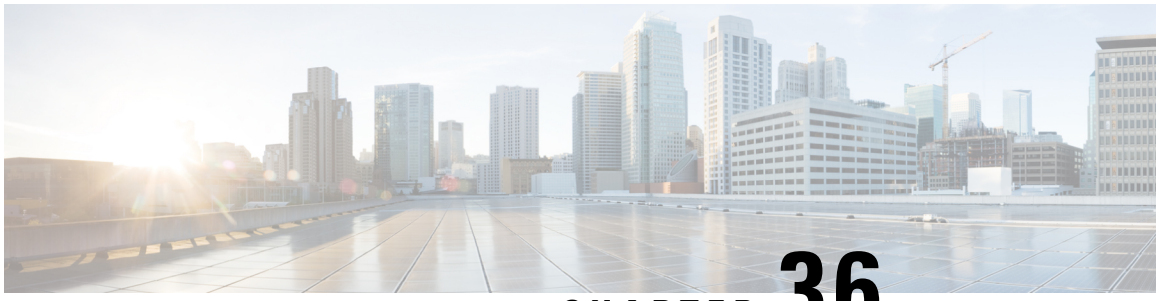
The following is a sample configuration:

```
switch(config)# router bgp 100
  address-family 12vpn evpn
    advertise-pip
interface nve 1
  advertise virtual-rmac
```

IP Prefix Advertisement in vPC Setup

There are 3 types of Layer-3 routes that can be advertised by BGP EVPN. They are:

- Local host routes—These routes are learned from the attached servers or hosts.
- Prefix routes—These routes are learned via other routing protocol at the leaf, border leaf and border spine switches.
- Leaf switch generated routes—These routes include interface routes and static routes.



CHAPTER 36

Configuring vPC Fabric Peering

This chapter contains these sections:

- Information About vPC Fabric Peering, on page 765
- Guidelines and Limitations for vPC Fabric Peering , on page 766
- Configuring vPC Fabric Peering, on page 768
- Migrating from vPC to vPC Fabric Peering, on page 772
- Verifying vPC Fabric Peering Configuration, on page 775

Information About vPC Fabric Peering

vPC Fabric Peering provides an enhanced dual-homing access solution without the overhead of wasting physical ports for vPC Peer Link.

The following lists the vPC Fabric Peering solution:

- vPC Fabric Peering port-channel with virtual members (tunnels).
- vPC Fabric Peering (tunnel) with removal of the physical peer link requirement.
- vPC Fabric Peering up/down events are triggered based on route updates and fabric up/down.
- Uplink tracking for extended failure coverage.
- vPC Fabric Peering reachability via the routed network, such as the spine.
- Increased resiliency of the vPC control plane over TCP-IP (CFSoIP).
- Data plane traffic over the VXLAN tunnel.
- Communication between vPC member switches uses VXLAN encapsulation.
- Failure of all uplinks on a node result in vPC ports going down on that switch. In that scenario, vPC peer takes up the primary role and forwards the traffic.
- Uplink tracking with state dependency and up/down signalization for vPCs.
- Positive uplink state tracking drives vPC primary role election.
- For border leafs and spines, there is no need for per-VRF peering since network communication uses the fabric.
- Enhance forwarding to orphans hosts by extending the VIP/PIP feature to Type-2 routes.

- Infra-VLAN is not required for vPC fabric peering.

Note The vPC Fabric Peering counts as three VTEPs unlike a normal vPC which counts as one VTEP.

Guidelines and Limitations for vPC Fabric Peering

The following are the vPC Fabric Peering guidelines and limitations:

Configuration recommendations

- vPC Fabric Peering requires TCAM carving of the region **ing-flow-redirect**. TCAM carving requires saving the configuration and reloading the switch prior to using the feature.

Note This requirement only applies to Cisco Nexus 9300-FX, 9300-FX2, 9300-FX3 platform switches.

- When using vPC fabric peering, you cannot create routing over SVIs for such vPC pairs.
- Prior to reconfiguring the vPC Fabric Peering source and destination IP, the vPC domain must be shut down. Once the vPC Fabric Peering source and destination IP have been adjusted, the vPC domain can be enabled (**no shutdown**).
- The source and destination IP supported in **virtual peer-link destination** command are class A, B, and C. Class D and E are not supported for vPC Fabric Peering.
- Immediately after converting from fabric peering to a physical peer link, make the following changes on both peers:
  1. Globally configure a TCAM region using the **hardware access-list tcam region ing-flow-redirect 0** command.
  2. Optionally, allocate the free space to other classes. For more information, see Understand How to Carve Nexus 9000 TCAM Space.
  3. Save the running configuration using the **copy running-config startup-config** command.
  4. Reload the switch.
- The vPC Fabric Peering peer-link is established over the transport network (the spine layer of the fabric). As communication between vPC peers occurs in this manner, control plane information CFS messages used to synchronize port state information, VLAN information, VLAN-to-VNI mapping, host MAC addresses are transmitted over the fabric. CFS messages are marked with the appropriate DSCP value, which should be protected in the transport network. The following example shows a sample QoS configuration on the spine layer of Cisco Nexus 9000 Series switches.

  Classify traffic by matching the DSCP value (DSCP 56 is the default value):

```
class-map type qos match-all CFS
  match dscp 56
```

Set traffic to the qos-group that corresponds with the strict priority queue for the appropriate spine switch. In this example, the switch sends traffic to qos-group 7, which corresponds to the strict priority queue (Queue 7). Note that different Cisco Nexus platforms might have a different queuing structure.

```
policy-map type qos CFS
  class CFS
    Set qos-group 7
```

Assign a classification service policy to all interfaces toward the VTEP (the leaf layer of the network):

```
interface Ethernet 1/1
  service-policy type qos input CFS
```

- Layer 3 Tenant Routed Multicast (TRM) is supported. Layer 2/Layer 3 TRM (Mixed Mode) is not supported.
- If Type-5 routes are used with this feature, the **advertise-pip** command is a mandatory configuration.
- Enhance forwarding to orphan hosts by extending the VIP/PIP feature to Type-2 routes.
- An orphan Type-2 host is advertised using PIP. A vPC Type-2 host is advertised using VIP. This is the default behavior for a Type-2 host.

  To advertise an orphan Type-5 route using PIP, you need to advertise PIP under BGP.
- For orphan ports, it is highly recommended to configure **vpc orphan-port suspend** command on both vPC nodes, to avoid traffic disruption during NVE failure scenarios.
- Traffic from remote VTEP to orphan hosts would land on the actual node which has the orphans. Bouncing of the traffic is avoided.



Note When the vPC leg is down, vPC hosts are still advertised with the VIP IP.

- ARP behavior differs between vPC fabric peering and physical peer links for orphan ports. vPC fabric peering does not sync ARP entries as it does with physical peer links. An orphan Type-2 host is advertised using the PIP in vPC Fabric peering.
- Beginning with Cisco NX-OS Release 10.6(1)F, vPC auto-recovery requires at least one operational fabric uplink. No auto-recovery occurs if all are down.

Supported feature, release and platforms

- Cisco Nexus9300-FX/FXP/FX2/FX3/GX/GX2/H2R/H1 platform switches support vPC Fabric Peering. Cisco Nexus 9500 platform switches do not support vPC Fabric Peering.
- Beginning with Cisco NX-OS Release 10.1(1), FEX Support is provided with vMCT for IPv4 underlay on Cisco Nexus 9300-FX/FX2/FX3 platform switches.
- Beginning with Cisco NX-OS Release 10.2(2)F, FEX Support is provided with vMCT for IPv4 underlay on Cisco Nexus 9300-GX platform switches.

- Beginning with Cisco NX-OS Release 10.1(1), vPC Fabric Peering supports FEX in Straight Through and Active-Active (dual home) modes in N9K-C9336C-FX2-E, N9K-C93108TC-FX, N9K-C93180YC-FX, N9K-C93216TC-FX2, N9K-C93240YC-FX2, N9K-C93360YC-FX2, N9K-C9336C-FX2, N9K-C93180YC-FX3, N9K-C93180YC-FX3S platform switches.

  Refer to *Cisco Nexus 2000 Series NX-OS Fabric Extender Configuration Guide for Cisco Nexus 9000 Series Switches* for details on FEX (Straight Through and Active-Active modes).

- Beginning with Cisco NX-OS Release 10.2(3)F, ND-ISSU and LXC-ISSU are supported with vMCT for IPv4 underlay on Cisco Nexus 9300-FX/FXP/FX2/FX3/GX/GX2 ToR switches.
- Beginning with Cisco NX-OS Release 10.3(2)F, the vPC Fabric Peering is supported for IPv6 underlay on Cisco Nexus 9300-FX/FXP/FX2/FX3/GX/GX2 ToR switches.
- Beginning with Cisco NX-OS Release 10.3(2)F, ND-ISSU and LXC-ISSU are supported with vMCT for IPv6 underlay on Cisco Nexus 9300-FX/FXP/FX2/FX3/GX/GX2 ToR switches.
- Beginning with Cisco NX-OS Release 10.4(1)F, the vPC Fabric Peering is supported for IPv6 underlay on Cisco Nexus 9332D-H2R switches.
- Beginning with Cisco NX-OS Release 10.4(2)F, the vPC Fabric Peering is supported for IPv6 underlay on Cisco Nexus 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.4(3)F, vPC Fabric Peering is supported for IPv6 underlay on Cisco Nexus 9364C-H1 switches.
- Beginning with Cisco NX-OS Release 10.6(1)F, Standalone leaf, vPC leaf, and vPC Fabric Peering leaf are supported on Cisco Nexus 9336C-SE1 switches.

Unsupported features

- The vPC Fabric Peering domain is not supported in the role of a Multi-Site vPC BGW.
- vMCT for IPv6 underlay does not support attaching FEX to it.
- VTEPs behind vPC ports are not supported. This means that virtual peer-link peers cannot act as a transit node for the VTEPs behind the vPC ports.
- SVI and sub-interface uplinks are not supported.

Configuring vPC Fabric Peering

Ensure the vPC Fabric Peering DSCP value is consistent on both vPC member switches. Ensure that the corresponding QoS policy matches the vPC Fabric Peering DSCP marking.

All VLANs that require communication traversing the vPC Fabric Peering must have a VXLAN enabled (vn-segment); this includes the native VLAN.



Note For MSTP, VLAN 1 must be extended across vPC Fabric Peering if the peer-link and vPC legs have the default native VLAN configuration. This behavior can be achieved by extending VLAN 1 over VXLAN (vn-segment). If the peer-link and vPC legs have non-default native VLANs, those VLANs must be extended across vPC Fabric Peering by associating the VLANs with VXLAN (vn-segment).

Use the **show vpc virtual-peerlink vlan consistency** command for verification of the existing VLAN-to-VXLAN mapping used for vPC Fabric Peering.

peer-keepalive command for vPC Fabric Peering is supported with one of the following configurations:

- Management interface
- Dedicated Layer 3 link in default or non-default VRF
- Loopback interface reachable using the spine.

Configuring Features

Example uses OSPF as the underlay routing protocol.

```
configure terminal
nv overlay evpn
feature ospf
feature bgp
feature pim
feature interface-vlan
feature vn-segment-vlan-based
feature vpc

feature nv overlay
```

vPC Configuration



Note To change the vPC Fabric Peering source or destination IP, the vPC domain must be shutdown prior to modification. The vPC domain can be returned to operation after the modifying by using the **no shutdown** command.

Configuring TCAM Carving

```
hardware access-list tcam region ing-racl 0
hardware access-list tcam region ing-sup 768
hardware access-list tcam region ing-flow-redirect 512
```



Note

- When configuring fabric vPC peering, the minimum size for Ingress-Flow-redirect TCAM region size is 512. Also ensure that the TCAM region size is always configured in multiples of 512.
-
- Switch reload is required for the TCAM carving to take effect.

Configuring the vPC Domain

For IPv4

```
vpc domain 100
peer-keepalive destination 192.0.2.1
virtual peer-link destination 192.0.2.100 source 192.0.2.20/32 [dscp <dscp-value>]
Warning: Appropriate TCAM carving must be configured for virtual peer-link vPC
peer-switch
peer-gateway
ip arp synchronize
```

```
ipv6 nd synchronize
exit
```

For IPv6

```
vpc domain 100
peer-keepalive destination 192:0:2::1
virtual peer-link destination 192:0:2::100 source 192:0:2::20/32 [dscp <dscp-value>]
Warning: Appropriate TCAM carving must be configured for virtual peer-link vPC
peer-switch
peer-gateway
ipv6 arp synchronize
ipv6 nd synchronize
exit
```



Note The **dscp** keyword in optional. Range is 1 to 63. The default value is 56.

Configuring vPC Fabric Peering Port Channel

No need to configure members for the following port channel.

```
interface port-channel 10
switchport
switchport mode trunk
vpc peer-link
```

```
interface loopback0
```



Note This loopback is not the NVE source-interface loopback (interface used for the VTEP IP address).

For IPv4

```
interface loopback 0
ip address 192.0.2.20/32
ip router ospf 1 area 0.0.0.0
```

For IPv6

```
interface loopback 0
ipv6 address 192:0:2::20/32
ipv6 router ospfv3 1 area 0.0.0.0
```



Note You can use the loopback for BGP peering or a dedicated loopback. This lookback must be different that the loopback for peer keep alive.

Configuring the Underlay Interfaces

Both L3 physical and L3 port channels are supported. SVI and sub-interfaces are not supported.

For IPv4

```
router ospf 1
interface Ethernet1/16
port-type fabric
ip address 192.0.2.2/24
ip router ospf 1 area 0.0.0.0
```

```
no shutdown
interface Ethernet1/17
port-type fabric
ip address 192.0.2.3/24
ip router ospf 1 area 0.0.0.0
no shutdown
interface Ethernet1/40
port-type fabric
ip address 192.0.2.4/24
ip router ospf 1 area 0.0.0.0
no shutdown
interface Ethernet1/41
port-type fabric
ip address 192.0.2.5/24
ip router ospf 1 area 0.0.0.0
no shutdown
```

For IPv6

```
router ospfv3 1
interface Ethernet1/16
port-type fabric
ipv6 address 192:0:2::2/24
ipv6 router ospfv3 1 area 0.0.0.0
no shutdown
interface Ethernet1/17
port-type fabric
ipv6 address 192:0:2::3/24
ipv6 router ospfv3 1 area 0.0.0.0
no shutdown
interface Ethernet1/40
port-type fabric
ipv6 address 192:0:2::4/24
ipv6 router ospfv3 1 area 0.0.0.0
no shutdown
interface Ethernet1/41
port-type fabric
ipv6 address 192:0:2::5/24
ipv6 router ospfv3 1 area 0.0.0.0
no shutdown
```

Note All ports connected to spines must be port-type fabric.

VXLAN Configuration

Note Configuring **advertise virtual-rmac** (NVE) and **advertise-pip** (BGP) are required steps. For more information, see the Configuring vPC Multi-Homing, on page 763 chapter.

Configuring VLANs and SVI

```
vlan 10
vn-segment 10010
vlan 101
vn-segment 10101

interface Vlan101
no shutdown
```

```
mtu 9216
vrf member vxlan-10101
no ip redirects
ip forward
ipv6 address use-link-local-only
no ipv6 redirects
interface vlan10
no shutdown
mtu 9216
vrf member vxlan-10101
no ip redirects
ip address 192.0.2.102/24
ipv6 address 2001:DB8:0:1::1/64
no ipv6 redirects
fabric forwarding mode anycast-gateway
```

Configuring Virtual Port Channel

```
interface Ethernet1/3
switchport
switchport mode trunk
channel-group 100
no shutdown
exit
interface Ethernet1/39
switchport
switchport mode trunk
channel-group 101
no shutdown
interface Ethernet1/46
switchport
switchport mode trunk
channel-group 102
no shutdown
interface port-channel100
vpc 100
interface port-channel101
vpc 101
interface port-channel102
vpc 102
exit
```

Migrating from vPC to vPC Fabric Peering

This procedure contains the steps to migration from a regular vPC to vPC Fabric Peering.

Any direct Layer 3 link between vPC peers should be used only for peer-keep alive. This link should not be used to advertise paths for vPC Fabric Peering loopbacks.



Note This migration is disruptive.

Before you begin

We recommend that you shut all physical Layer 2 links between the vPC peers before migration. We also recommend that you map VLANs with vn-segment before or after migration.

SUMMARY STEPS

1. **configure terminal**
2. **show vpc**
3. **show port-channel summary**
4. **interface ethernet** *slot/port*
5. **no channel-group**
6. Repeat steps 4 and 5 for each interface.
7. **show running-config vpc**
8. **vpc domain** *domain-id*
9. **virtual peer-link destination** *dest-ip* **source** *source-ip*
10. **interface {ethernet | port-channel}** *value*
11. **port-type fabric**
12. (Optional) **show vpc fabric-ports**
13. **virtual peer-link destination** *dest-ip* | *dest_ipv6* **source** *source-ip* | *source_ipv6* **dhcp** *dhcp_val*
14. **hardware access-list tcam region ing-flow-redirect** *tcam-size*
15. **copy running-config startup-config**
16. **reload**

DETAILED STEPS

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**show vpc** **Example:** `switch(config)# show vpc`	Determine the number of members in the port channel.
Step 3	**show port-channel summary** **Example:** `switch(config)# show port-channel summary`	Determine the number of members.
Step 4	**interface ethernet** *slot/port* **Example:** `switch(config)# interface ethernet 1/4`	Specifies the interface you are configuring. **Note** This is the peer link port channel.
Step 5	**no channel-group** **Example:** `switch(config-if)# no channel-group`	Remove vPC peer-link port-channel members. **Note** Disruption occurs following this step.
Step 6	Repeat steps 4 and 5 for each interface.	

	Command or Action	Purpose
	Example:	
Step 7	**show running-config vpc** **Example:** `switch(config-if)# show running-config vpc`	Determine the vPC domain.
Step 8	**vpc domain** *domain-id* **Example:** `switch(config-if)# vpc domain 100`	Enter vPC domain configuration mode.
Step 9	**virtual peer-link destination** *dest-ip* **source** *source-ip* **Example:** `switch(config-vpc-domain)# virtual peer-link destination 192.0.2.1 source 192.0.2.100`	Specify the destination and source IP addresses for vPC fabric peering.
Step 10	**interface {ethernet | port-channel}** *value* **Example:** `switch(config-if)# interface Ethernet1/17`	Specifies the L3 underlay interface you are configuring.
Step 11	**port-type fabric** **Example:** `switch(config-if)# port-type fabric`	Configures port-type fabric for underlay interface. **Note** All ports connected to spines must be port-type fabric.
Step 12	(Optional) **show vpc fabric-ports** **Example:** `switch# show vpc fabric-ports`	Displays the fabric ports connected to spine.

Step 13	**virtual peer-link destination** *dest-ip* | *dest_ipv6* **source** *source-ip* | *source_ipv6* **dhcp** *dhcp_val*
	Example:
	For IPv4
	`switch(config-vpc-domain)# virtual peer-link destination 192.0.2.1 source 192.0.2.100 dhcp 56`
	Example:
	For IPv6
	`switch(config-vpc-domain)# virtual peer-link destination 6001:aaa::11 source 6001:aaa::22 dhcp 56`
		Specify the destination and source IPv4/IPv6 addresses for vPC fabric peering.
		Note
		The IPv4 vPC Fabric peering config works only with the IPv4 VXLAN underlay and the IPv6 vPC Fabric peering config will work only with the IPv6 VXLAN underlay.

Step 14	**hardware access-list tcam region ing-flow-redirect** *tcam-size* **Example:** `switch(config-vpc-domain)# hardware access-list tcam region ing-flow-redirect 512`	Perform TCAM carving. The minimum size for Ingress-Flow-redirect TCAM region size is 512. Also ensure it is configured in multiples of 512.

	Command or Action	Purpose
Step 15	**copy running-config startup-config** **Example:** `switch(config-vpc-domain)# copy running-config startup-config`	Copies the running configuration to the startup configuration.
Step 16	**reload** **Example:** `switch(config-vpc-domain)# reload`	Reboots the switch.

Verifying vPC Fabric Peering Configuration

To display the status for the vPC Fabric Peering configuration, enter one of the following commands:

Table 13: vPC Fabric Peering Verification Commands

Command	Purpose
show vpc fabric-ports	Displays the fabric ports state.
show vpc	Displays information about vPC Fabric Peering mode.
show vpc virtual-peerlink vlan consistency	Displays the VLANs which are not associated with vn-segment.

Example of the show vpc fabric-ports Command

```
switch# show vpc fabric-ports
Number of Fabric port : 9
Number of Fabric port active : 9

Fabric Ports State
------------------------------------
Ethernet1/9 UP
Ethernet1/19/1 ( port-channel151 ) UP
Ethernet1/19/2 ( port-channel151 ) UP
Ethernet1/19/3 UP
Ethernet1/19/4 UP
Ethernet1/20/1 UP
Ethernet1/20/2 ( port-channel152 ) UP
Ethernet1/20/3 ( port-channel152 ) UP
Ethernet1/20/4 ( port-channel152 ) UP
```

Example of the show vpc Command

```
switch# show vpc
Legend:
                (*) - local vPC is down, forwarding via vPC peer-link

vPC domain id                     : 3
Peer status                       : peer adjacency formed ok
vPC keep-alive status             : peer is alive
```

```
Configuration consistency status  : success
Per-vlan consistency status       : success
Type-2 consistency status         : success
vPC role                          : primary
Number of vPCs configured         : 1
Peer Gateway                      : Enabled
Dual-active excluded VLANs        : -
Graceful Consistency Check        : Enabled
Auto-recovery status              : Enabled, timer is off.(timeout = 240s)
Delay-restore status              : Timer is off.(timeout = 30s)
Delay-restore SVI status          : Timer is off.(timeout = 10s)
Operational Layer3 Peer-router    : Disabled
Virtual-peerlink mode             : Enabled

vPC Peer-link status
---------------------------------------------------------------------
id    Port   Status Active vlans
--    ----   ------ --------------------------------------------------
1     Po100  up     1,56,98-600,1001-3401,3500-3525

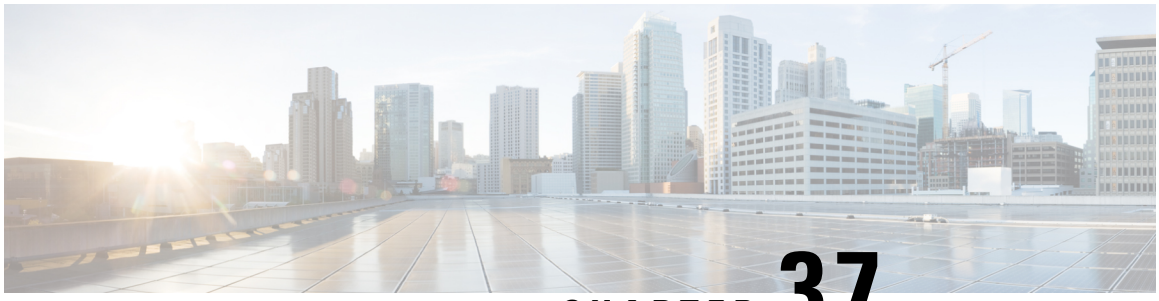
vPC status
----------------------------------------------------------------------------
Id    Port          Status Consistency Reason                Active vlans
--    ------------  ------ ----------- ------                ---------------
101   Po101         up     success     success               98-99,1001-280
                                                           0

Please check "show vpc consistency-parameters vpc <vpc-num>" for the
consistency reason of down vpc and for type-2 consistency reasons for
any vpc.

ToR_B1#
```

Example of the show vpc virtual-peerlink vlan consistency Command

```
switch# show vpc virtual-peerlink vlan consistency
Following vlans are inconsistent
23
switch#
```



CHAPTER 37

Interoperability with EVPN Multi-Homing Using ESI

Beginning Cisco NX-OS Release 10.6(1)F, Cisco Nexus 9300-FX2/FX3/GX/GX2/H2R/H1 Series switches and Cisco Nexus 9500 Series switches with 9700-GX/FX3 line cards provides full support of EVPN multi-homing feature.

Note For more information on the EVPN multi-homing functionality, see the EVPN Ethernet Segment Identifier Multi-Homing, on page 783 chapter.

However, as discussed in the following section, Cisco Nexus 9000 switches can be integrated in the same VXLAN EVPN fabric with switches that fully support the EVPN multi-homing functionality.

This chapter contains these sections:

- Interoperability with EVPN Multi-Homing Using ESI, on page 777
- Guidelines and Limitations for Interoperability with EVPN Multi-Homing using ESI, on page 778
- Example of EVPN Multi-Homing Using ESI, on page 779

Interoperability with EVPN Multi-Homing Using ESI

Beginning Cisco NX-OS Release 10.2(2)F, EVPN MAC/IP routes (Type 2) with non-reserved and with reserved ESI (0 or MAX-ESI) values are evaluated for forwarding (a functionality usually referred to as "ESI RX"). The definition of the EVPN MAC/IP route resolution is defined in RFC 7432 Section 9.2.2.

EVPN MAC/IP routes (Type 2):

- With reserved ESI value (0 or MAX-ESI) is resolved solely by the MAC/IP route alone (BGP next-hop within Type 2).
- With non-reserved ESI value is resolved only if an accompanied per-ES Ethernet Auto-Discovery route (Type 1, per-ES EAD) is present.

The EVPN MAC/IP route resolution with non-reserved ESI values is supported on Cisco Nexus 9300-FX/FX2/FX3/GX Platform Switches.

This means that those switches, while still using vPC multi-homing for locally connected devices (as discussed in the previous , EVPN Ethernet Segment Identifier Multi-Homing, on page 783 and Configuring vPC Fabric

Peering, on page 765sections), can coexist in a VXLAN EVPN fabric with other switches that use EVPN multi-homing for the connectivity of local devices. MAC and IP addresses of remote endpoints are learned from those remote switches using the EVPN control plane messages listed above and get assigned multiple next-hop IP addresses (the unique VTEP addresses identifying each of the switches implementing EVPN multi-homing).

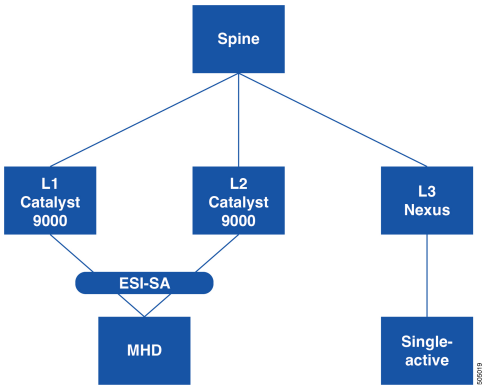
Guidelines and Limitations for Interoperability with EVPN Multi-Homing using ESI

- Until Cisco NX-OS Release 10.4(1)F, Cisco Nexus 9300-FX/FX2/FX3/GX/GX2 switches and 9500 switches with X97160YC-EX, 9700-FX/GX line cards can coexist in a VXLAN fabric with other switches that support ESI multi-homing only in All-active mode. Cisco Nexus 9300-FX switches supports only one active path. However, Single-active mode is not supported.
- Beginning with Cisco NX-OS Release 10.5(2)F, Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line cards can coexist in a VXLAN fabric with other switches that support ESI multi-homing only in All-active mode. However, Single-active mode is not supported.
- Beginning with Cisco NX-OS Release 10.4(1)F, coexistence with switches that support ESI multi-homing in Single-active mode is introduced for Cisco Nexus 9300-FX/FX2/FX3/GX/GX2 switches and 9500 switches with X97160YC-EX, 9700-FX/GX line cards.
- Beginning with Cisco NX-OS Release 10.4(2)F, coexistence with switches that support ESI multi-homing in both All-active and Single-active modes is available also for Cisco Nexus 9332D-H2R and 93400LD-H1 switches.
- Beginning with Cisco NX-OS Release 10.5(2)F, coexistence with switches that support ESI multi-homing in both All-active and Single-active modes is available also for Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.
- Beginning with Cisco NX-OS Release 10.4(3)F, coexistence with switches that support ESI multi-homing in both All-active and Single-active modes is available also for Cisco Nexus 9364C-H1 switches.
- The Cisco NX-OS devices as remote node accepts MAC route from ESI active node, and EAD-ES and EAD-EVI routes from both ESI Active and Standby nodes. Using these routes, Cisco NX-OS devices calculates the primary and backup paths for a given endpoint's MAC or IP address. In steady state L2 traffic will be forwarded using primary path and in case of primary failure, traffic will be switched to backup path.
- Beginning with Cisco NX-OS Release 10.5(2)F, EVPN multi-homing is supported on Cisco Nexus 9500 Series switches with N9K-X9736C-FX3 line card.
- Maintenance mode (GIR) on ESI only supports custom profiles to bring down uplinks.
- You must configure maximum-path under the EVPN address-family on the Cisco NX-OS nodes. This enables BGP to select multi-paths for EAD per ES, EAD per EVI routes. In particular, Remote VTEPs require this configuration for ESI-RX functionality, which allows the fabric to support multi-homing on the ESI-VTEPs.

Example of EVPN Multi-Homing Using ESI

Example of EVPN Route Type

Figure 78: ESI Single-Active Multihoming



In this topology, the Leaf 3 is a Cisco Nexus 9000 device which acts as remote VTEP to Cat9k (Leaf1, Leaf2) devices that support ESI multi-homing connectivity to local devices. It has the following capabilities:

- Accepts the MAC, EAD per ES, EAD per EVI routes from ESI-active node and EAD per ES, EAD per EVI routes from ESI-standby node(s).
- Defines whether the ESI is single-active based on flag set in EAD per ES routes.
- Defines whether the ESI single-active is two-way attached or n-way attached based on EAD per ES and EAD per EVI received from how many nodes.

The following example shows sample output from Leaf 3 device for the BGP L2 EVPN Route-Type-1 (EAD/ES or EAD/EVI), You must configure **maximum-path** under the EVPN address-family on the Cisco Nexus 9000 nodes. This enables BGP to select all the paths as best-path or multi-paths for EAD per ES, EAD per EVI routes and download all next-hops to L2RIB.

```
show bgp l2vpn evpn route-type 1
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 51.51.51.51:3907 (EAD-ES [03de.affe.ed00.0b00.0000 3907])
BGP routing table entry for [1]:[03de.affe.ed00.0b00.0000]:[0xffffffff]/152, version 71
Paths: (1 available, best #1)
Flags: (0x000002) (high32 00000000) on xmit-list, is not in l2rib/evpn

Advertised path-id 1
Path type: local, path is valid, is best path, no labeled nexthop, has esi_gw
AS-Path: NONE, path locally originated
51.51.51.51 (metric 0) from 0.0.0.0 (51.51.51.51)
Origin IGP, MED not set, localpref 100, weight 32768
Received label 0
Extcommunity: RT:12000:1000002 RT:12000:1000003 RT:12000:1000012
RT:12000:1000013 ENCAP:8 ESI:1:000000

Path-id 1 advertised to peers:
111.111.46.1 111.111.47.1
```

In `ESI:1:000000 –> 1` field, the value indicates the mode, where 1 represent **single-active** and 0 represents **all-active**.

Example of Single-Active MAC entries

The following example shows sample output from Leaf 3 device for the MAC address table command which is enhanced to display single-active MAC entries.

In case of Single Active ESI MAC entries, the **Ports** value displays two VTEPs where **A** represents Active ESI Path and **S** represents Standby ESI Path.

For example: nve1(**A**:11.11.11.11 **S**:22.22.22.22)

```
switch# show mac address-table
Legend:
        * - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC
        age - seconds since last seen,+ - primary entry using vPC Peer-Link,
        (T) - True, (F) - False, C - ControlPlane MAC, ~ - vsan,
        (NA)- Not Applicable, A – Active ESI Path, S – Standby ESI Path
  VLAN     MAC Address      Type    age  Secure NTFY Ports
-------+-----------------+-------+-----+------+----+------------------
C  100    0000.6666.6661  dynamic  NA    F      F   nve1(A:11.11.11.11 S:22.22.22.22)
C  101    0000.6666.6662  dynamic  NA    F      F   nve1(A:11.11.11.11 S:22.22.22.22)
C  101    0000.6666.6663  dynamic  NA    F      F   nve1(A:11.11.11.11 S:22.22.22.22)
C  102    0000.6666.6664  dynamic  NA    F      F   nve1(A:22.22.22.22 S:11.11.11.11)
C  103    0000.6666.6665  dynamic  NA    F      F   nve1(33.33.33.33 44.44.44.44)
C  104    0000.6666.6666  dynamic  NA    F      F   nve1(33.33.33.33 44.44.44.44)
C  105    0000.6666.6667  dynamic  NA    F      F   nve1(33.33.33.33 44.44.44.44)
G    -    0091.f3e7.1b08  static   -     F      F   sup-eth1(R)
switch#
```

Example of L2 Route Path List

Beginning with Cisco NX-OS Release 10.5(3)F, the labeled nexthop and an asymmetric VNI flag is added as shown. For symmetric VNI, labels and flags will not be shown as part of nexthops for EAD and PL.

```
switch# show l2route evpn path-list all detail
(R) = Remote Global EAD NH Peerid resolved,
(UR) = Remote Global EAD NH Peerid unresolved
Flags - (A):All-Active (Si):Single-Active

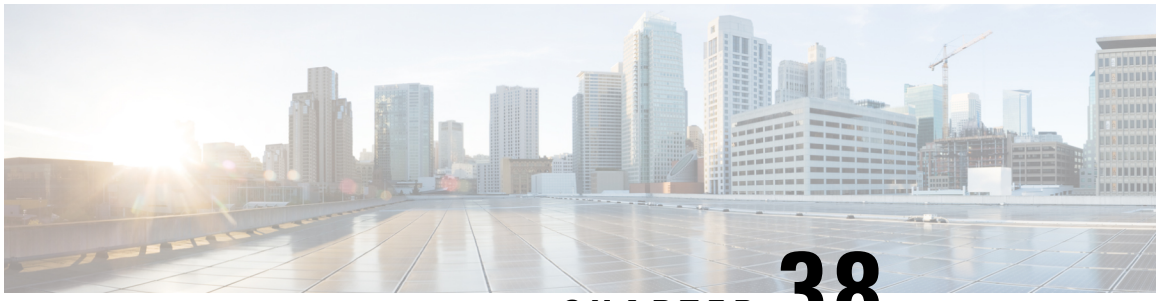
NH Flags: Asy = Asymmetric VNI
Topology ID  Prod   ESI                        ECMP Label        Flags  Client Ctx  MACs
    Sent To
------------ ------ -------------------------- ---------- ------ ----------- ----------
   ----------
 100          None   aaaa.aaaa.aaaa.aaaa.99aa  0          A             0           1
     UFDM
                    CP Next-Hops: 5.5.5.5, 6.6.6.6
                    Gbl EAD Next-Hops: 5.5.5.5 (5,R)
                                       6.6.6.6 (6,R)
                    Res Next-Hops: 5.5.5.5 (Label: 20000)(Flags: Asy)
                                   6.6.6.6 (Label: 10000)(Flags: Asy)
                    Bkp Next-Hops:
                    Res Next-Hops from UFDM: 5.5.5.5 (Label: 20000)(Flags: Asy)
                                             6.6.6.6 (Label: 10000)(Flags: Asy)
```

Example of L2 Route EVPN EAD

Beginning with Cisco NX-OS Release 10.5(3)F, the labeled nexthop and an asymmetric VNI flag is added as shown. For symmetric VNI, labels and flags will not be shown as part of nexthops for EAD and PL.

```
switch# show l2route evpn ead all detail
Flags -(A):All-Active (Si):Single-Active (V):Virtual ESI (D):Del Pending
(S):Stale
```

```
Topology ID   Prod   ESI                       Sent To    Num PLs  Flags
------------- ------ ------------------------- ---------- -------- ----------
100           BGP    aaaa.aaaa.aaaa.aaaa.99aa   -          1       A
                Next-Hops: 5.5.5.5 (Label: 20000)(Flags: Asy)
4294967294    BGP    aaaa.aaaa.aaaa.aaaa.99aa   -          1       A
                Next-Hops: 5.5.5.5
                           6.6.6.6
```



CHAPTER 38

EVPN Ethernet Segment Identifier Multi-Homing

Ethernet Segment Identifier (ESI) multi-homing represents a more modern and scalable approach, leveraging the industry-standard BGP Ethernet VPN (EVPN) control plane. Unlike vPC, ESI multi-homing does not require a dedicated peer-link. Instead, coordination and state synchronization (for example, MAC or ARP) between participating switches (VTEPs) occur over the VXLAN fabric using BGP-EVPN signaling. A key advantage of ESI multi-homing is its ability to support redundancy with more than two switches, offering up to four-way redundancy for a single host connection, making it suitable for larger, more distributed environments.

- EVPN ESI multi-homing, on page 783
- IGMP or MLD Snooping with ESI - EVPN Type-7 and Type-8 route, on page 852
- ECMP reuse, on page 860
- Layer 2 Gateway Spanning Tree Protocol, on page 861

EVPN ESI multi-homing

This section describes how to configure the EVPN ESI multi-homing on Cisco NX-OS devices.

EVPN ESI multi-homing solutions

EVPN ESI multi-homing solutions are standardized EVPN multi-homing technologies that:

- use BGP EVPN-based communication between ESI peers to synchronize data plane state,
- provide redundancy similar to vPC and vPC fabric peering, and
- support 2-way, 3-way, and 4-way multi-homing.

Note vPC and vPC fabric peering only support 2-way multi-homing.

The main functions supported by ESI multi-homing include:

- ESI multi-homing with Primary IP (PIP) or Unicast next-hop,
- ESI multi-homing with Virtual IP (VIP) or Anycast next-hop,
- Butterfly topology,

- IGMP or MLD Snooping (Type-7 and Type-8), and
- ECMP optimization.

The following sections provide detailed explanations of the ESI multi-homing functions.

EVPN multi-homing terms and definitions

This table defines key EVPN multi-homing terms.

Table 14: EVPN multi-homing terms and definitions

Term	Description
EVI	An EVPN instance represented by the VNI.
MAC-VRF	A container housing virtual forwarding tables for MAC addresses. Each MAC-VRF can have a unique route distinguisher and import or export target.
ES	An Ethernet Segment composed of a set of bundled links.
EAD/ES	An Ethernet Auto Discovery route per ES (Type-1), used for fast traffic convergence during access failures. It uses Ethernet Tag value 0xFFFFFFFF.
EAD/EVI	An Ethernet Auto Discovery route per EVI (Type-1), used for aliasing and load balancing. It cannot use Ethernet Tag value 0xFFFFFFFF.
Aliasing	A method for load balancing traffic to all switches connected to an Ethernet Segment using the Type-1 EAD/EVI route, regardless of the switch where hosts are learned.
Mass withdrawal	A mechanism for fast convergence during access failure using the Type-1 EAD/ES route.
Designated Forwarder (DF) election	A process to prevent forwarding loops and duplicates by allowing only a single switch to decapsulate and forward traffic for an Ethernet Segment.
Split horizon	A method to prevent forwarding loops between VTEPs on the same Ethernet Segment and avoid duplicate BUM traffic.
Ethernet Segment Identifier (ESI)	A 10-byte value assigned to each switch under a bundled link shared with a multi-homed neighbor. It is used for DF election for BUM traffic via a Type-4 route.

Term	Description
LACP bundling	LACP is used to detect ESI misconfiguration in multi-homed port-channel bundles. LACP sends the configured ESI MAC address to the access switch. Cisco recommends running LACP to detect and act on misconfigured ES IDs.

EVPN multi-homing information

The EVPN overlay draft introduces changes to the BGP MPLS-based EVPN solution. These changes allow the solution to operate as a network virtualization overlay using VXLAN encapsulation. In BGP MPLS EVPN, the Provider Edge (PE) node acts as the VTEP or Network Virtualization Edge device (NVE). VTEPs use control plane learning and distribute remote addresses through BGP, instead of relying on data plane learning.

The supported route types:

- Type-1: Ethernet Auto-Discovery (EAD) route
- Type-2: MAC advertisement route
- Type-3: Inclusive multicast route
- Type-4: Ethernet Segment route
- Type-5: IP prefix route
- Type-6: Selective Multicast (SMET) route

  For more information, see the Optimized Layer 2 Overlay Multicast, on page 445 chapter.
- Type-7: Multicast membership report synch route

  For more information, see the IGMP or MLD Snooping with ESI, on page 853 section.
- Type-8: Multicast leave synch route

  For more information, see the IGMP or MLD Snooping with ESI, on page 853 section.

BGP EVPN running on Cisco NX-OS uses route Type-2 to advertise MAC and IP (host) information. Route Type-3 specifically carries VTEP information for ingress replication. EVPN route Type-5 allows the advertisements of IPv4 or IPv6 prefixes in a Network Layer Reachability Information (NLRI) without MAC addresses in the route key.

With EVPN multi-homing, Cisco NX-OS software uses the EAD route. In this route, the Ethernet Segment Identifier and the Ethernet Tag ID are part of the prefix in the NLRI.

The BGP control plane learns the reachability of end points. As a result, network convergence time depends on how many MAC or IP routes the VTEP must withdraw during a failure scenario. To handle such conditions, each VTEP advertises one or more Ethernet Auto-Discovery per ES routes for every locally attached Ethernet Segment. If an attached segment fails, the VTEP withdraws the corresponding set of routes. For information on fast convergence, see the Layer 2 fast reroute, on page 792 section.

Cisco NX-OS software also uses the Ethernet Segment Route with EVPN multi-homing, primarily for DF election for BUM traffic. If the Ethernet Segment is multi-homed, the presence of multiple DFs could cause forwarding loops and packet duplication. Therefore, the Ethernet Segment Route (Type-4) is used to elect the DF and apply split horizon filtering. All VTEPs or PEs that are configured with an Ethernet Segment originate this route.

ESI multi-homing operation modes

ESI multi-homing supports the following operation modes:

- **PIP mode**: Use unique loopback addresses on each ESI-VTEP to identify the reachability of a multi-homed host.
- **VIP mode**: Use a common Anycast IP address (VIP), configured across all ESI peer VTEPs, to identify the reachability of a multi-homed host.

ESI multi-homing PIP modes

ESI multi-homing PIP modes are operational modes in VXLAN EVPN fabrics that:

- enable a host to be multi-homed to multiple ESI peer VTEPs,
- identify the reachability of a multi-homed host using unique loopback addresses (PIPs) on individual ESI-VTEPs, and
- operate without requiring a common or Anycast IP address across ESI peers in a PIP setup.

How ESI multi-homing PIP modes work

- **PIP assignment**: Each VTEP that is part of an ESI cluster is assigned a distinct loopback address, referred to as its PIP.
- **Host reachability advertisement**: A multi-homed host connects to multiple ESI peer VTEPs. The BGP EVPN control plane advertises the host's reachability through the individual PIP addresses of all connected ESI peer VTEPs to other remote VTEPs in the VXLAN fabric.
- **Traffic forwarding**: When a remote VTEP needs to send traffic to the multi-homed host, it performs a hashing operation to select one of the available ESI peer VTEPs (identified by their unique PIPs) for Equal-Cost Multi-Path (ECMP) forwarding. The selected PIP then becomes the destination IP in the outer VXLAN header of the encapsulated traffic.
- **Fault tolerance and ECMP adjustment**: If an ESI goes down on a particular ESI switch, the BGP EVPN control plane signals this status change throughout the VXLAN fabric. Remote VTEPs update their forwarding tables and remove the PIP of the failed VTEP from the list of available paths for the multi-homed host. Subsequent traffic to the host is then re-balanced using ECMP across the remaining active ESI peers. If traffic lands on a VTEP where ESI is down, the Layer 2 Fast Reroute (L2FRR) mechanism forwards the traffic to a VTEP where ESI is up. For more information on L2FRR, see the Layer 2 fast reroute, on page 792 section.



Note

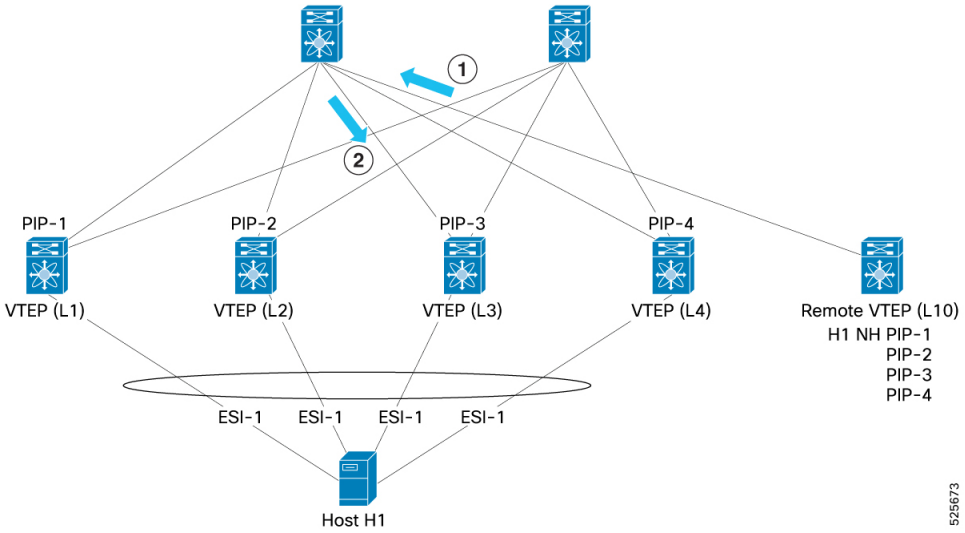
- A switch can operate in either VIP mode or PIP mode, but not both simultaneously.
- Ensure that all ESI peers operate in the same mode to avoid inconsistencies.

Examples

Consider four ESI peer VTEPs (L1, L2, L3, and L4), with their respective unique PIP addresses (PIP-1, PIP-2, PIP-3, and PIP-4). A host (H1), is multi-homed to all four of these VTEPs. A remote VTEP (L10), learns that H1 is reachable through PIP-1, PIP-2, PIP-3, and PIP-4. When L10 sends traffic to H1, it uses hashing to

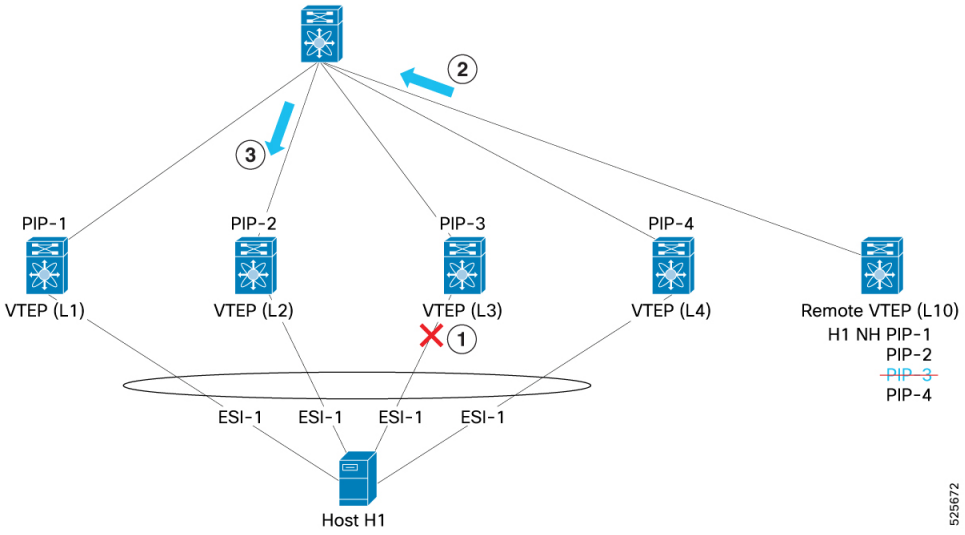
select one of these PIPs (PIP-3) and encapsulates the traffic, setting PIP-3 as the outer VXLAN destination IP. If the ESI on L3 experiences failure, the BGP EVPN control plane informs L10. L10 then removes PIP-3 from its list of paths for H1 and subsequently distributes traffic to H1 using ECMP, selecting from PIP-1, PIP-2, and PIP-4, in this example PIP-2 is selected.

Figure 79: Steady state scenario



1. L10 sends packet to Host with destination IP, PIP-3
2. Spine forwards packet to L3

Figure 80: Failure scenario



1. ESI goes down on L3
2. L10 makes adjustment to send traffic to PIP-2 (and not PIP-3)
3. Spine forwards packet to L2

ESI multi-homing VIP modes

ESI multi-homing VIP modes are operational modes in VXLAN EVPN fabric that:

- identify the reachability of a multi-homed host through a common Anycast IP address (VIP) configured across all participating ESI peer VTEPs,
- provide a highly resilient and scalable solution for host connectivity in VXLAN EVPN fabrics, and
- simplify the forwarding logic for remote VTEPs by presenting a single, stable IP address for multi-homed hosts. This approach reduces control plane churn across the fabric.

How ESI multi-homing VIP modes work

- **VIP assignment**: Each ESI-VTEP is configured with a unique PIP. Additionally, all ESI-VTEPs within a given ESI cluster are configured with the same VIP.
- **Host reachability advertisement**: A multi-homed host connects to multiple ESI peer VTEPs. The BGP EVPN control plane advertises the host's reachability to remote VTEPs. Critically, it advertises reachability via the common VIP rather than the individual PIPs.
- **Traffic forwarding**: When a remote VTEP needs to send traffic to the multi-homed host, it encapsulates the traffic with the VIP as the destination IP in the outer VXLAN header. The remote VTEP does not need to know the individual PIPs of the ESI peers.

  The spine switches or other intermediate devices in the VXLAN fabric are aware that the VIP is reachable through multiple VTEPs (the ESI peer VTEPs). The spine switch performs an ECMP hash to select one of the available ESI peer VTEPs and forwards the encapsulated traffic to that selected VTEP.

  Once the traffic reaches the selected ESI peer VTEP, the VTEP forwards it to the multi-homed host.
- **Fault tolerance**: If an ESI goes down on a particular ESI switch, the BGP EVPN control plane signals this information. However, the remote VTEP continues to send traffic to the host using the VIP. No changes occur to forwarding at the spine. If the traffic lands on the VTEP whose ESI is down, the Layer 2 Fast Reroute (L2FRR) mechanism forwards the traffic to one of the VTEPs where ESI is UP. For more information on L2FRR, see the Layer 2 fast reroute, on page 792 section.
- **Impact on VIP or PIP switching**: The status of ESIs is signaled via BGP-EVPN. Remote VTEPs use this information to dynamically switch between VIP forwarding mode (when two or more multi-homed ESI-VTEPs are up) and PIP forwarding mode (when only a single ESI-VTEP is active), ensuring continuous traffic flow despite changes in ESI availability.

  When the ESI-VTEPs are configured in VIP (Anycast) mode, the Remote VTEP can operate with VIP next-hop or with PIP next-hop, depending on how many ESI-VTEPs are currently active for the multi-homed host.

  Remote VTEPs in the VXLAN fabric track the number of VTEPs that are active for a particular ESI based on the BGP-EVPN signaling across the fabric. For example, when an ESI is down on a particular ESI-VTEP, route Type-1 , such as a BGP EVPN Route-Update, is signaled across the VXLAN fabric.

  If more than one ESI-VTEP is active, a Remote VTEP using VIP next-hop operates more efficiently. When only one ESI-VTEP is active, using VIP next-hop increases the chance that the flow reaches an ESI-VTEP whose ESI is down. In this case, traffic is rerouted to an ESI-active, which increases latency.

  Thus, the remote VTEP operates primarily in VIP next-hop mode and occasionally in PIP next-hop mode, depending on the status of the ESI-VTEPs. The L2FRR mechanism on the ESI-VTEPs ensures that traffic

is forwarded to one of the VTEPs during transitions between PIP and VIP modes. For more information on L2FRR, see the Layer 2 fast reroute, on page 792 section.

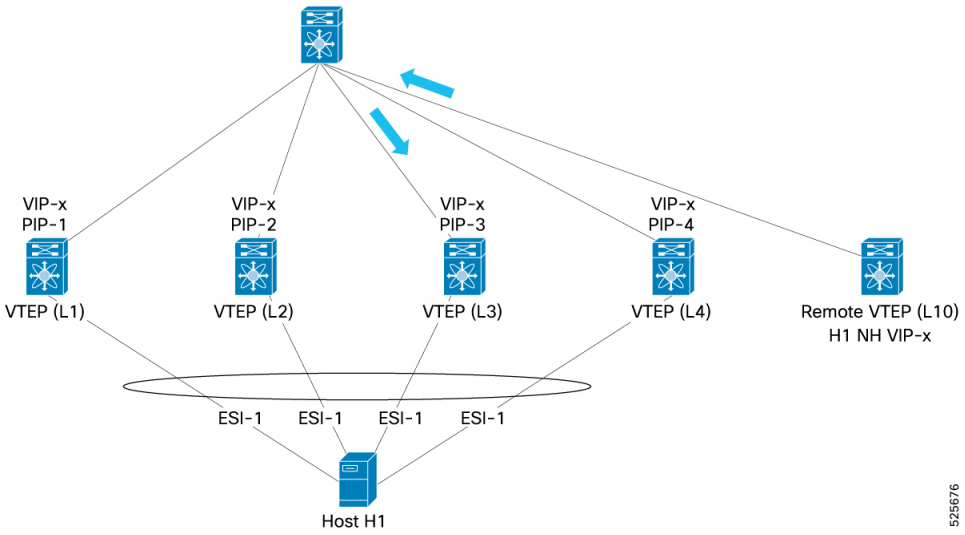
Note

- A switch can operate in either VIP mode or PIP mode, but not both simultaneously.
- Ensure all ESI peers operate in the same mode to avoid inconsistencies.
- VIP or Anycast next-hop is specified in draft evpn-anycast-aliasing.

Examples

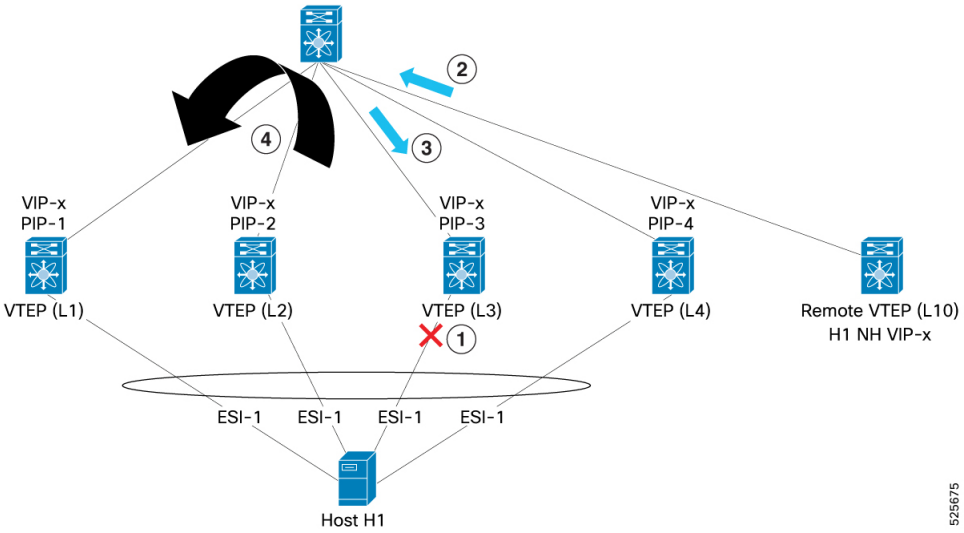
Consider four ESI peer (L1, L2, L3, and L4), with individual PIPs (PIP-1, PIP-2, PIP-3, PIP-4), and a common VIP (VIP-x). A host (H1), is multi-homed to all four VTEPs. A remote VTEP (L10), learns that H1 is reachable through VIP-x. L10 encapsulates traffic to H1 with VIP-x as the outer VXLAN destination IP. The spine switch, knowing that VIP-x is reachable via L1, L2, L3, and L4 uses ECMP to select one (for example, L3) and forwards the traffic to L3. If the ESI on L3 fails, BGP-EVPN route Type 1 updates are sent to the fabric. L10 continues to send traffic to VIP-x, and the spine continues to ECMP the traffic to L1, L2, L3, and L4. If traffic lands on L3, then L2FRR sends the traffic to any one of the remaining VTEPs without requiring L10 to update its forwarding state. In this example L1 is chosen.

Figure 81: Steady state scenario



1. L10 sends packet to Host with Dest-IP VIP-x
2. Spine selects L3 for VIP-x and forwards to L3

Figure 82: Failure scenario

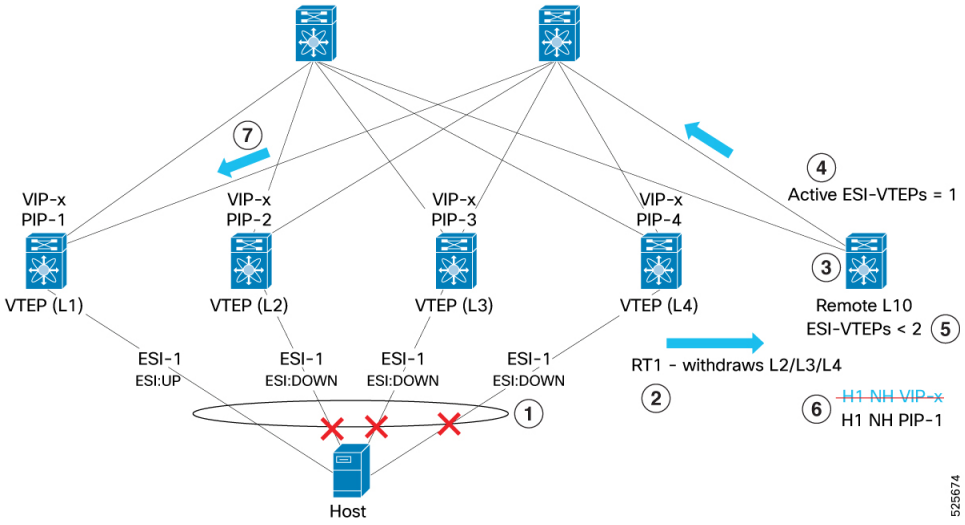


1. ESI goes down on L3
2. L10 continues to send with VIP-x
3. Traffic reaches L3 where ESI is down
4. L3 uses L2FRR to redirect the traffic to L1 where ESI-1 is still up.

Example for VIP to PIP transition

1. Multiple ESI links (L2, L3, and L4) go down, causing several ESI failures to be detected.
2. BGP-EVPN Route Type 1 (RT-1) withdrawals are signaled across the fabric to indicate these changes.
3. Remote VTEP L10 receives updates about the ESI status changes.
4. As a result, the number of active ESI-VTEPs drops to one—only L1 remains active.
5. L10 detects that fewer than two ESI-VTEPs are active and triggers a transition from VIP-next-hop mode to PIP-next-hop mode.
6. The next-hop is updated from the shared VIP-x to the individual IP address of L1 (PIP-1).
7. All traffic is now sent directly to L1 without ECMP distribution among multiple VTEPs.
8. Since there is only one active ESI-VTEP, no fast reroute is needed; traffic reaches the remaining active VTEP (L1) directly.

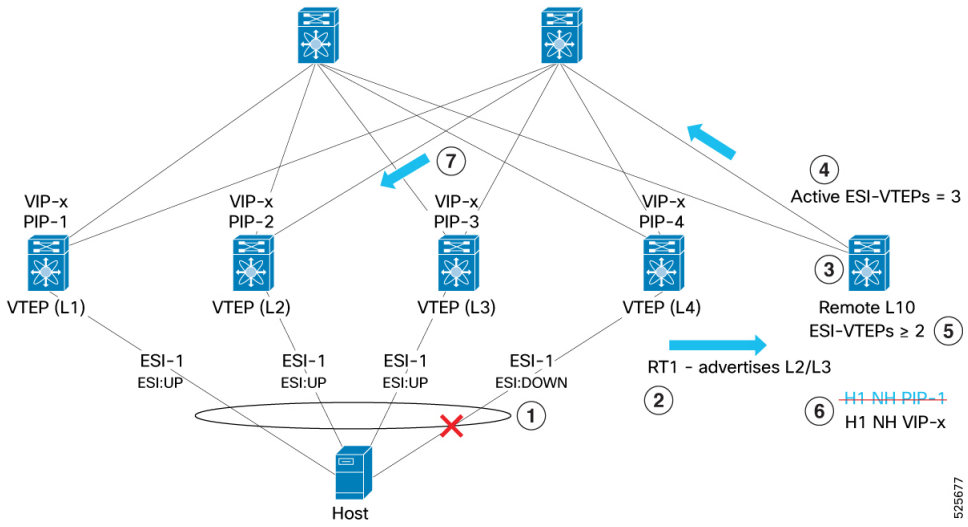
Figure 83: VIP to PIP transition



Example for PIP to VIP transition

1. ESI recovery occurs as the L2 and L3 ESI links come back online.
2. BGP-EVPN Route Type 1 (RT1) advertisements are signaled across the fabric to announce the recovery.
3. Remote VTEP L10 receives updates about the restored ESI status.
4. The count of active ESI-VTEPs increases to three, where L1, L2, and L3 are all active.
5. L10 detects that two or more ESI-VTEPs are now active and initiates a transition from PIP-next-hop mode to VIP-next-hop mode.
6. The next-hop is updated from the individual IP address of L1 (PIP-1) to the shared anycast IP (VIP-x).
7. With ECMP load balancing enabled, traffic is distributed across L1, L2, and L3.
8. As a result, bandwidth utilization is optimized through even traffic distribution.

Figure 84: PIP to VIP transition



Layer 2 fast reroute

Layer 2 Fast Reroute (L2FRR) is a network protection mechanism that"

- minimizes traffic disruption and prevents traffic from being discarded during failure conditions of an ESI connection,
- ensures that traffic destined for a multi-homed host, that might still arrive at a failed ESI switch, is redirected to an active ESI peer within the same cluster, and
- enhances network resilience by maintaining traffic flow during convergence or when remote VTEPs are not yet updated.

How L2FRR work

- **General principle**: When an ESI connection to a multi-homed host fails on a specific ESI switch, the L2FRR functionality is activated on that failing switch. The switch internally redirects any traffic that still reaches it for the affected multi-homed host to another healthy ESI peer switch in the same multi-homing cluster.

  Assign a unique identifier to the fast reroute packet. This allows you to handle fast reroute packets in specific ways. For example, if a packet has already been fast rerouted once, it should not be fast rerouted again to another ESI-VTEP in case of double failures. To ensure unique identification, assign a special Source IP, called the fast reroute anycast source IP, to the fast reroute packet. For detailed configuration steps, see the Enable EVPN ESI multi-homing, on page 804 section.

Note Ensure that you configure the fast reroute anycast source IP address uniquely on all the switches that are part of the same ESI multi-homing cluster.

- **In PIP mode**:

- **Failure event**: If an ESI connection goes down on an ESI switch (for example, L3), the BGP-EVPN control plane signals this event across the VXLAN fabric.
- **Convergence delay**: There is a brief period during which remote VTEPs (for example, L10) are converging and updating their forwarding tables to remove the failed VTEP (L3) from the list of reachable paths for the multi-homed host.
- **L2FRR action**: During this convergence period, traffic from L10 might still be directed to L3. Upon arrival at L3, the L2FRR functionality detects the ESI connection failure and internally redirects this traffic to one of the remaining active ESI peers (for example, L1, L2, or L4).

  For each ESI, a reroute peer is elected in a round-robin fashion from the active peer list whenever an ESI peer comes up. This approach minimizes oversubscription in scenarios where multiple ESIs are down and require fast rerouting.
- **Outcome**: This prevents traffic from being discarded while the network converges, ensuring continuous service until remote VTEPs fully update their forwarding paths to bypass the failed VTEP.

For more information, see the Examples, on page 786.

- **In VIP mode**:
  - **Failure event**: If an ESI connection goes down on a particular ESI switch (for example, L3), the BGP-EVPN control plane signals this event.
  - **Remote VTEP behavior**: In VIP-mode, remote VTEPs (for example, L10) always send traffic to the common VIP for the multi-homed host, regardless of individual ESI connection status. The spine switch is responsible for ECMP'ing traffic to the underlying active VTEPs for that VIP.
  - **L2FRR action**: Even though ESI is down on L3, the spine switch will still have the VIP reachable via L3, and traffic may still be directed to L3 (for example, due to hashing or specific network conditions). The L2FRR functionality on L3 intercepts any traffic destined for the multi-homed host and redirects it to one of the remaining active ESI peers (for example, L1, L2, or L4).
  - **Outcome**: This ensures that traffic is never discarded at the failed ESI switch, providing robust fault tolerance without requiring any changes or churn on the remote VTEP.

For more information, see Examples, on page 789.

Designated Forwarder elections

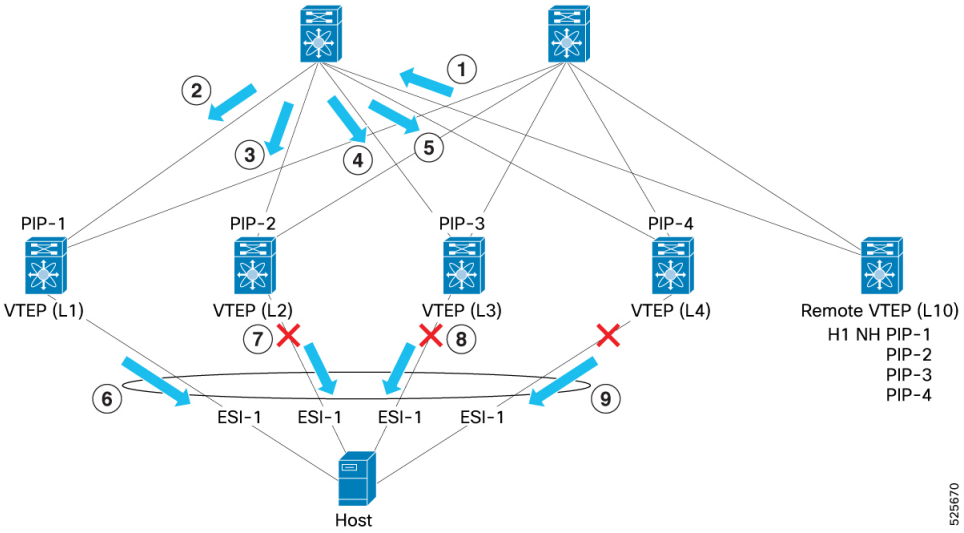
Designated Forwarder (DF) elections are network mechanisms in multi-homing VXLAN EVPN fabrics that:

- determine which VTEP forwards BUM traffic to a shared Ethernet Segment,
- avoid redundant forwarding and network loops, and
- support configurable methods, such as per-flow hashing and modulo-based selection, to optimize traffic handling and scalability.

How DF elections works

DF elections are crucial in multi-homing environments to prevent duplicate forwarding of BUM traffic. DF elections select one VTEP from those connected to an Ethernet Segment to forward such traffic. Other VTEPs do not forward BUM traffic for that segment.

Figure 85: DF election traffic flows



(1): VTEP-L10 sends broadcast packet (for example, ARP)

(2), (3), (4), and (5): Spine sends to all VTEPs (L1-L4)

In this example, the fabric uses a multicast underlay. If the underlay uses IR, the Remote VTEP (L10) sends a separate copy of the packet to each ESI-VTEP. In both cases, broadcast data packet reaches all the ESI-VTEPs.

(6): L1 is the DF and sends packet to Host

(7), (8), and (9): L2, L3, and L4 are non-DFs and do not forward the packet to Host

Supported DF election methods

Per-flow (Packet Hash-based) method (default)

- **Mechanism**: This is the default method. The DF is determined based on hashing the fields of each incoming packet. Each ESI peer independently performs this hash.

  The DF election algorithm per-flow uses the Ethernet Segment route type (Type-4).

  The BGP route Type-4 helps you determine the number of ESI-VTEPs for an ESI and their respective loopback IPs that are associated with each ESI. Based on this information from route Type-4, the hash result is divided equally among the ESI peer VTEPs.

  After calculating the packet hash, each VTEP determines whether it must forward the packet to the multi-homed host or if another ESI peer is responsible.

  All VTEPs participating in the ESI peer cluster must use the same per-flow DF. If any VTEPs do not use per-flow DF, the cluster uses modulo-based DF election.

- **Traffic handling**:
  - Each ESI peer receives BUM traffic, which is based on PIP, from the remote VTEP, regardless of the underlay replication mode (ingress replication or multicast).
  - Upon receiving the traffic, each ESI peer calculates the packet hash.
  - Based on the hash result, one peer is designated as the forwarder for that specific flow and sends the traffic to the multi-homed host.

- Other peers not designated for that flow drop the traffic towards the multi-homed host.
- If traffic is received from an ESI peer (for example, from another VTEP in the same ESI cluster), the ESI ports are pruned to prevent loops.

- **Scalability**: This method offers excellent scalability for L2VNIs because the DF election is determined per-flow, rather than per-VLAN, and provides better BUM flow distribution across the multi-homed VTEPs.

Modulo method (fallback)

- **Mechanism**: This method utilizes Ethernet Segment routes, which are also employed in Anycast BGWs and multi-homing scenarios. It assigns DF responsibility based on a modulo calculation.

  The BGP route Type-4 helps you determine the number of ESI-VTEPs for an ESI and their respective loopback IPs. Based on this information, the VLANs are distributed equally among the ESI peer-VTEPs. In this way, for each VLAN, only one ESI-VTEP is responsible to forward the traffic to the multi-homed host and is called the DF for that VLAN. This process is the modulo method.

- **Fallback use case**: The modulo method acts as a fallback DF election mechanism if members of the same Ethernet Segment are configured with different DF election modes.
- **Scalability implication**: This method typically requires the use of alternative multicast indices (Alt-mcindex), which are derived from a shared global multicast index pool. This can impact the overall scale of supported VLANs and IGMP or MLD groups on the platform.
- **Traffic handling**: When BUM traffic arrives at an ESI-VTEP and the Modulo Method is used:
  - The ESI-VTEP forwards the BUM traffic to all orphan ports associated with that VLAN, regardless of whether it is the DF for that VLAN.
  - The ESI-VTEP then evaluates whether to send the BUM traffic to the ESI port, which connects to the multi-homed host. If the ESI-VTEP is the DF for that VLAN, it forwards the traffic to the multi-homed host through the ESI port. If it is not the DF, it does not send the traffic to the multi-homed host.

  This process ensures that only one ESI-VTEP sends BUM traffic to the multi-homed host, preventing duplication.

EVPN EAD-EVI routes

EVPN EAD-EVI route is a type of BGP EVPN route that:

- advertises per-EVI information to enable flexible communication between Ethernet segments in an EVPN network,
- supports standards compliance for multi-homed host scenarios, and
- enables interoperability while being disabled by default in Cisco NX-OS implementations unless explicitly enabled.

In the current Cisco NX-OS implementation, when a host is learned on one ESI-VTEP, it is advertised to the peer ESI VTEPs. The peer ESI VTEPs also send out a Re-Origination (RO) to the fabric. This approach ensures that a remote VTEP has all the information needed to route traffic through multiple paths to a multi-homed host. Since RO is mandatory in NX-OS, a remote VTEP does not need to use EAD-EVI

information to determine the ESI VTEPs for a multi-homed host. The EAD-EVI route remains available for standards compliance.

Type-2 route re-origination

A Type-2 route re-origination is an EVPN mechanism that

- enables VTEP peers to originate new copies of peer-synced MAC and MAC-IP routes,
- ensures redundancy for multi-homed hosts in ESI clusters, and
- prevents loss of reachability for remote VTEPs if a local VTEP or Ethernet Segment goes down.

Example

A host (H1) is multi-homed to VTEPs L1, L2, L3, and L4. If H1's MAC and MAC-IP routes are learned locally on L1, then L1 advertises MAC and MAC-IP reachability. As part of the same ESI cluster, L2, L3, and L4 treat those as peer-synced routes and program reachability to the local Ethernet Segment. L2, L3, and L4 also re-originate the MAC and MAC-IP routes and advertise them in EVPN update messages.

A remote VTEP, L10, learns that H1 is reachable through PIP-1, PIP-2, PIP-3, and PIP-4 (in PIP mode), or through a virtual IP (VIP) with 4 paths from L1–L4 (in VIP mode).

If re-origination is not used, remote VTEP L10 receives only a single path (from L1) for H1's reachability. If the ES goes down on L1 and the other VTEPs have not yet learned H1's MAC and MAC-IP routes, IP traffic to H1 is lost until the routes are learned elsewhere. With re-origination, the remaining VTEPs prevent this traffic loss by advertising alternate paths.

Ethernet Segment delay-restore timers

The Ethernet Segment (ES) delay-restore timer is a mechanism for managing the recovery of Ethernet Segments in EVPN that

- activates when an NVE interface flaps (for example, during Anycast IP configuration, core isolation, or device reload), causing all associated ESIs to go down,
- delays the restoration of ESIs, bringing them back online only after the timer expires, and
- prevents traffic from being attracted to the VIP while hardware programming is still in progress, avoiding potential traffic instability during convergence.

The VIP interface remains down until the ES delay-restore timer expires. Once all ESI POs are operational, the VIP interface activates. After the ES delay-restore timer ends, a 480-second VIP anycast delay-restore timer starts. This additional timer ensures that if a PO fails to come online after the ES delay-restore timer expires, the VIP interface will eventually activate.

How ESI multi-homing topologies work

ESI multi-homing topologies are data center network designs that provide redundancy and load balancing through the use of Ethernet Segment Identifiers. These designs support scalable deployments by allowing multiple ESI interfaces in various configurations. Examples include 2-way, 4-way, and butterfly topologies. These designs also support flexible and interoperable network architectures that provide high availability.

The supported ESI multi-homing topologies include:

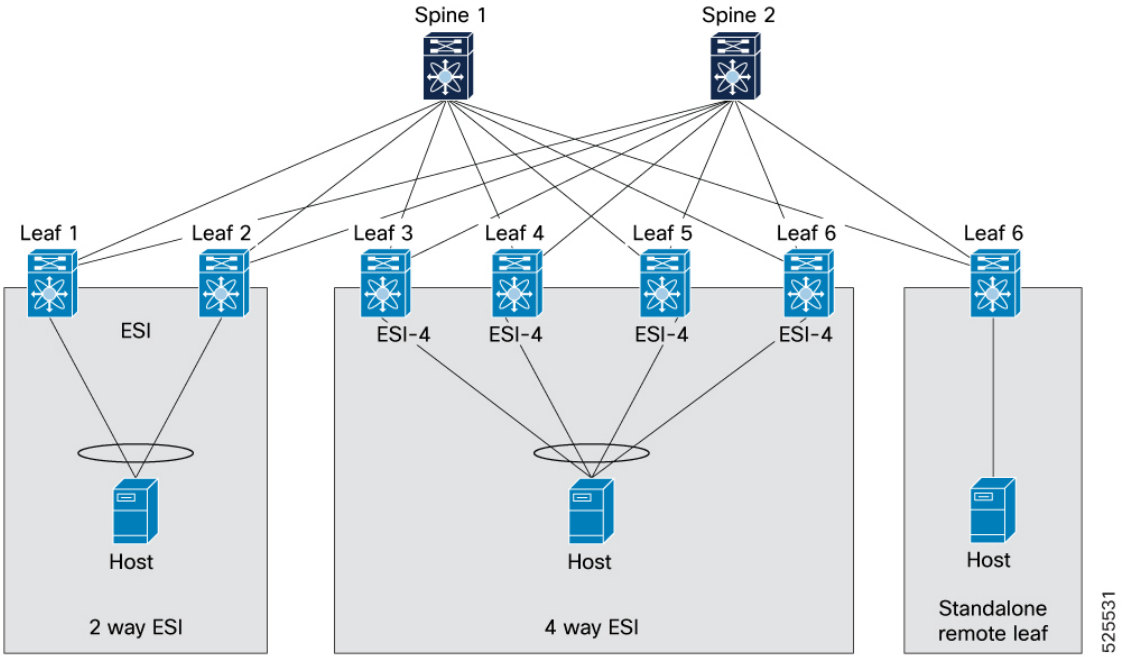
- ESI multi-homing with 2-way, 4-way and standalone remote leaf,
- Butterfly topology, and
- Interoperability with different ESI multi-homing solutions.

The key components involved in the process are:

- **Spine switches (Spine 1 and Spine 2)**: The spine switches interconnect all leaf switches and provide high-speed forwarding between network segments.
- **Leaf switches (Leaf 1-10)**: These switches act as the access layer, connecting directly to end hosts and providing connectivity to the spine layer. Leaf switches participate in ESI groups to support multi-homed hosts.
- **ESI**: A logical construct that groups multiple physical links from a host to different leaf switches, enabling multi-homing for redundancy and load balancing.
- **Hosts**: End devices (for example: servers, virtual machines) that require network connectivity and can be configured for single-homed or multi-homed attachment.
- **2-way ESI configuration**: A setup where a host is multi-homed to two leaf switches, providing dual-path redundancy. This configuration is also called dual homing.
- **4-way ESI configuration**: A setup where a host is multi-homed to four leaf switches, offering enhanced redundancy and greater load-balancing capabilities.
- **Standalone remote leaf (Leaf 7)**: A leaf switch that provides single-homed connectivity to a host, operating independently without multi-homing.
- **vPC or vPC fabric peering multi-homing**: Represents various multi-homing approaches, including physical link aggregation, vPC fabric peering, and vPC for redundant host attachments.

ESI multi-homing topology with 2-way, 4-way and standalone remote leaf

Figure 86: Topology of 2-way ESI, 4-way ESI and standalone remote leaf



ESI is a crucial component in modern data center networks, particularly in EVPN deployments. It provides multi-homing capabilities for end hosts. This mechanism ensures high availability and efficient traffic distribution.

The process involves these stages:

- **Host multi-homing**: A host connects its physical interfaces to two or more leaf switches. This configuration creates redundant physical paths, allowing the host to access the network.
- **ESI group formation**: Leaf switches that are configured with the same ESI for a specific host form a logical group. The leaf switches exchange control plane information to coordinate traffic forwarding for the multi-homed host.
- **Ingress traffic handling**: When the multi-homed host sends traffic, it is received by one of the active leaf switches within its ESI group. The receiving leaf switch then forwards the traffic toward the spine layer for routing to its destination. The receiving leaf switch also forwards the traffic to local interfaces that are part of the ESI group or to orphan interfaces.
- **Egress traffic handling**: When BUM (multi-destination) traffic is destined for a multi-homed host, all leaf switches in the host's ESI group may receive the traffic from the spine layer. The ESI mechanism ensures that only one designated leaf switch forwards traffic to the host and prevents duplicate frames from reaching the host.
- **Load balancing and redundancy**: The ESI mechanism balances traffic to and from the multi-homed host across the active links and leaf switches in the ESI group. If a link or leaf switch fails, the remaining active ESI group members seamlessly take over traffic forwarding to ensure continuous host connectivity.
- **Standalone leaf operation**: For hosts connected to a standalone remote leaf (for example, Leaf 7), traffic forwarding follows a direct, single-path route between the host, the leaf switch, and the spine switches, as these hosts are not part of an ESI multi-homing group.

Result: Ethernet Segments provide high availability, redundancy, and efficient load balancing for your hosts in a spine-leaf network architecture. This ensures robust and scalable data center connectivity.

Butterfly topology

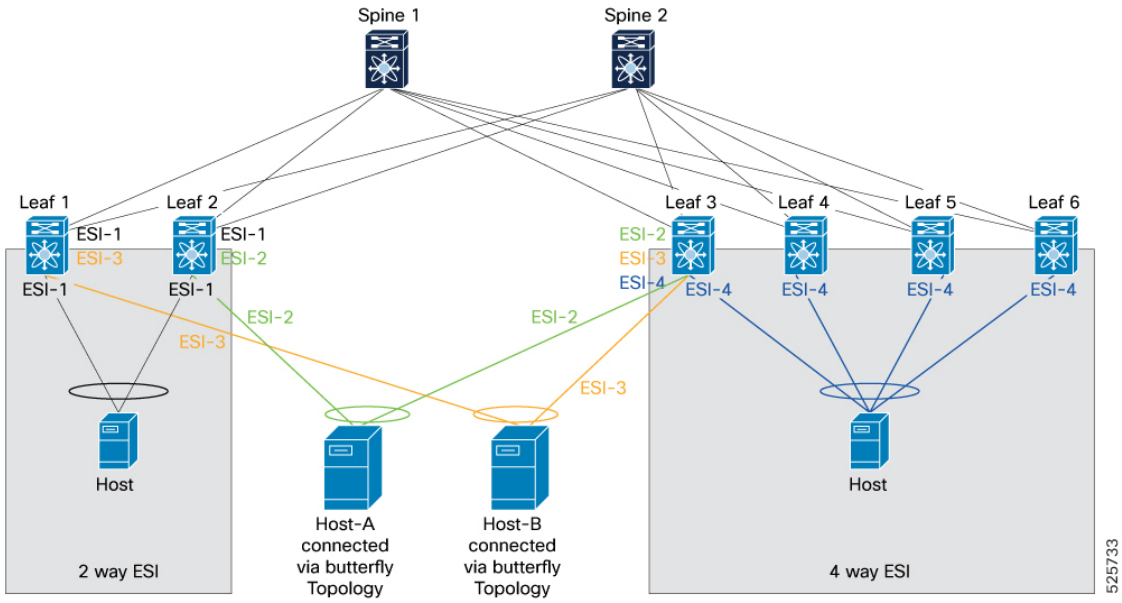
The butterfly topology connects hosts through Ethernet Segments between 2-way and 4-way ESI clusters. This approach allows seamless traffic forwarding between different multi-homing clusters in complex network topologies.

The supported cluster combinations in butterfly topology include:

- Multiple hosts connected via butterfly topology,
- 2-way butterfly topology, and
- 4-way butterfly topology.

Multiple hosts connected via butterfly topology

Figure 87: Butterfly topology of multiple hosts



Key components in butterfly topology are:

- 2-way ESI cluster: Leaf 1 and Leaf 2.
- 4-way ESI cluster: Leaf 3, Leaf 4, Leaf 5, and Leaf 6.

Host connections:

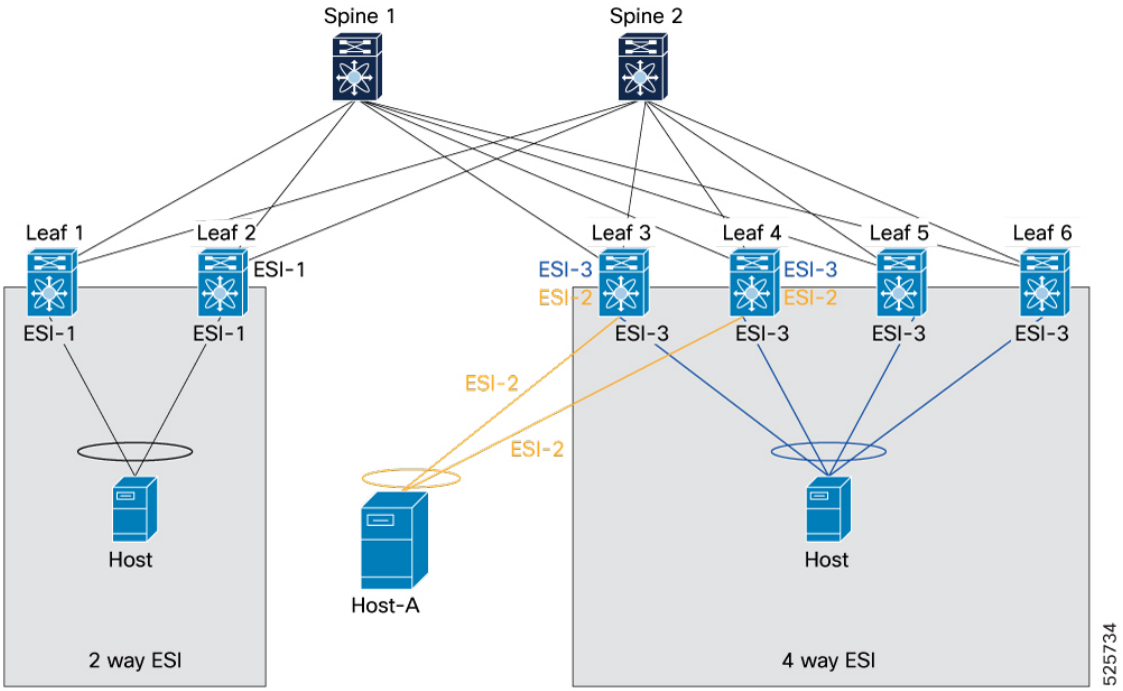
- Host-A is connected to the 2-way ESI cluster (via Leaf 2) using ESI-2 and to the 4-way ESI cluster (via Leaf 3) using ESI-2 as well.
- Host-B is connected to the 2-way ESI cluster (via Leaf 1) using ESI-3 and to the 4-way ESI cluster (via Leaf 3) also using ESI-3 as well.

Result:

- This allows Host 1 to send and receive traffic through both clusters using the ESI (ESI-2).
- Host 2 uses ESI-3 for both clusters, enabling similar redundancy and flexible traffic forwarding.

2-way butterfly topology

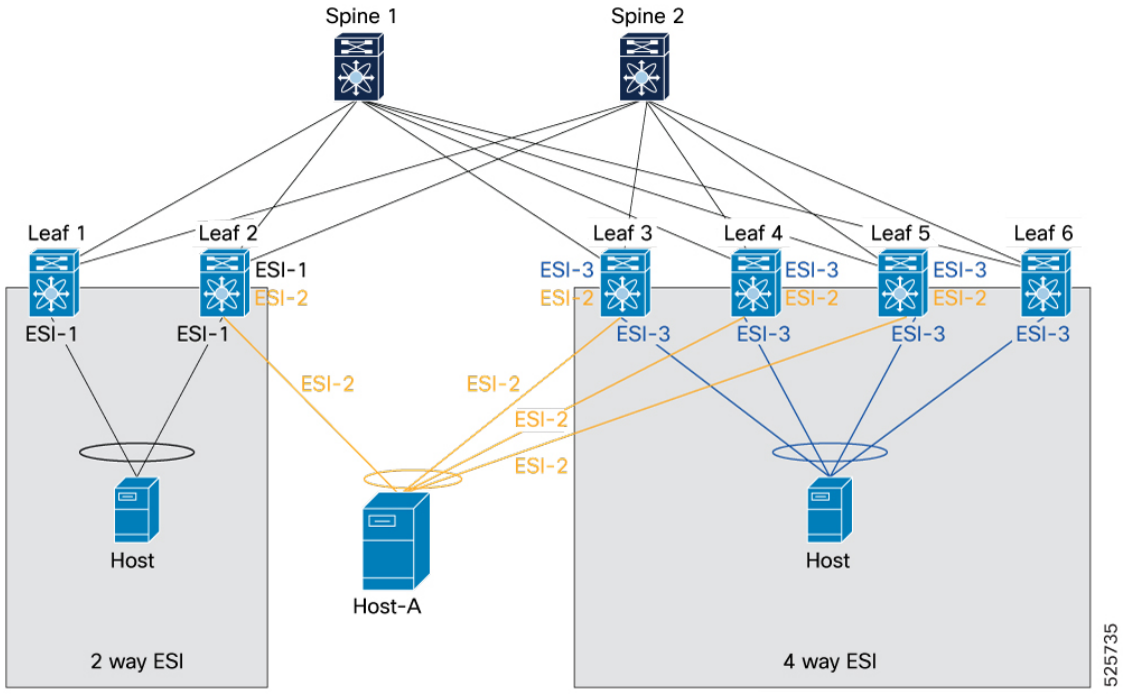
Figure 88: Butterfly topology of 2-way ESI cluster



In this topology, Host-A connects to the 4-way ESI cluster by using ESI-2 through Leaf 3 and Leaf 4. This setup forms a 2-way butterfly topology.

4-way butterfly topology

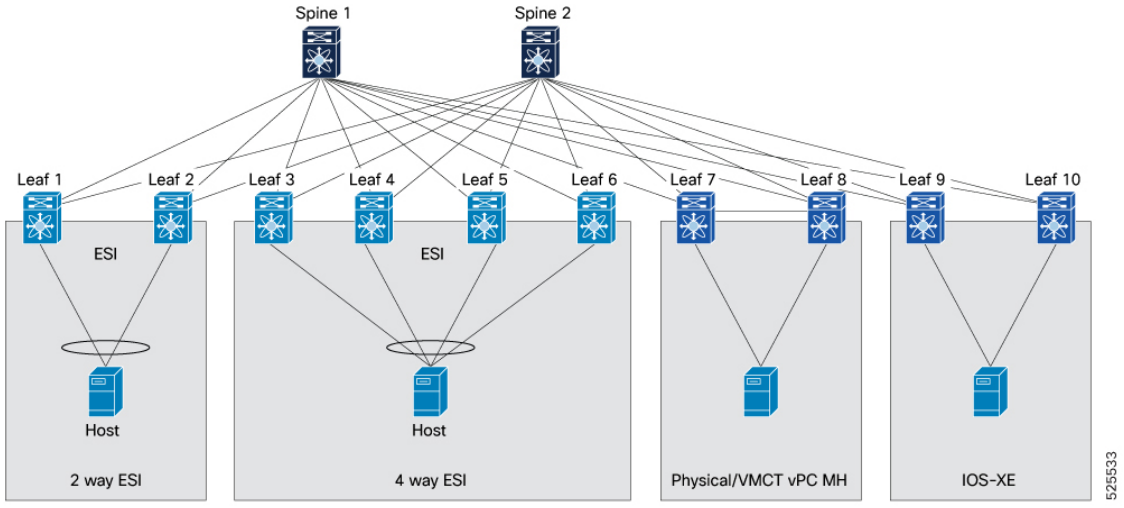
Figure 89: Butterfly topology of 4-way ESI cluster



In this topology, Host-A connects to the 2-way ESI cluster through Leaf 2 and to the 4-way ESI cluster through Leaf 3, Leaf 4, and Leaf 5, using ESI-2. This setup forms a 4-way butterfly topology.

Interoperability with different ESI multi-homing solutions

Figure 90: ESI multi-homing topology with different ESI multi-homing solutions



ESI multi-homing supports interoperability with different solutions, such as vPC and vPC fabric peering, to establish robust and redundant host connectivity. The use of various multi-homing techniques and foundational fabric designs achieves this goal. These methods ensure high availability, efficient traffic distribution, and scalability for connected devices.

The process involves the following stages:

- **vPC or vPC fabric peering multi-homing**: In environments where hosts connect through an intermediate Layer 2 device, leaf switches (for example: Leaf 9 and Leaf 10) provide multi-homed connectivity. This typically involves physical link aggregation or vPC configurations, ensuring redundant paths from the L2 device to the leaf switches and ultimately to the spine fabric.

Result: The implementation of these diverse connectivity methods ensures that hosts within a spine-leaf network achieve high availability, optimal traffic distribution, and robust fault tolerance, adapting to various application demands and infrastructure designs.

Guidelines and limitations for ESI multi-homing

This section outlines the requirements, constraints, and supported behaviors for configuring ESI multi-homing on Cisco NX-OS.

Supported releases and platforms

Table 15: Supported releases and platforms

Release	Platforms
10.6(1)F	Cisco Nexus 9300-FX2/FX3/GX/GX2/H2R/H1 Series switches Cisco Nexus 9500 Series switches with 9700-GX/FX3 line cards **Note** Cisco Nexus 9700-FX line cards support only core links on the ESI fabric.

Supported and unsupported features

Table 16: Supported and unsupported features

Feature	Supported Release	Comments
Multi-site	10.6(1)F	Multi-Homing on BGW is not supported.
DF algorithm	10.6(1)F	Supported
Host mobility	10.6(1)F	Supported
Multicast underlay and ingress replication for IPv4	10.6(1)F	Supported
IGMP and MLD Snooping, Type-7 and 8	10.6(1)F	Supported
Full STP support	10.6(1)F	Supported

Feature	Supported Release	Comments
PIP and VIP NH	10.6(1)F	Supported
VXLAN PBR	10.6(1)F	Supported
N-way ESI	10.6(1)F	4-way ESI is supported, but configurations exceeding 4-way (for example, 5-way or more) are not supported.
Butterfly topology	10.6(1)F	Only PIP mode is supported. VIP mode is not supported.
ESI Type-0 and 3	10.6(1)F	Supported
VXLAN TE	-	Not supported
TRM	-	Not supported
ISSU on EOR	-	Not supported
IPv6 underlay	-	Not supported
DSVNI	-	Not supported

Configuration requirements

- The **maximum-path** must be configured under the EVPN address-family on all participating NX-OS nodes. This allows BGP to select multiple paths for EAD per ES and EAD per EVI routes. Remote VTEPs require this configuration for ESI-RX functionality, supporting multi-homing on ESI-VTEPs.
- vPC or vPC fabric peering can coexist. However, each VTEP operates in either vPC-VTEP or ESI-VTEP mode, not both simultaneously.
- The new L3-VNI must be configured on all ESI peers to extend the VNI range.
- Beginning with Cisco NX-OS Release 10.6(1)F, link-local addresses are advertised on devices with ESI multi-homing enabled. Only devices participating in the given Ethernet Segment accept the route. Devices that are not part of the Ethernet Segment drop the route.
- The fast reroute anycast source IP address must be unique across the fabric to handle the fast reroute packet appropriately.

  For example, if a packet has already undergone fast rerouting once, do not reroute it again to another ESI-VTEP in the event of double failures. Assign a specific source IP address, known as the fast reroute anycast source IP, to reroute the packet and ensure unique identification. For more information about configuration, see the Enable EVPN ESI multi-homing, on page 804 section.
- ISSU is supported for VTEPs part of a multi-homing cluster.
- All multi-homed hosts across ESIs must be configured to the same set of ESI-VTEPs, and the esi-id should be consistent on each relevant port-channel; discrepancies are not flagged by the consistency checker.

- The ESI Multi-homing feature requires TCAM carving of the region **ing-flow-redirect**. Save the configuration and reload the switch to apply the changes.



Note The **ing-flow-redirect** TCAM region must be explicitly carved to a size of 512.

- A secondary IP address (VIP) is required only when either vPC or EVPN ESI multi-homing is configured.
- Ethernet Segment and ESI configurations are supported only on Layer 2 port channel interfaces.
- To detect duplicate host addresses and set host move detection parameters, use these commands:
  - **l2rib dup-host-mac-detection** *num-of-host-moves dup-detection-timeout*
  - **fabric forwarding dup-host-ip-addr-detection** *num-of-host-moves dup-detection-timeout*

  For more information, see the Duplicate Detection for IP and MAC Addresses, on page 143 section.
- If a user performs **clear ip arp force-delete** on a VRF, and then clear mac address-table dynamic, the number of ARP entries on the ESI nodes may become unsynchronized. To prevent this inconsistent state, synchronize the ARP entries on all ESI nodes by using the **ip arp suppression-cache download remote vlan** *vlan_id* command.

PIP or VIP mode

- If re-origination occurs, you do not need to send the EAD/EVI route. By default, the EAD/EVI route is disabled for both PIP and VIP modes. Enable EAD/EVI using the **ead-evi route** command if required. This configuration optimizes the network by reducing the number of routes. For configuration details, see the EVPN EAD-EVI routes, on page 795 section.
- An L2FRR entry is programmed for the duration of the ESI downtime. This action redirects traffic to another functional ESI peer.
- In PIP mode, FRR must be configured to ensure that L2FRR functions correctly.
- In VIP next-hop mode, traffic may land on a VTEP whose ESI is down. This occurs because the remote VTEP continues to use the VIP as the next hop.
- **Mode transition behavior (from the standpoint of a remote VTEP)**
  - The VTEP operates in VIP next-hop mode when ESI is active on two or more VTEP peers.
  - It transitions to PIP next-hop mode when only one ESI peer remains active for a multi-homed host.
  - The system reverts to VIP next-hop mode if two or more ESI peers become active again.

  For more information, see the ESI multi-homing PIP modes, on page 786 and ESI multi-homing VIP modes, on page 788 sections.

Enable EVPN ESI multi-homing

Enable ESI-based multi-homing using the **evpn esi multihoming** command. The ESI multi-homing command enables Ethernet-segment configurations and generates Ethernet-segment routes on the switches.

Receiving Type-1 and Type-2 routes with a valid ESI, as well as path-list resolution, does not depend on the **evpn esi multihoming** command. If the switch receives MAC or MAC-IP routes with a valid ESI when the command is not enabled, ES-based path resolution logic still applies to these remote routes. This behavior ensures interoperability between vPC-enabled switches and ESI-enabled switches.



Note Cisco NX-OS supports either vPC-based EVPN multi-homing or ESI-based EVPN multi-homing, but not both simultaneously.

Follow these steps to enable the EVPN ESI multi-homing feature:

Before you begin

- Make sure TCAM carving is configured.
  - If TCAM carving is not already configured, use the following command to configure it:

    ```
    hardware access-list tcam region ing-flow-redirect 512
    ```

  - Reload the switch to apply the configuration.
- Configure maximum-path under the EVPN address-family on the Cisco NX-OS nodes. This configuration enables BGP to select multiple paths for EAD per ES and EAD per EVI routes. Remote VTEPs require this setting for ESI-RX functionality. This configuration allows the fabric to support multi-homing on ESI-VTEPs.

  Example:

  ```
  Leaf-7(config)# router bgp 1
  Leaf-7(config-router)# address-family l2vpn evpn
  Leaf-7(config-router-af)# maximum-paths ibgp 4

  Leaf-7(config-router)# vrf 3001
  Leaf-7(config-router-vrf)# address-family ipv4 unicast
  Leaf-7(config-router-vrf-af)# maximum-paths ibgp 4
  ```

Procedure

Step 1 Enter global configuration mode.

Use the **configure terminal** command.

Example:

```
switch# configure terminal
switch(config)#
```

Step 2 Enable EVPN ESI multi-homing.

Use the **evpn multihoming** command.

Example:

```
switch(config)# evpn multihoming
switch(config-evpn-mh)#
```

To disable EVPN ESI multi-homing, use the **no evpn multihoming** command.

Step 3 (Optional) Configure system MAC address.

Use the **system-mac** *system-mac* command with your chosen format to inherit globally configured MAC for all ESI.

Example:

```
switch(config-evpn-mh)# system-mac 1200.0000.0000
```

Supported formats for the global system MAC address are:

- system-mac E.E.E,
- system-mac EE-EE-EE-EE-EE-EE,
- system-mac EE:EE:EE:EE:EE:EE, and
- system-mac EEEE.EEEE.EEEE

Step 4 (Optional) Set DF election mode.

Use the **df-election mode {modulo | per-flow}** command.

Example:

```
switch(config-evpn-mh)# df-election mode modulo
```

- Default mode: per-flow.
- To restore the default, use the **no df-election mode** command.

Step 5 (Optional) Configure DF election timer.

Use the **df-election time** *secs* command to set the DF election timer for all ES.

Example:

```
switch(config-evpn-mh)# df-election time 1.0
```

- Range: 1 to 3 seconds. Default: 3 seconds.
- To restore the default, use the **no df-election time** *secs* command.

Tip
The default three second DF election timer ensures stable DF elections. However, a one second timer can accelerate convergence and reduce failover time, suitable for low-latency, critical environments.

Step 6 (Optional) Enable EAD-EVI route advertisement.

Use the **ead-evi route** command.

Example:

```
switch(config-evpn-mh)# ead-evi route
```

- By default, EAD-EVI route advertisement is disabled.
- To disable, use the **no ead-evi route** command.

Step 7 (Optional) Configure Ethernet segment delay restore time.

Use the **ethernet-segment delay-restore time** *secs* command.

Example:

```
switch(config-evpn-mh)# ethernet-segment delay-restore time 180
```

- Range: 30 to 1000 seconds. Default: 180 seconds.
- To restore the default, use the **no ethernet-segment delay-restore time** *secs* command.

Step 8 Configure fast reroute anycast source IP (L2FRR).

Use the **frr anycast source-ip** *ip-address* command.

Example:

```
switch(config-evpn-mh)# frr anycast source-ip 10.0.0.13
```

Example:

Note

- For L2FRR to work properly, ensure the VIP or fast reroute anycast source IP is configured correctly.
- Fast reroute anycast source IP is mandatory for PIP mode.
- In VIP mode of operation, this command is optional and uses the VIP as the anycast source IP. The configured anycast source IP takes precedence over the VIP address.
- If 4-way is used, configure the same IP on all four nodes.
- The fast reroute anycast source IP address must be unique across the entire fabric. For PIP mode, configuration on a specific interface is not required.

Step 9 Exit EVPN multi-homing configuration mode.

Use the **exit** command.

Example:

```
switch(config-evpn-mh)# exit
switch(config)#
```

Step 10 Enable core-links tracking on L3 interface toward spine.

In interface config mode, use the **evpn multihoming core-tracking** command.

Example:

```
switch(config)# interface Ethernet1/1
switch(config-if)# description intf_site_1_spine_1_1
switch(config-if)# evpn multihoming core-tracking
switch(config-evpn-mh)#
```

To remove core-link tracking, use the **no evpn multihoming core-tracking** command.

Note
This step is required; without it, the ES is not UP.

Step 11 (Optional) Verify configuration.

Use the **show running-config nv overlay** command.

Example:

```
switch# show running-config nv overlay
evpn multihoming
  ethernet-segment delay-restore time 120
  ead-evi route
```

```
  frr anycast source-ip 10.0.0.13
  df-election mode modulo
  df-election time 1.0
  system-mac 1200.0000.0000
advertise evpn multicast
switch#
```

EVPN ESI multi-homing is enabled and operational. Verification confirms the configuration is active and correct.

Configure ESI under a port channel interface

Configure an ESI for a Layer 2 port channel interface to enable multi-homing and redundancy in an EVPN environment. The port channel can then participate in multi-chassis link aggregation, supporting features such as active-active redundancy and seamless failover between devices.

Follow these steps to configure an ESI for a port channel interface.

Before you begin

- Before you begin, ensure that Port channel is configured as a Layer 2 port channel.
- To inherit the globally configured system MAC in a Type-3 ESI, you must first configure the global system MAC.

Procedure

Step 1 Enter global configuration mode.

Use the **configure terminal** command.

Example:

```
switch# configure terminal
switch(config)#
```

Step 2 Enter interface mode.

Use the **interface port-channel** *value* command.

Example:

```
switch(config)# interface port-channel 1
switch(config-if)#
```

Step 3 Enable ES configuration.

Use the **ethernet-segment** command.

Example:

```
switch(config-if)# ethernet-segment
switch(config-if-ethernet-segment)#
```

Step 4 Configure ESI under port channel.

Use the **esi** *value* | **esi system-mac** [*system-mac*] *local-discriminator* command.

The supported ESI route types are:

- Type-0: Configure a 10-byte ESI value manually, with the first byte set to 00.

  Example:

  ```
  esi 004a.230a.03c8.8299.58e0
  ```

- Type-3: These options are available to configure Type-3 ESI.

  The ESI Type-3 consists of the PE system MAC address combined with a local discriminator (unique 32-bit value). The local discriminator distinguishes Ethernet segments that share the same PE system MAC address, ensuring each Ethernet Segment Identifier is unique.

Table 17: Type-3 ESI configuration options

If	Then	Example
Global system MAC inherited without local identifier (local discriminator)	System MAC is derived from global system MAC. **Note** If the `local-Identifier` is omitted, port channel number from the port channel interface will be inherited as local discriminator value.	`switch(config-if-ethernet-segment)# esi system-mac`
Global system MAC inherited with local identifier (local discriminator)	The system MAC is derived from the global system MAC. Additionally, configure a local identifier (local discriminator) with the system MAC. The valid range is 0 to 16,777,215.	`switch(config-if-ethernet-segment)# esi system-mac 1`
Manually configured system MAC without local identifier (local discriminator)	The user defines the system MAC. **Note** If the `local-Identifier` is omitted, port channel number from the port channel interface will be inherited as local discriminator value.	For a manually configured system MAC without local discriminator, the local discriminator is inherited from the port channel number. `switch(config-if-ethernet-segment)# esi system-mac 0334.5600.0000`
Manually configured system MAC with local identifier (local discriminator)	The user defines the system MAC. Additionally local identifier (local discriminator) is configured along with user defined system MAC. Range: 0 - 16777215.	`switch(config-if-ethernet-segment)#esi system-mac 0334.5600.0000 1`

Step 5 Exit until you reach configuration mode.

Use the **exit** command.

Example:

```
switch(config-if-ethernet-segment)# exit
switch(config-evpn-mh)# exit
switch(config)#
```

Step 6 Configure the NVE interface.

Use the **interface nve** *value* command.

Example:

```
switch(config)# interface nve 1
(config-if-nve)#
```

Step 7 (Optional) Configure Anycast VIP interface.

Use the **source-interface loopback** *Id* **anycast loopback** *Id* command in NVE interface.

Example:

```
switch(config)# interface nve 1
(config-if-nve)# source-interface loopback1 anycast loopback2
```

To remove the NVE source and anycast loopback interface mapping, use the **no source-interface loopback** *Id* **anycast loopback** *Id* command.

Note
To remove only anycast loopback interface mapping, use the **source-interface loopback** *Id* command.

Step 8 (Optional) Verify the port channel configuration.

Use the **show running-config interface port-channel** *value* command.

Example:

```
switch# show running-config interface port-channel
interface port-channel1
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 2-6
  ethernet-segment
    esi system-mac 1
  spanning-tree port type edge trunk

interface Ethernet1/14/4

  description intf_site_1_l2sw_1_esia_1
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 2-6
  channel-group 1 mode active
  no shutdown
```

Step 9 (Optional) Verify whether the ES status is up on all the ESI multi-homing nodes after ES configuration.

Use the **show nve ethernet-segment** command.

Example:

```
switch# show nve ethernet-segment s
ESI                             Parent interface   ES State
------------------------------   ------------------ ----------
0334.5600.0000.0000.0001         port-channel1      Up
0334.5600.0000.0000.0002         port-channel2      Up
switch# show nve ethernet-segment

ESI: 0080.7e59.e1e4.5459.e022
   Parent interface: port-channel201
  ES State: Up                      !Verify ES Status is UP
  Port-channel state: Up
  NVE Interface: nve1
   NVE State: Up
   Host Learning Mode: control-plane
```

```
  Anycast IP: 192.0.2.4
  Active Vlans: 2-41
   DF Vlans: 3,5,7,9,11,13,15,17,19,21,23,25,27,29,31,33,35,37,39,41
   Active VNIs: 2000002-2000041
  DF BDs: N/A
   DF VNIs: N/A
   Number of ES members: 2
  My ordinal: 1
  DF timer start time: 00:00:00
  DF timer expiry: 08:08:09
  Config State: config-applied
  DF List: 192.0.2.12 192.0.2.13
  Bounce peer : 192.0.2.12
  ES route added to L2RIB: True
  EAD/ES routes added to L2RIB: True
  EAD/EVI route timer age: not running [Disabled]
  EAD/EVI timer duration: 00:05:00
  ESI type: Ether-segment
  ESI DF election mode: Modulo
switch#
```

Step 10 (Optional) Verify the NVE interface configuration details.

Use the **show nve interface nve 1 detail** command.

Example:

```
switch# show nve interface nve 1 detail
Interface: nve1, State: Up, encapsulation: VXLAN
 VPC Capability: VPC-VIP-Only [not-notified]
 Local Router MAC: f4ee.31e8.f2fd
 Host Learning Mode: Control-Plane
 Source-Interface: loopback1 (primary: 192.0.2.11)          !Verify source interface loopback IP
address
 Anycast-Interface: loopback2 (secondary: 192.0.2.1)           !Verify Anycast interface loopback
 IP address
 Source Interface State: Up                                  !Verify source interface Status
 Anycast Interface State: Up                                    !Verify source interface Status
 ESI multihoming anycast-restore time left: 0 seconds
 Virtual RMAC Advertisement: No
 NVE Flags:
 Interface Handle: 0x49000001
 Source Interface hold-down-time: 180
 Source Interface hold-up-time: 30
 Remaining hold-down time: 0 seconds
 Virtual Router MAC: 0200.0101.0101
 Interface state: nve-intf-add-complete
 ESI multihoming delay-restore time: 180 seconds
 ESI multihoming delay-restore time left: 0 seconds
 ESI multihoming FRR anycast source IP: 192.0.2.1          !Verify source interface loopback IP
address
 Fabric convergence time: 135 seconds
 Fabric convergence time left: 0 seconds
```

The port channel is successfully configured with the specified ESI, enabling it for multi-homing in an EVPN environment. The system recognizes the ESI configuration, and verification commands confirm that the ESI and related interfaces are operational, supporting redundancy and resiliency for Layer 2 services.

What to do next

- Verify connectivity and redundancy.

- Monitor the status of the ESI and port channel interfaces and confirm they remain in the "Up" state.

Verification commands for ESI multi-homing configuration

Use these commands to view ESI multi-homing configuration information and verify the ESI multi-homing setup on your device.

Table 18: Verification commands for ESI multi-homing configuration

Command	Purpose
show bgp evi	Displays BGP EVPN routes learned by the switch.
show bgp l2vpn evpn route-type *value*	Displays BGP route type state.
show l2route topology all	Displays a comprehensive view of the Layer 2 forwarding table with information about MAC addresses, associated VLANs, and next-hops.
show l2route evpn mac all	Displays all learned MAC addresses within the EVPN domain.
show l2route evpn mac-ip all	Displays all MAC and IP addresses that the switch learned through BGP EVPN.
show l2route evpn imet all	Displays the MAC addresses, their associated VNI, and the VTEP from which the addresses were learned.
show l2route evpn fl all	Displays all Layer 2 EVPN routes the switch learned, both locally and remotely.
show l2route evpn ead all	Displays information about all EAD routes in the EVPN routing table.
show l2route evpn ethernet-segment {**esi** *esi-id* | **all**}[**bgp** | **vxlan**] [**detail**]	Displays information about Ethernet Segments in an EVPN environment, specifically within a VXLAN fabric.
show l2route evpn ethernet-segment {**esi** *esi-id* | **all**}[**detail**]	Displays information about Ethernet Segments in an EVPN environment on the switch.
show l2route smet {**topology** *topo-id* | **all**} [**detail**]	Displays information about Layer 2 Multicast (SMET) routes within a VXLAN EVPN fabric.
show l2routereport-sync{**topology** *topo-id* | **all**} [**esi***esi- id*] [**detail**]	Displays details about the synchronization of Layer 2 EVPN routes reports.
show l2routeleave-sync{**topology** *topo-id* | **all**} [**esi** *esi-id*] [**detail**]	Displays details about leave synchronization for Layer 2 EVPN routes.
show nve interface nve1 detail	Displays information about the Network Virtualization Edge interface nve1.
show nve ethernet-segment	Displays detailed information about configured NVE Ethernet segments.

Command	Purpose
show nve core-links	Displays details about the configured Ethernet segments, including their status and associated VNIs.
show [ip | ipv6] [igmp | mld] snooping groups [vlan] [vlan ID]	Displays the corresponding group address, version, group type, and port list information, depending on IGMP or MLD.
show ipv6 mld groups	Displays information about MLD snooping.
show ip igmp snooping evpn	Displays information about the IP IGMP snooping configuration and statistics.
show ipv6 mld snooping evpn	Displays information about the IPv6 MLD snooping configuration.
show ip igmp snooping evpn esi	Displays the IGMP snooping configuration and status for the EVPN context, including ESI details, port-channel, ESI ID, and type.
show ipv6 mld snooping evpn esi	Displays information about MLD snooping within the EVPN context, including ESI details, port-channel, ESI ID, and type.
show ip igmp snooping evpn report-sync	Displays IGMP snooping and EVPN report-sync information for VLAN, VNI, group, source, local/remote, and version flags.
show ipv6 mld snooping evpn report-sync	Displays IPv6 IGMP snooping and EVPN report-sync information for VLAN, VNI, group, source, local/remote, and version flags.
show ip igmp snooping evpn leave-sync	Displays the IGMP snooping leave synchronization status in an EVPN environment.
show ipv6 mld snooping evpn leave-sync	Displays the IPv6 IGMP snooping leave synchronization status in an EVPN environment.
show ip igmp snooping remote groups	Displays the remote IGMP group address, version, group type, and port list at VLAN level.
show ipv6 mld snooping remote groups	Displays the remote MLD group address, version, group type, and port list at VLAN level.
show bgp l2vpn evpn route-type 7	Displays the state of BGP L2VPN EVPN route Type-7 and the report-sync routing table information for all VRFs.
show bgp l2vpn evpn route-type 8	Displays the state of BGP L2VPN EVPN route Type-8 and leave-sync routing table information for all VRFs.
show ip adjacency vrf *vrf-id*	Displays IP adjacency information for a specific VRF.
show ipv6 adjacency vrf *vrf-id*	Displays IPv6 adjacency information within a specific VRF.
show ip arp vrf *vrf-id*	Displays the ARP table for a specific VRF instance.

Command	Purpose
show ip arp suppression-cache vlan *vlan-id*	Displays ARP entries being suppressed for the specified VLAN, including IP address, MAC address, and interface for each entry.
show ipv6 neighbor vrf *vrf_name*	Displays the IPv6 neighbors learned via NDP within the specific VRF.
show ipv6 icmp neighbor vrf *vrf_name*	Displays the IPv6 neighbor entries for a specified VRF, typically learned via ICMPv6 Neighbor Discovery.
show mac address-table dynamic	Displays all MAC addresses learned dynamically by the switch, along with associated interface and VLAN.
show mac address-table count es all	Displays the total number of MAC addresses learned for each Ethernet Segment.
show forwarding route *ipaddr* **platform**	Displays the routing table, including learned routes, next hops, and associated interfaces.

ESI multi-homing configuration examples

This section explains how to configure and verify the ESI multi-homing (MH) feature on Cisco devices, covering baseline setup, ESI enablement, and NVE interface configuration for both PIP and VIP modes.

Pre configure for ESI multi-homing

Before configuring ESI multi-homing, ensure the following steps are complete

1. Create a TCAM carving. TCAM carving is required to enable the ESI-MH feature.

```
Leaf-3(config)# hardware access-list tcam region ing-flow-redirect 512
```

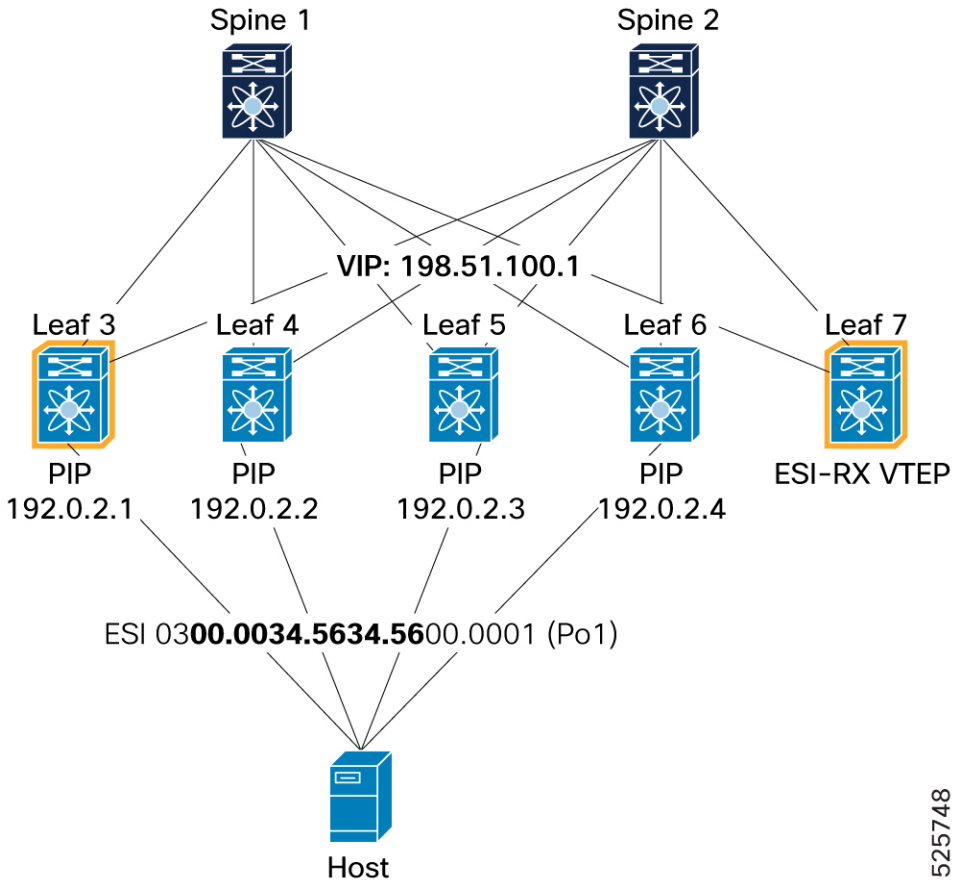
2. Enable maximum-paths. The`maximum-paths` is required on ESI-RX VTEPs to enable ECMP.

```
Leaf-7(config)# router bgp 1
Leaf-7(config-router)# address-family l2vpn evpn
Leaf-7(config-router-af)# maximum-paths ibgp 4

Leaf-7(config-router)# vrf 3001
Leaf-7(config-router-vrf)# address-family ipv4 unicast
Leaf-7(config-router-vrf-af)# maximum-paths ibgp 4
```

Configure ESI multi-homing

Figure 91: Enable ESI MH



1. Enable ESI multi-homing with system MAC. The `system-mac` is required to generate the Ethernet Segment ID (ESI).

```
Leaf-3(config)# evpn multihoming
Leaf-3(config-evpn-mh)# system-mac 0000.3456.3456
```

2. Enable L2 Fast Reroute (L2FRR). L2FRR is designed to minimize traffic disruption during ESI failure conditions. This command is mandatory in PIP mode but optional in VIP mode, where VIP is used as the anycast source IP.

```
Leaf-3(config-evpn-mh)# frr anycast source-ip 203.0.113.1
```

3. Enable core links tracking on L3 interfaces to check the operational state of the core links. When all core links are down, ESI port-channel interfaces also go down to avoid traffic null route.

```
Leaf-3(config)# interface Ethernet1/1-2
Leaf-3(config-if)# evpn multihoming core-tracking
```

Enable Ethernet Segment

1. Enable Ethernet Segment (ES) on Port-Channel interface.

```
Leaf-3(config)# interface port-channel 1
Leaf-3(config-if)# ethernet-segment
```

2. Configure ESI
   - Type 0: Arbitrary 9-byte values

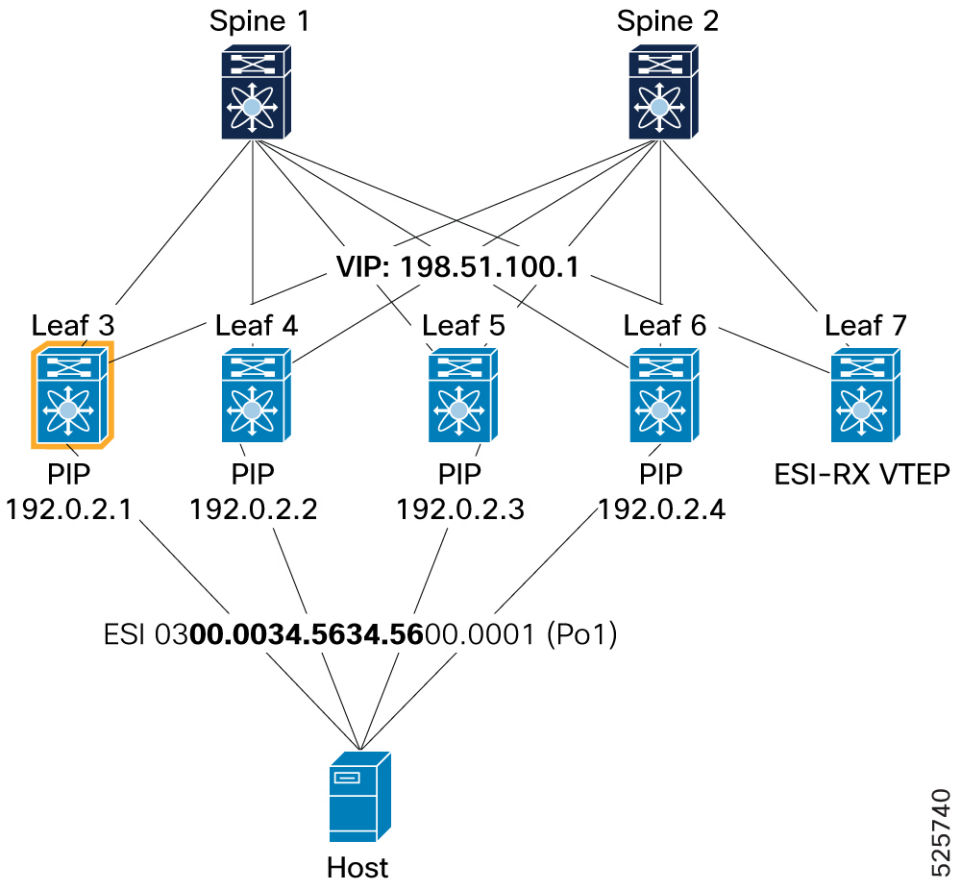
     ```
     Leaf-3(config-if-ethernet-segment)# esi 0021.0000.0000.2100.0222
     ```

   - Type 3: System MAC address (6 bytes) and Local Discriminator (3 bytes). For more details of Type-3 system MAC options, see the Configure ESI under a port channel interface, on page 808 section.

     ```
     Leaf-3(config-if-ethernet-segment)# esi system-mac
     ```

Enable ESI multi-homing PIP or VIP mode

Figure 92: VIP and PIP mode configuration



1. Configuration for PIP mode
   - Create loopback interface with NVE source IP.

     ```
     Leaf-3(config)# interface loopback 1
     Leaf-3(config-if)# ip address 192.0.2.1/32
     ```

   - Enable PIP on NVE interface.

     ```
     Leaf-3(config)# interface nve1
     Leaf-3(config-if-nve)# source-interface loopback1
     ```

2. Configuration for PIP mode

   - Create loopback interface with NVE source IP.

     ```
     Leaf-3(config)# interface loopback 1
     Leaf-3(config-if)# ip address 192.0.2.1/32
     ```

   - Create loopback interface with Anycast VIP.

     ```
     Leaf-3(config)# interface loopback 2
     Leaf-3(config-if)# ip address 198.51.100.1/32
     ```

   - Enable Anycast VIP on NVE interface.

     ```
     Leaf-3(config)# interface nve1
     Leaf-3(config-if-nve)# source-interface loopback1 anycast loopback2
     ```

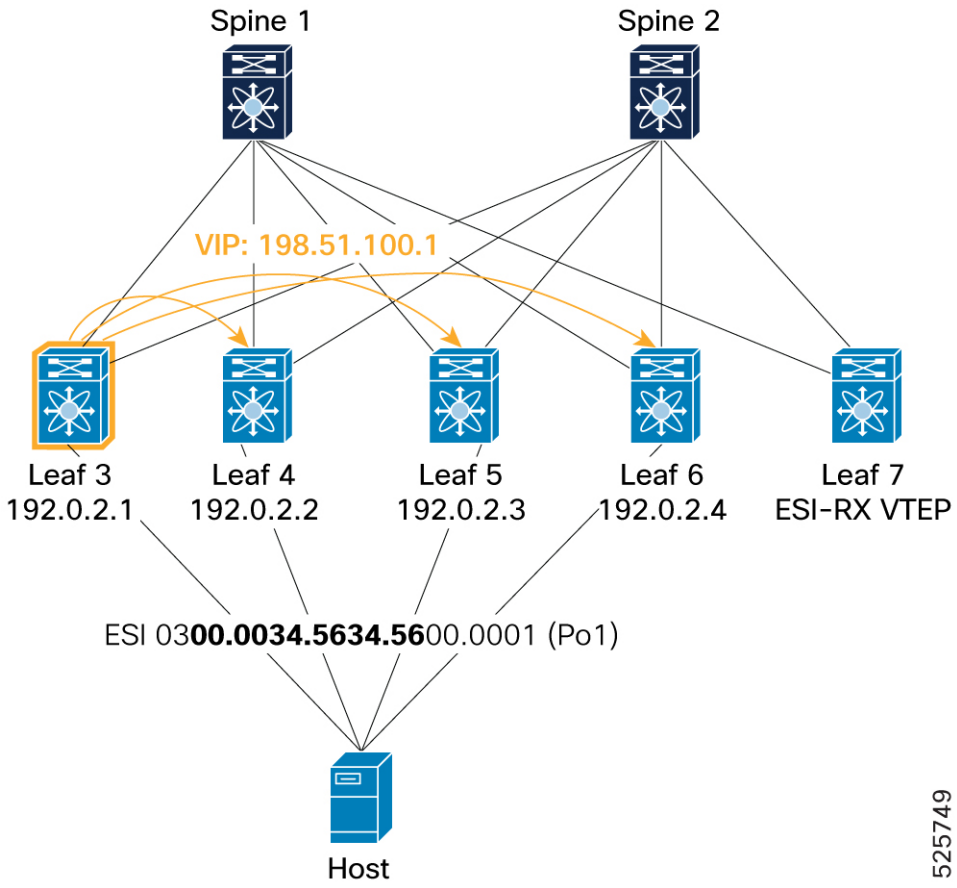
Verification of VIP mode

This section details the verification steps for ESI multi-homing VIP Mode. The validation covers the advertisement and synchronization of EVPN Type-1 (EAD-ES) and Type-2 (MAC-IP) routes across local and remote VTEPs using a VIP.

1. EVPN EAD-ES route advertisements

   - **Local route (from Leaf 3)**: Shows that Leaf 3, as an ESI node, advertises its local EAD-ES route (Type-1) with a VIP Next-Hop (198.51.100.1) to its ESI peers (Leaf 4, Leaf 5 and Leaf 6) and BGP. The `show l2route evpn ead es` and `show bgp l2vpn evpn route-type 1` outputs confirm the local origination and the VIP as the egress next-hop.

Figure 93: Verification of VIP mode – Local route



Example for BGP component:

```
Leaf3# show bgp l2vpn evpn route-type 1

Route Distinguisher: 192.0.13.1:7817    (EAD-ES [0300.0034.5634.5600.0001 7817])
BGP routing table entry for [1]:[0300.0034.5634.5600.0001]:[0xffffffff]/152, version
 154
Paths: (1 available, best #1)
Flags: (0x000002) (high32 00000000) on xmit-list, is not in l2rib/evpn
Multipath: eBGP iBGP

  Advertised path-id 1
  Path type: local, path is valid, is best path, no labeled nexthop
  AS-Path: NONE, path locally originated
    192.0.2.1 (metric 0) from 0.0.0.0 (192.0.13.1)
      Origin IGP, MED not set, localpref 100, weight 32768
      Received label 0
      Extcommunity: RT:1:2001001 RT:1:2001002 ENCAP:8 ESI:32:000000
      Tunnel Encapsulation Attribute: Type: 8 Length: 12 Sub Type: 6
         Egress Nexthop: 198.51.100.1

  Path-id 1 advertised to peers:
    192.0.2.21          192.0.2.22
```

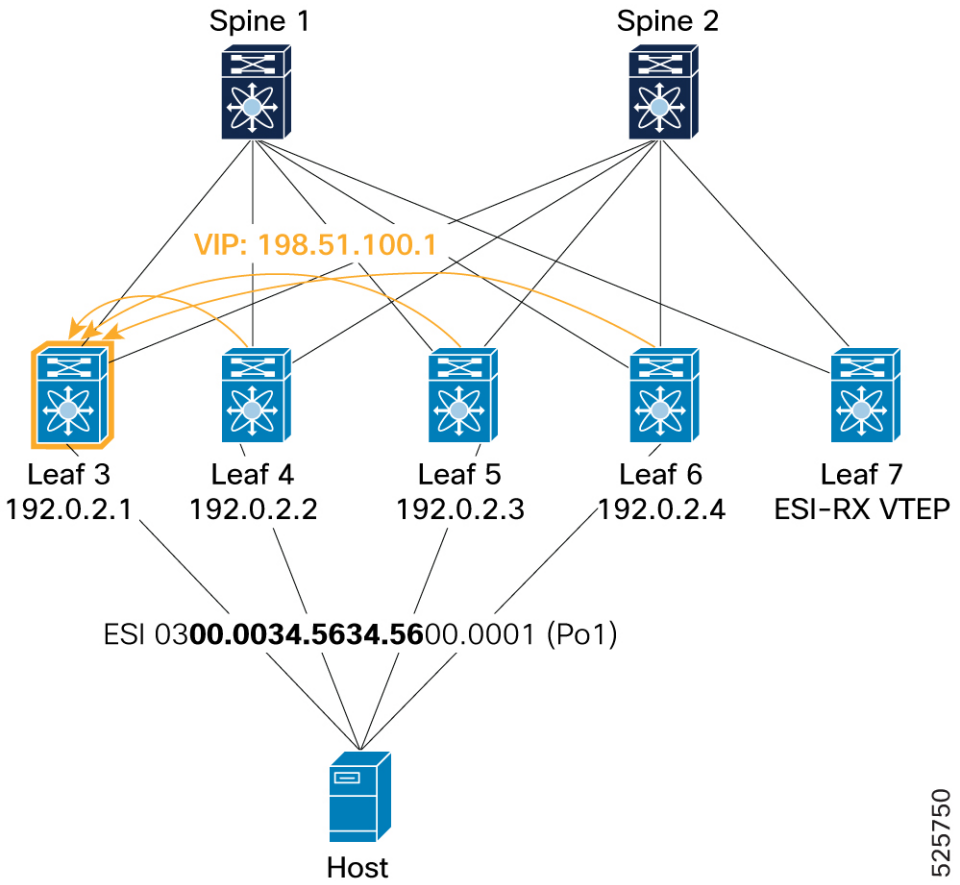
Example for L2RIB component:

```
Leaf3# show l2route evpn ead es

Topology ID   Prod   ESI                        Sent To    Num PLs  Flags
------------- ------ -------------------------- ---------- -------- ----------
4294967294    VXLAN  0300.0034.5634.5600.0001   BGP        0        A
                     Next-Hops: 192.0.2.1
                     VIP Next-Hop: 198.51.100.1
```

- **Remote route (received by Leaf3)**: Shows Leaf 3 receiving Type-1 routes from other remote ESI peers (Leaf 4, Leaf 5, and Leaf 6) within the 4-way ESI cluster. These routes also indicate the VIP (198.51.100.1) as the egress next-hop, and Leaf3's L2RIB correctly populates the next-hops for these remote ESI members.

Figure 94: Verification of VIP mode – Remote route



Example for BGP component:

```
Leaf3# show bgp l2vpn evpn route-type 1

Route Distinguisher: 192.0.13.1:65534    (L2VNI 0)<SNIP>
  Path type: internal, path is valid, not best reason: Router Id, multipath, no
labeled nexthop, in rib
             Imported from
192.0.16.1:7821:[1]:[0300.0034.5634.5600.0001]:[0xffffffff]/152
  AS-Path: NONE, path sourced internal to AS
    192.0.2.4 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
```

```
      Extcommunity: RT:1:2002001 RT:1:2002002 ENCAP:8 ESI:32:000000
      Originator: 192.0.16.1 Cluster list: 192.0.2.21
      Tunnel Encapsulation Attribute: Type: 8 Length: 12 Sub Type: 6
         Egress Nexthop: 198.51.100.1

  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop, in rib
             Imported from
192.0.14.1:7821:[1]:[0300.0034.5634.5600.0001]:[0xffffffff]/152
  AS-Path: NONE, path sourced internal to AS
    192.0.2.2 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Extcommunity: RT:1:2002001 RT:1:2002002 ENCAP:8 ESI:32:000000
      Originator: 192.0.14.1 Cluster list: 192.0.2.21
      Tunnel Encapsulation Attribute: Type: 8 Length: 12 Sub Type: 6
         Egress Nexthop: 198.51.100.1

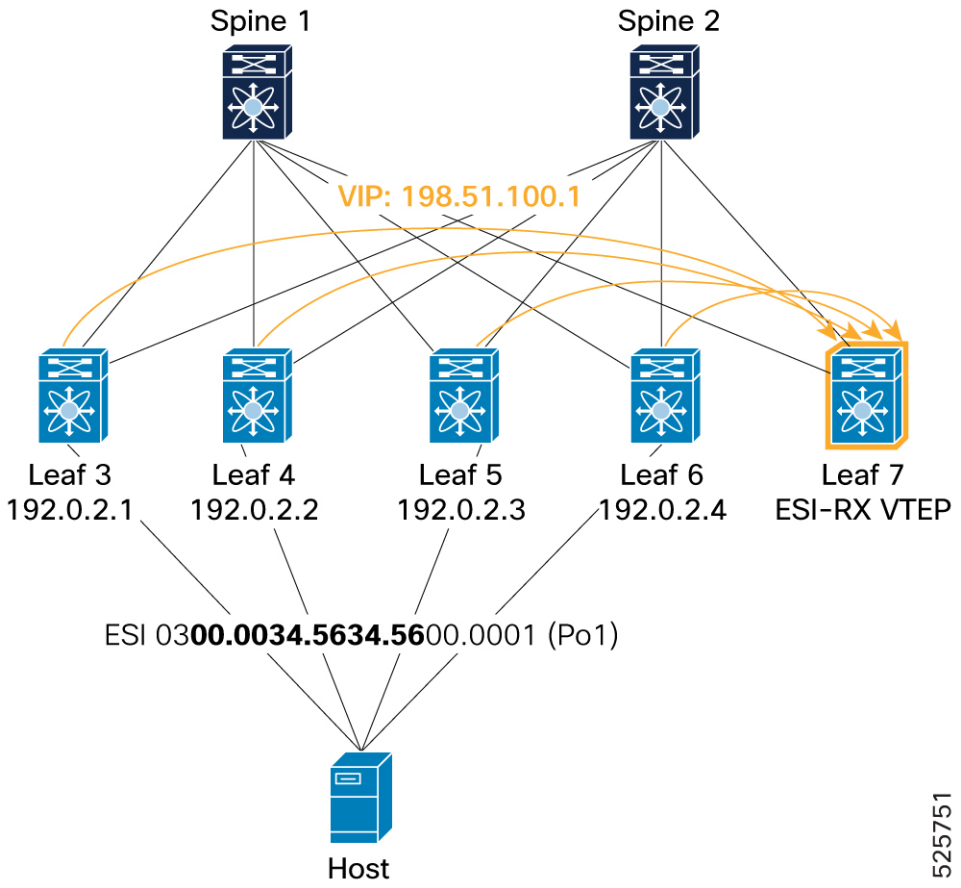
  Path type: internal, path is valid, not best reason: Router Id, multipath, no
labeled nexthop, in rib
             Imported from
192.0.2.10:7821:[1]:[0300.0034.5634.5600.0001]:[0xffffffff]/152
  AS-Path: NONE, path sourced internal to AS
    192.0.2.3 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Extcommunity: RT:1:2002001 RT:1:2002002 ENCAP:8 ESI:32:000000
      Originator: 192.0.2.10 Cluster list: 192.0.2.21
      Tunnel Encapsulation Attribute: Type: 8 Length: 12 Sub Type: 6
         Egress Nexthop: 198.51.100.1
```

Example for L2RIB component:

```
Leaf3# show l2route evpn ead es
Topology ID   Prod   ESI                        Sent To    Num PLs  Flags
------------- ------ -------------------------- ---------- -------- ----------
4294967294    BGP    0300.0034.5634.5600.0001   -          0        A
              Next-Hops: 192.0.2.2
                         192.0.2.3
                         192.0.2.4
```

- **Remote VTEP (ESI-RX, Leaf 7)**: Verifies that a remote VTEP (Leaf 7) receives Type-1 EAD-ES route advertisements from all four ESI nodes (Leaf 3, Leaf 4, Leaf 5, and Leaf 6) within the 4- way ESI cluster, with the VIP (198.51.100.1) consistently being the egress next-hop. Leaf 7's L2RIB reflects all four VTEP IPs as next-hops for the ESI.

Figure 95: Verification of VIP mode - Remote VTEP (ESI-RX)



Example for BGP component:

```
Leaf 7# show bgp l2vpn evpn route-type 1

Route Distinguisher: 192.0.17.1:65534    (L2VNI 0)
<SNIP>
  Path type: internal, path is valid, not best reason: Router Id, multipath, no
labeled nexthop, in rib
             Imported from
192.0.16.1:7821:[1]:[0300.0034.5634.5600.0002]:[0xffffffff]/152
  AS-Path: NONE, path sourced internal to AS
    192.0.2.4 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Extcommunity: RT:1:2002001 RT:1:2002002 ENCAP:8 ESI:32:000000
      Originator: 192.0.16.1 Cluster list: 192.0.2.21
      Tunnel Encapsulation Attribute: Type: 8 Length: 12 Sub Type: 6
         Egress Nexthop: 198.51.100.1

  Path type: internal, path is valid, not best reason: Router Id, multipath, no
labeled nexthop, in rib
             Imported from
192.0.14.1:7821:[1]:[0300.0034.5634.5600.0002]:[0xffffffff]/152
  AS-Path: NONE, path sourced internal to AS
    192.0.2.2 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
```

```
      Received label 0
      Extcommunity: RT:1:2002001 RT:1:2002002 ENCAP:8 ESI:32:000000
      Originator: 192.0.14.1 Cluster list: 192.0.2.21
      Tunnel Encapsulation Attribute: Type: 8 Length: 12 Sub Type: 6
         Egress Nexthop: 198.51.100.1

  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop, in rib
             Imported from
192.0.13.1:7821:[1]:[0300.0034.5634.5600.0002]:[0xffffffff]/152
  AS-Path: NONE, path sourced internal to AS
    192.0.2.1 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Extcommunity: RT:1:2002001 RT:1:2002002 ENCAP:8 ESI:32:000000
      Originator: 192.0.13.1 Cluster list: 192.0.2.21
      Tunnel Encapsulation Attribute: Type: 8 Length: 12 Sub Type: 6
         Egress Nexthop: 198.51.100.1

  Path type: internal, path is valid, not best reason: Router Id, multipath, no
labeled nexthop, in rib
             Imported from
192.0.2.10:7821:[1]:[0300.0034.5634.5600.0002]:[0xffffffff]/152
  AS-Path: NONE, path sourced internal to AS
    192.0.2.3 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Extcommunity: RT:1:2002001 RT:1:2002002 ENCAP:8 ESI:32:000000
      Originator: 192.0.2.10 Cluster list: 192.0.2.21
      Tunnel Encapsulation Attribute: Type: 8 Length: 12 Sub Type: 6
         Egress Nexthop: 198.51.100.1
```

Example for L2RIB component:

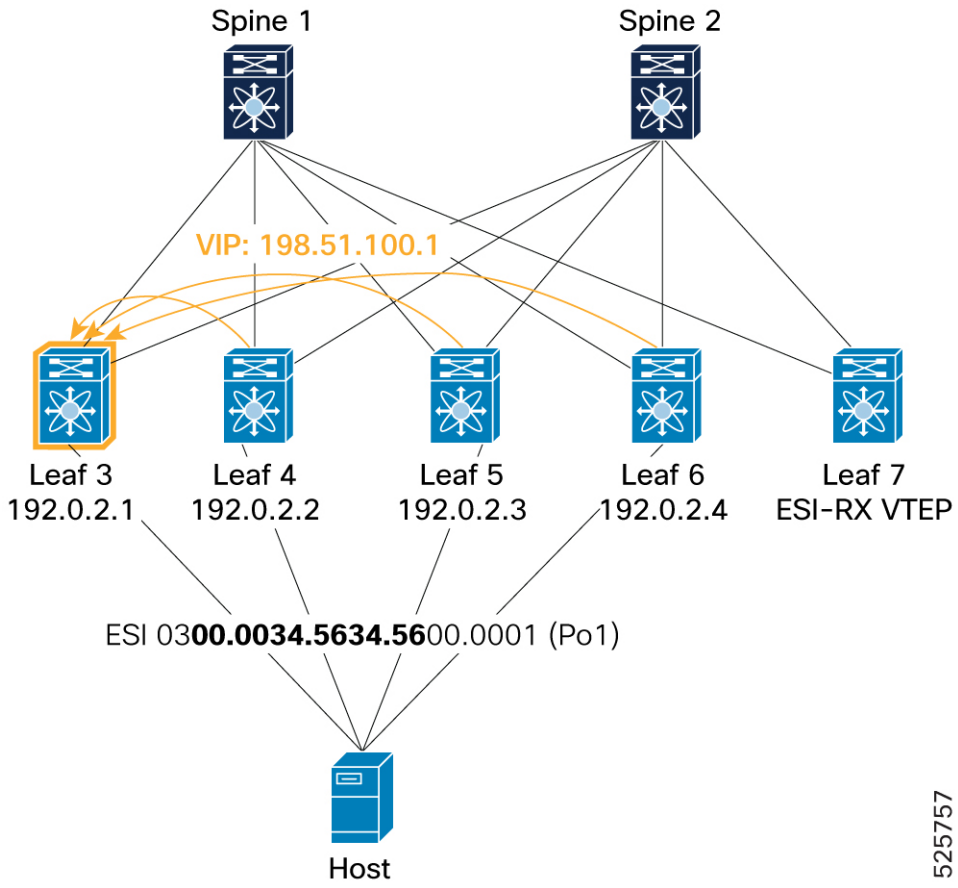
```
Leaf 7# show l2route evpn ead es

Topology ID   Prod   ESI                         Sent To    Num PLs  Flags
------------- ------ --------------------------- ---------- -------- ----------
4294967294    BGP    0300.0034.5634.5600.0001   -          2       A
              Next-Hops: 192.0.2.1
                         192.0.2.2
                         192.0.2.3
                         192.0.2.4
```

2. Verification of NVE Ethernet Segment

   Confirms the Ethernet Segment (ESI: 0300.0034.5634.5600.0001) on Leaf3 is "Up" and associated with port-channel1. The `show nve ethernet-segment` and `show nve interface nve 1` detail commands on Leaf 3 confirm the configuration of the Anycast IP: 198.51.100.1 and list all four ESI members (192.0.2.1, 192.0.2.2, 192.0.2.3, 192.0.2.4) in the Designated Forwarder (DF) list. The NVE interface detail also shows the Anycast-Interface as loopback2 (secondary: 198.51.100.1).

Figure 96: VIP mode -NVE Ethernet Segment verifications




```
!Shows NVE status
Leaf3# show nve ethernet-segment summary
ESI                              Parent interface   ES State
-------------------------------  ------------------ ----------
0300.0034.5634.5600.0001         port-channel1      Up

!Shows ES status and ES configuration information
Leaf3# show nve ethernet-segment

ESI: 0300.0034.5634.5600.0001
   Parent interface: port-channel1
  ES State: Up
  Port-channel state: Up
  NVE Interface: nve1
   NVE State: Up
   Host Learning Mode: control-plane
  Anycast IP: 198.51.100.1
  Active Vlans: 1001-1002
   DF Vlans:
   Active VNIs: 2001001-2001002
  DF BDs: N/A
   DF VNIs: N/A
   Number of ES members: 4
  My ordinal: 0
  DF timer start time: 00:00:00
  DF timer expiry: 16:05:31
```

```
  Config State: config-applied
  DF List: 192.0.2.1 192.0.2.2 192.0.2.3 192.0.2.4
  Bounce peer : 192.0.2.4
  ES route added to L2RIB: True
  EAD/ES routes added to L2RIB: True
  EAD/EVI route timer age: not running [Disabled]
  EAD/EVI timer duration: 00:05:00
  ESI type: Ether-segment
  ESI DF election mode: Per-flow

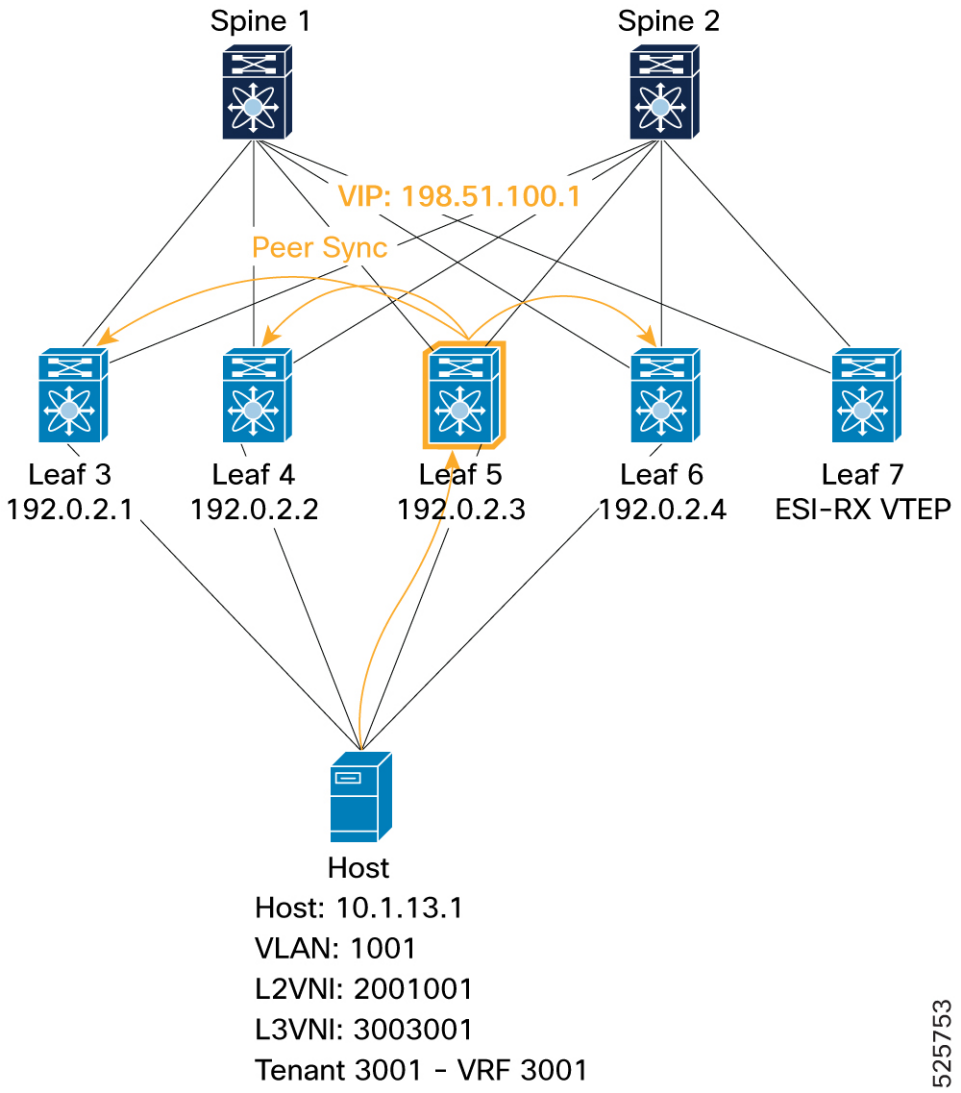
!Shows ES configuration information in detial
Leaf3# show nve interface nve 1 detail

Interface: nve1, State: Up, encapsulation: VXLAN
 VPC Capability: VPC-VIP-Only [not-notified]
 Local Router MAC: 4880.0290.0727
 Host Learning Mode: Control-Plane
 Source-Interface: loopback1 (primary: 192.0.2.1)
 Anycast-Interface: loopback2 (secondary: 198.51.100.1)
 Source Interface State: Up
 Anycast Interface State: Up
 ESI multihoming anycast-restore time left: 0 seconds
 Virtual RMAC Advertisement: No
 NVE Flags:
 Interface Handle: 0x49000001
 Source Interface hold-down-time: 180
 Source Interface hold-up-time: 30
 Remaining hold-down time: 0 seconds
 Virtual Router MAC: 0200.6701.8601
 Interface state: nve-intf-add-complete
 ESI multihoming delay-restore time: 180 seconds
 ESI multihoming delay-restore time left: 0 seconds
 ESI multihoming FRR anycast source IP: 203.0.113.1
 Fabric convergence time: 135 seconds
 Fabric convergence time left: 0 seconds
```

3. EVPN Type-2 Route Advertisement (MAC-IP)

   - **Local VTEP to ESI peers (peer sync)**: Demonstrates a host (MAC: 0010.0100.1301, IP: 10.1.13.1) learned locally on Leaf 5. Leaf 5 then advertises this Type-2 route as "locally originated" with its ESI. Other ESI peers (Leaf 3, Leaf 4, and Leaf 6) receive this route as "Peer Synced" (`PS` flag in L2RIB), indicating they learned it from a peer within the same ESI.

Figure 97: Verification of VIP mode - Local VTEP to ESI peers (peer sync)



L2FM:

```
Leaf5# show mac address-table vlan 1001
VLAN     MAC Address      Type      age     Secure NTFY Ports
---------+-----------------+--------+---------+------+----+------------------
C 1001     0010.0100.1301   dynamic  NA         F      F    Po1
```

ARP:

```
Leaf5# show ip arp vrf 3001

Address         Age       MAC Address     Interface       Flags
10.1.13.1       00:18:43  0010.0100.1301  Vlan1001
```

L2RIB:

```
Leaf5# show l2route evpn mac-ip evi 1001 detail

Topology    Mac Address    Host IP                                   Prod   Flags
       Seq No     Next-Hops
```

```
----------- -------------- ------------------------------------ ------
----------------- ---------- --------------------------
1001        0010.0100.1301 10.1.13.1                            HMM    L,
       0          Local
            L3-Info: 3003001
            Sent To: BGP
            ESI : 0300.0034.5634.5600.0001
```

BGP:

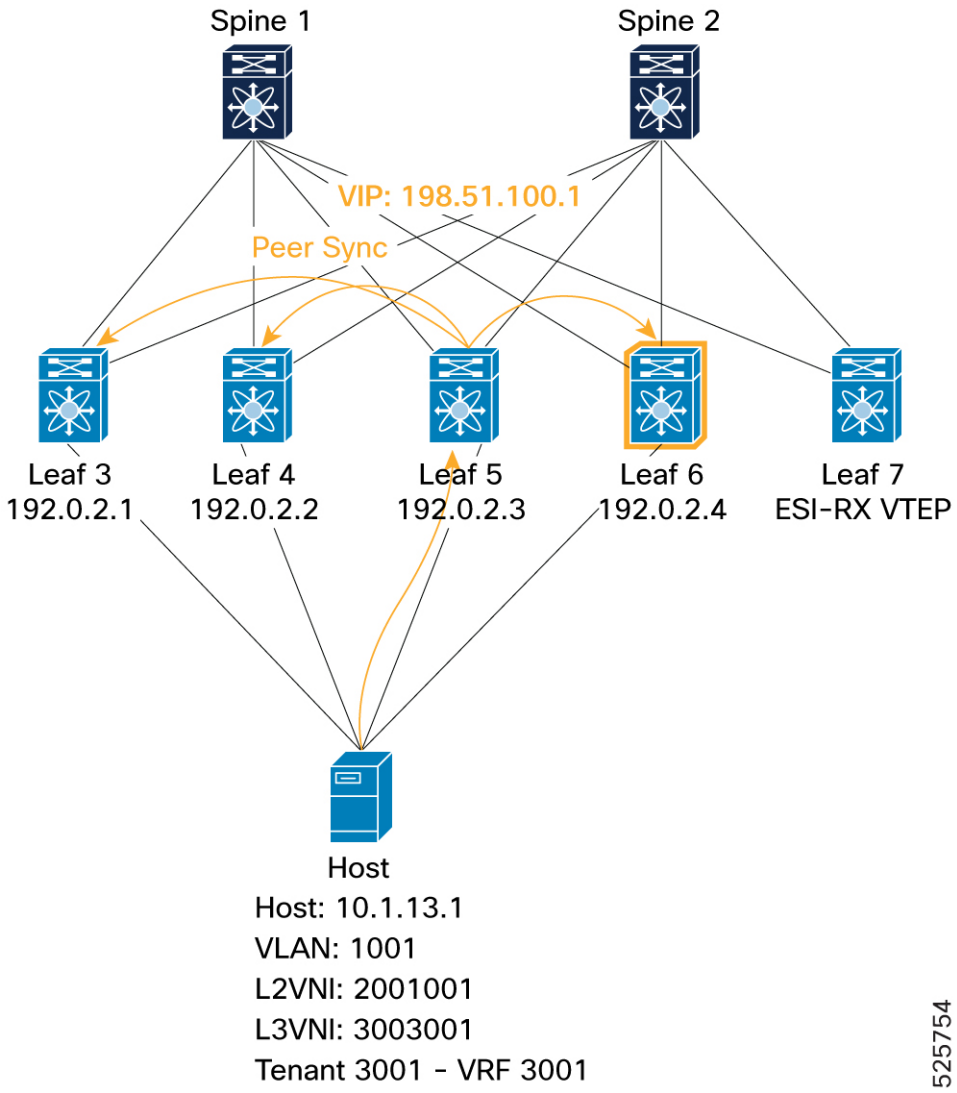
```
Leaf5# show bgp l2vpn evpn 10.1.13.1

Route Distinguisher: 192.0.2.10:33768    (L2VNI 2001001)
<SNIP>
  Advertised path-id 1
  Path type: local, path is valid, is best path, no labeled nexthop
  AS-Path: NONE, path locally originated
    192.0.2.3 (metric 0) from 0.0.0.0 (192.0.2.10)
      Origin IGP, MED not set, localpref 100, weight 32768
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 ENCAP:8 Router MAC:4880.0290.01af
      ESI: 0300.0034.5634.5600.0001

  Path-id 1 advertised to peers:
    192.0.2.21         192.0.2.22
```

- **ESI-MH Peer VTEP (Peer Sync)**: Shows successful peer synchronization of an EVPN Type-2 route for host 10.1.13.1 in one of 4-Way ESI MH node (example, Leaf 6). This is visible by the `mh_peer_synced` flag and "PS" flag.

Figure 98: Verification of VIP mode - ESI-MH Peer VTEP (Peer Sync)



BGP:

```
Leaf6# show bgp l2vpn evpn 10.1.13.1

Route Distinguisher: 192.0.16.1:33768    (L2VNI 2001001)
<SNIP>
  Path type: internal, path is valid, not best reason: Local ESI, mh_peer_synced,
no labeled nexthop, in rib
            Imported from
192.0.2.10:33768:[2]:[0]:[0]:[48]:[0010.0100.1301]:[32]:[10.1.13.1]/272
  AS-Path: NONE, path sourced internal to AS
    192.0.2.3 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 ENCAP:8 Router MAC:4880.0290.01af
      Originator: 192.0.2.10 Cluster list: 192.0.2.21
      ESI: 0300.0034.5634.5600.0001
```

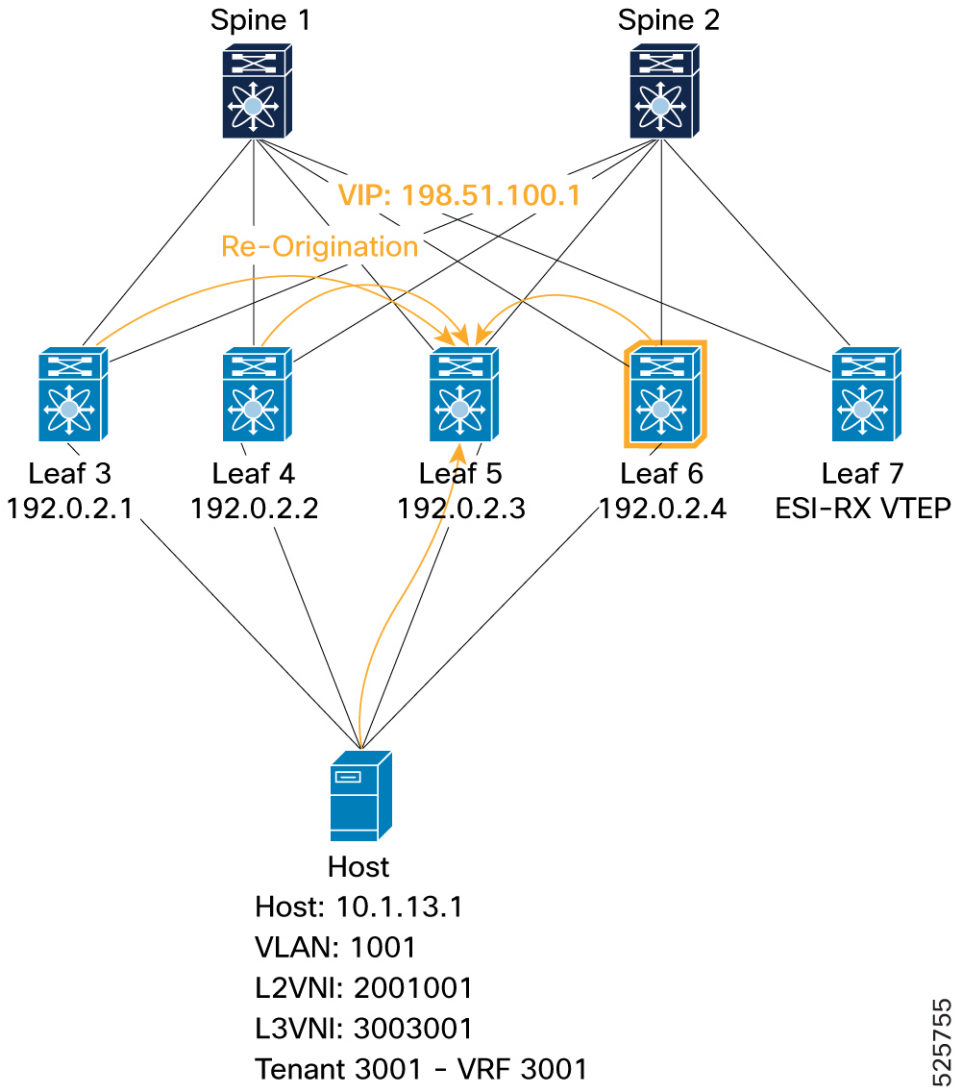
L2RIB:

```
Leaf6# show l2route evpn mac-ip evi 1001 detail

Topology    Mac Address    Host IP                                   Prod   Flags
       Seq No     Next-Hops
----------- -------------- ----------------------------------- ------
----------------- ---------- --------------------------
1001        0010.0100.1301 10.1.13.1                           BGP    PS,
       0          192.0.2.3 (Label: 2001001)
            Sent To: ARP
            ESI : 0300.0034.5634.5600.0001
            Port-Channel Info: Po1
            Encap-type:1
```

- **ESI-MH Remote VTEP (Re-Origination)**: Shows that an ESI member (Leaf 6) re-originates the Type-2 route for the locally learned host (even if learned via peer sync) to remote VTEPs. This is visible by a "locally originated" path from Leaf 6 itself, alongside the "Peer Synced" entry.

Figure 99: Verification of VIP mode - ESI-MH Remote VTEP (Re-Origination)



BGP:

```
Leaf6# show bgp l2vpn evpn 10.1.13.1
Route Distinguisher: 192.0.16.1:33768    (L2VNI 2001001)
<SNIP>
 Advertised path-id 1
 Path type: local, path is valid, is best path, no labeled nexthop
  AS-Path: NONE, path locally originated
    192.0.2.4 (metric 0) from 0.0.0.0 (192.0.16.1)
      Origin IGP, MED not set, localpref 100, weight 32768
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 SOO:192.0.16.1:256 ENCAP:8
          Router MAC:f839.184d.48c7
      ESI: 0300.0034.5634.5600.0001
 Path-id 1 advertised to peers:
    192.0.2.21         192.0.2.22
```

L2RIB:

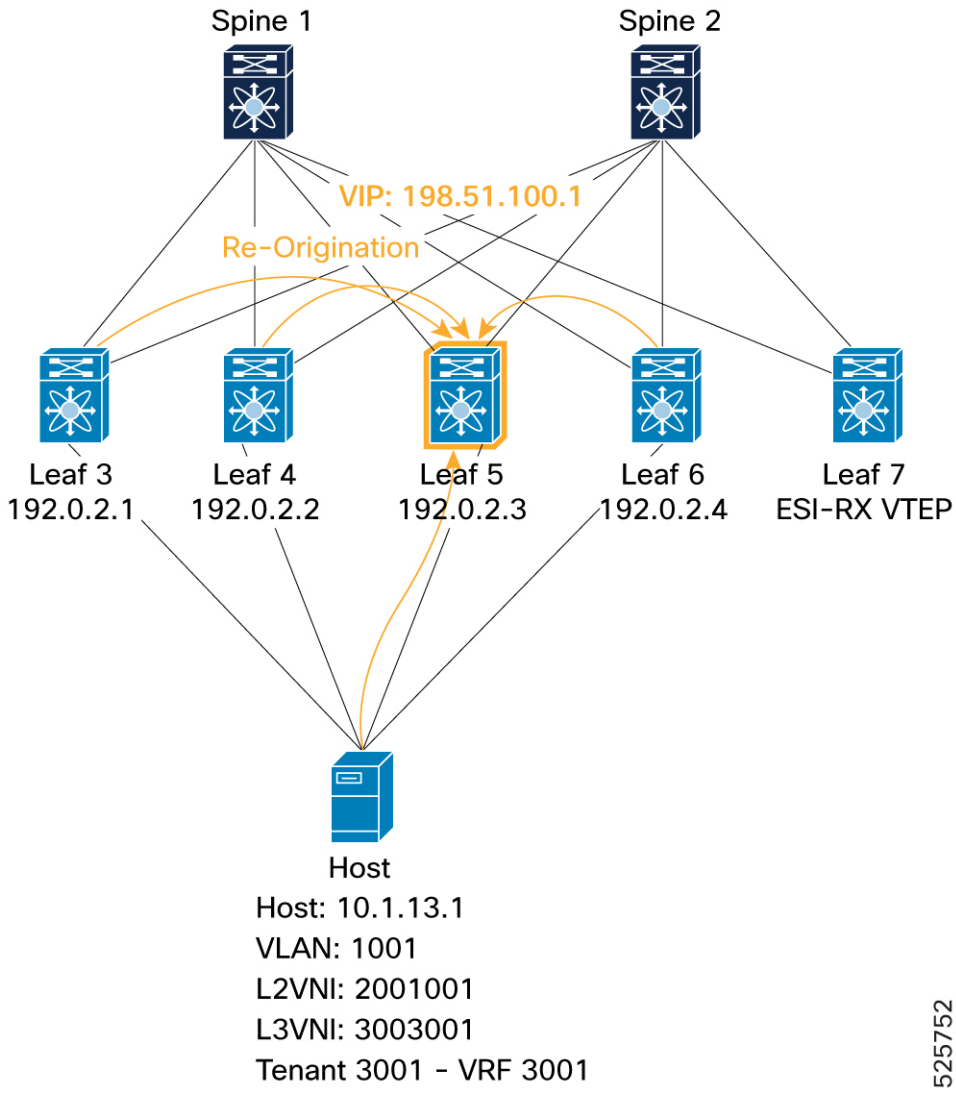
```
Leaf6# show l2route evpn mac-ip evi 1001 detail

Topology    Mac Address    Host IP                                 Prod   Flags
       Seq No     Next-Hops
----------- -------------- ------------------------------------ ------
----------------- ---------- --------------------------
1001        0010.0100.1301 10.1.13.1                               BGP    PS,
       0          192.0.2.4 (Label: 2001001)
            Sent To: ARP
            ESI : 0300.0034.5634.5600.0001
            Port-Channel Info: Po1
            Encap-type:1

1001        0010.0100.1301 10.1.13.1                               HMM    RO,
       0         Local
            L3-Info: 3003001
            Sent To: BGP
            ESI : 0300.0034.5634.5600.0001
```

- **Re-Origination Verification (Local VTEP)**: Verifies that the local VTEP (Leaf 5) receives re-originated Type-2 routes from other ESI members, marked with the `mh_peer_reoriginated` flag in BGP.

Figure 100: Verification of VIP mode - ESI-MH Local VTEP (Re-Origination)




```
Leaf5# show bgp l2vpn evpn 10.1.13.1

Route Distinguisher: 192.0.2.10:33768    (L2VNI 2001001)
<SNIP>
  Path type: internal, path is valid, not best reason: Router Id,
mh_peer_reoriginated, no labeled nexthop
             Imported from
192.0.14.1:33768:[2]:[0]:[0]:[48]:[0010.0100.1301]:[32]:[10.1.13.1]/272
  AS-Path: NONE, path sourced internal to AS
    192.0.2.2 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 SOO:192.0.14.1:256 ENCAP:8
          Router MAC:f839.1867.df6b
      Originator: 192.0.14.1 Cluster list: 192.0.2.21
      ESI: 0300.0034.5634.5600.0001

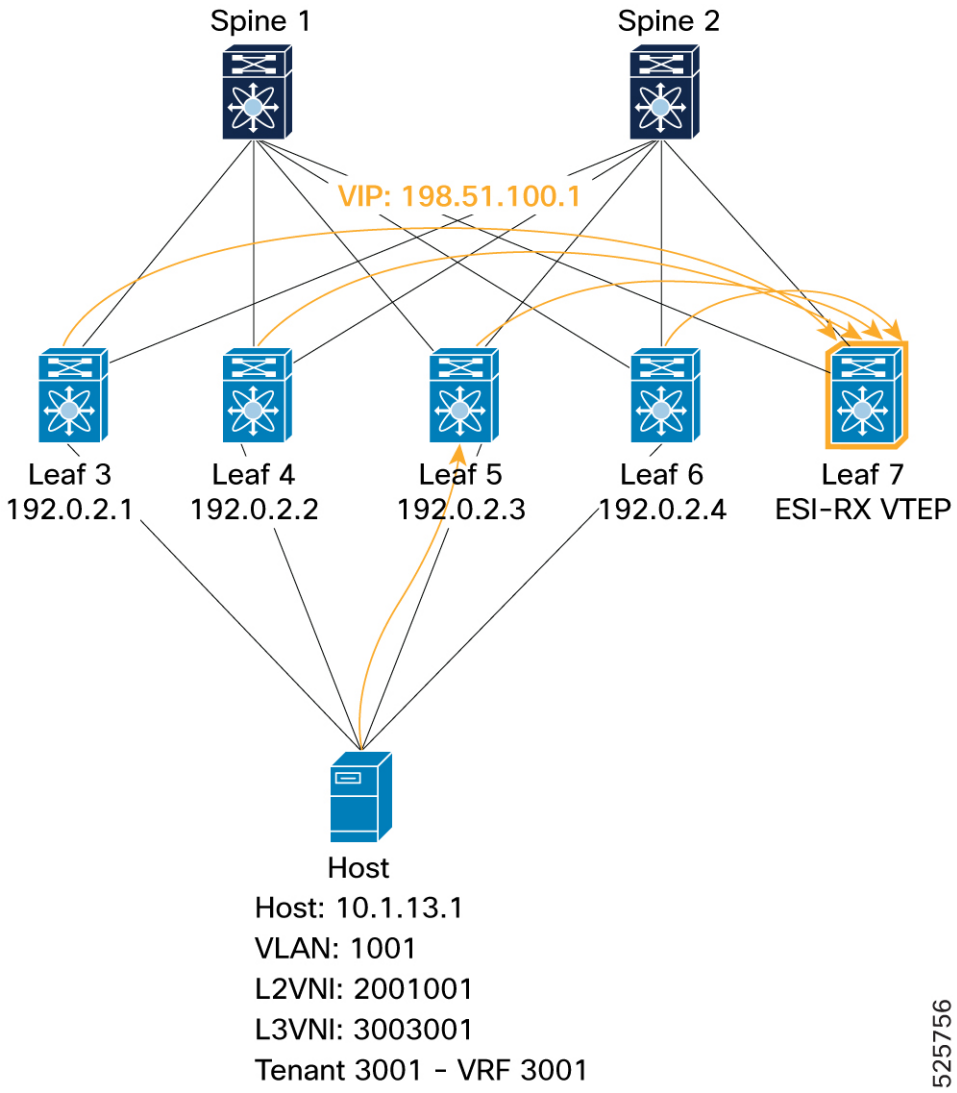
  Path type: internal, path is valid, not best reason: Router Id,
```

```
mh_peer_reoriginated, no labeled nexthop
            Imported from
192.0.16.1:33768:[2]:[0]:[0]:[48]:[0010.0100.1301]:[32]:[10.1.13.1]/272
  AS-Path: NONE, path sourced internal to AS
    192.0.2.4 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 SOO:192.0.16.1:256 ENCAP:8
          Router MAC:f839.184d.48c7
      Originator: 192.0.16.1 Cluster list: 192.0.2.21
      ESI: 0300.0034.5634.5600.0001

  Path type: internal, path is valid, not best reason: Local ESI,
mh_peer_reoriginated, no labeled nexthop
            Imported from
192.0.13.1:33768:[2]:[0]:[0]:[48]:[0010.0100.1301]:[32]:[10.1.13.1]/272
  AS-Path: NONE, path sourced internal to AS
    192.0.2.1 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 SOO:192.0.13.1:256 ENCAP:8
          Router MAC:4880.0290.0727
      Originator: 192.0.13.1 Cluster list: 192.0.2.21
      ESI: 0300.0034.5634.5600.0001
```

- **To Remote VTEP (ESI-RX, Leaf 7)**: Confirms that the remote ESI-RX VTEP (Leaf 7) receives the Type-2 route for the host (10.1.13.1) from all four ESI cluster members. All these paths consistently point to the VIP (198.51.100.1) as the egress next-hop. The L2RIB, MAC address table, IP route table, and FIB on Leaf 7 all show the host reachable via the VIP (198.51.100.1) through the NVE interface.

Figure 101: Verification of VIP mode - Remote VTEP (ESI-RX)



BGP:

```
Leaf7# show bgp l2vpn evpn 10.1.13.1

  Path type: internal, path is valid, not best reason: Router Id, no labeled nexthop,
 anycast NH
             Imported from
192.0.14.1:33768:[2]:[0]:[0]:[48]:[0010.0100.1301]:[32]:[10.1.13.1]/272
  AS-Path: NONE, path sourced internal to AS
    198.51.100.1 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 SOO:192.0.14.1:256 ENCAP:8
          Router MAC:f839.1867.df6b
      Originator: 192.0.14.1 Cluster list: 192.0.2.21
      ESI: 0300.0034.5634.5600.0001

  Path type: internal, path is valid, not best reason: Router Id, no labeled nexthop,
 anycast NH
```

```
            Imported from
192.0.16.1:33768:[2]:[0]:[0]:[48]:[0010.0100.1301]:[32]:[10.1.13.1]/272
  AS-Path: NONE, path sourced internal to AS
    198.51.100.1 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 SOO:192.0.16.1:256 ENCAP:8
          Router MAC:f839.184d.48c7
      Originator: 192.0.16.1 Cluster list: 192.0.2.21
      ESI: 0300.0034.5634.5600.0001

  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop, in rib,
anycast NH
            Imported from
192.0.13.1:33768:[2]:[0]:[0]:[48]:[0010.0100.1301]:[32]:[10.1.13.1]/272
  AS-Path: NONE, path sourced internal to AS
    198.51.100.1 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 SOO:192.0.13.1:256 ENCAP:8
          Router MAC:4880.0290.0727
      Originator: 192.0.13.1 Cluster list: 192.0.2.21
      ESI: 0300.0034.5634.5600.0001

  Path type: internal, path is valid, not best reason: Router Id, no labeled nexthop,
 anycast NH
            Imported from
192.0.15.1:33768:[2]:[0]:[0]:[48]:[0010.0100.1301]:[32]:[10.1.13.1]/272
  AS-Path: NONE, path sourced internal to AS
    198.51.100.1 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 ENCAP:8 Router MAC:4880.0290.01af
      Originator: 192.0.15.1 Cluster list: 192.0.2.21
      ESI: 0300.0034.5634.5600.0001
```

L2RIB:

```
Leaf7# show l2route evpn mac-ip evi 1001 detail

Topology    Mac Address    Host IP                                 Prod   Flags
       Seq No     Next-Hops
----------- -------------- ------------------------------------ ------
----------------- ---------- --------------------------
1001        0010.0100.1301 10.1.13.1                               BGP    --
       0         198.51.100.1 (Label: 2001001)
            Sent To: ARP
            Encap-type:1
```

L2FM

```
Leaf7# show mac address-table vlan 1001
VLAN     MAC Address      Type      age       Secure   NTFY   Ports
---------+-----------------+--------+---------+------+----+------------------
C 1001     0010.0100.1301   dynamic  NA         F      F    nve1(198.51.100.1)
```

URIB:

```
Leaf7# show ip route 10.1.13.1 detail vrf 3001

10.1.13.1/32, ubest/mbest: 1/0
    Extended Community: 0x1b 1c 01 03 65 01 0d 01 01 00 00 00 00 00 00 00 00 03
00 00 34 56 34 56 00 00 01
    *via 198.51.100.1%default, [200/0], 00:56:42, bgp-1, internal, tag 1, segid:
```

```
3003001 tunnelid: 0x67018601 encap: VXLAN

          BGP-EVPN: VNI=3003001 (EVPN)
         client-specific data: 579
         recursive next hop: 198.51.100.1/32%default
         extended route information: BGP origin AS 1 BGP peer AS 1
```

FIB:

```
Leaf7# show forwarding ipv4 route 10.1.13.1 vrf 3001

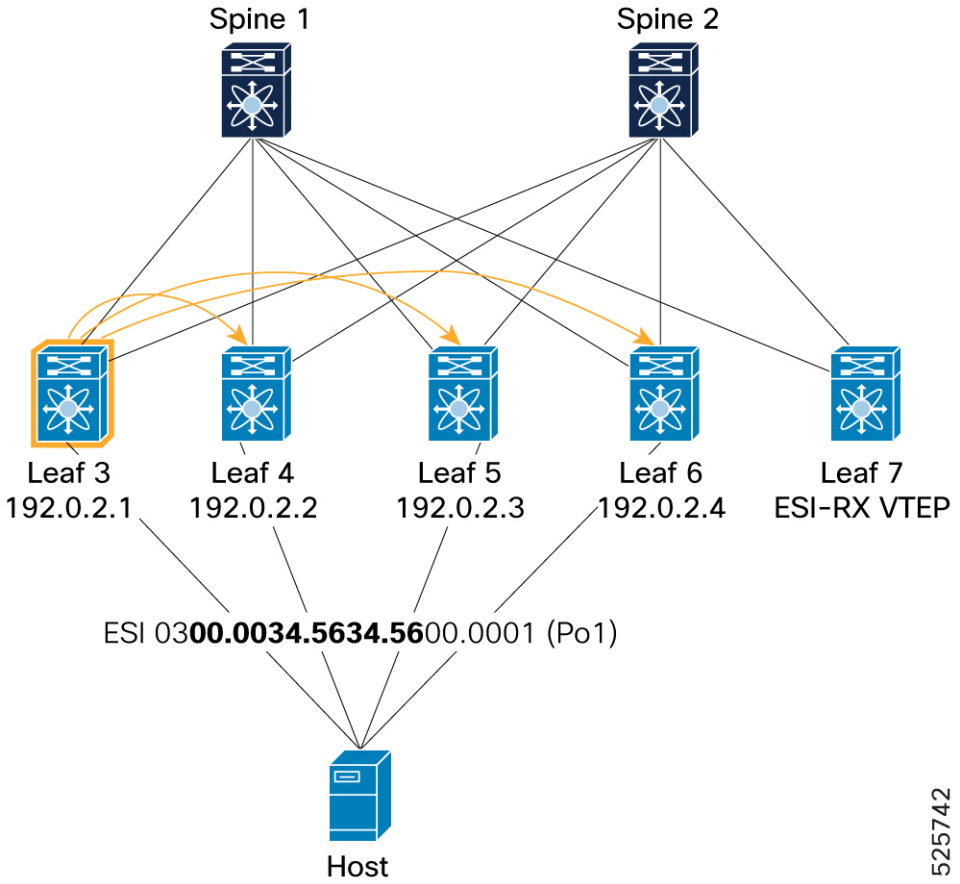
──────────────┼──────────────────────────────────┼──────────────┼─────────────┼──────────────
Prefix            | Next-hop                          | Interface
| Labels          | Partial Install
──────────────┼──────────────────────────────────┼──────────────┼─────────────┼──────────────
10.1.13.1/32         198.51.100.1                      nve1                   vni:
   3003001
```

Verification of PIP mode

This section details the verification steps for ESI multi-homing PIP Mode. The validation covers the advertisement and synchronization of EVPN Type-1 (EAD-ES) and Type-2 (MAC-IP) routes across local and remote VTEPs using a PIP.

1. EVPN EAD-ES route advertisements
   - **Local route (from Leaf3)**: Shows how a local Leaf (Leaf 3) advertises its Type-1 EAD-ES route for ESI 0300.0034.5634.5600.0001 to its BGP peers (Spines). The command `show l2route evpn ead es` on Leaf3 confirms the local route (Prod: VXLAN) with its next-hop. The command `show bgp l2vpn evpn route-type 1` on Leaf3 shows the locally originated Type-1 route with the correct ESI and next-hop.

Figure 102: Verification of PIP mode – Local route



Example for BGP component:

```
Leaf3# show bgp l2vpn evpn route-type 1

Route Distinguisher: 192.0.13.1:7817    (EAD-ES [0300.0034.5634.5600.0001 7817])
BGP routing table entry for [1]:[0300.0034.5634.5600.0001]:[0xffffffff]/152, version
 154
Paths: (1 available, best #1)
Flags: (0x000002) (high32 00000000) on xmit-list, is not in l2rib/evpn
Multipath: eBGP iBGP

  Advertised path-id 1
  Path type: local, path is valid, is best path, no labeled nexthop
  AS-Path: NONE, path locally originated
    192.0.2.1 (metric 0) from 0.0.0.0 (192.0.13.1)
      Origin IGP, MED not set, localpref 100, weight 32768
      Received label 0
      Extcommunity: RT:1:2001001 RT:1:2001002 ENCAP:8 ESI:0:000000

  Path-id 1 advertised to peers:
    192.0.2.21          192.0.2.22
```

Example for L2RIB component:

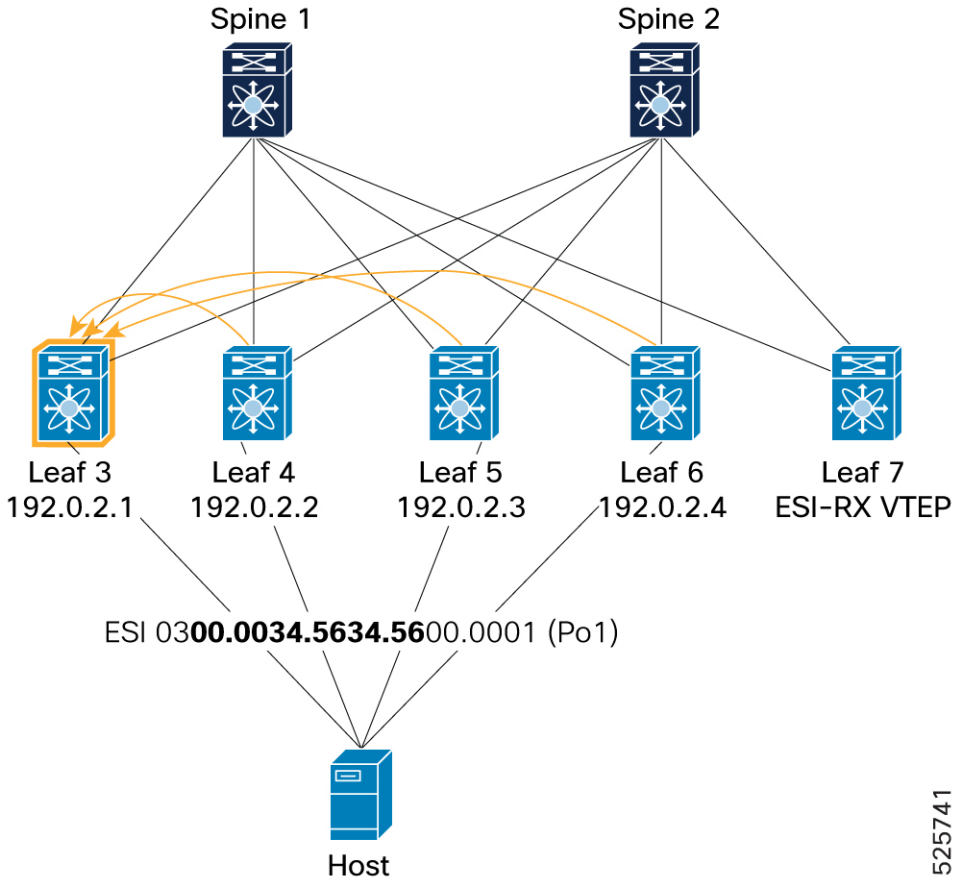
```
Leaf3# show l2route evpn ead es

Topology ID   Prod   ESI                        Sent To      Num PLs      Flags
```

```
------------ ------ ------------------------ ---------- -------- ----------
4294967294    VXLAN  0300.0034.5634.5600.0001   BGP        0       A
              Next-Hops: 192.0.2.1
```

- **Remote route (received by Leaf 3)**: Shows how a Leaf (Leaf 3) receives Type-1 EAD-ES routes from its ESI peers (Leaf 4, Leaf 5, and Leaf 6) within a 4-way ESI cluster. The command `show bgp l2vpn evpn route-type 1` on Leaf 3 displays multiple paths for the same ESI, originating from different ESI peers (192.0.2.2, 192.0.2.3, 192.0.2.4), indicating successful reception of remote routes. The command `show l2route evpn ead es` on Leaf 3 confirms the presence of these remote ESI routes (Prod: BGP) with multiple next-hops.

Figure 103: Verification of PIP mode – Remote route



Example for BGP component:

```
Leaf3# show bgp l2vpn evpn route-type 1

Route Distinguisher: 192.0.13.1:65534    (L2VNI 0)
BGP routing table entry for [1]:[0300.0034.5634.5600.0001]:[0xffffffff]/152, version
 3434
Paths: (3 available, best #3)
Flags: (0x000012) (high32 00000000) on xmit-list, is in l2rib/evpn, is not in HW
Multipath: eBGP iBGP

  Path type: internal, path is valid, not best reason: Router Id, multipath, no
labeled nexthop, in rib
             Imported from
192.0.2.10:7817:[1]:[0300.0034.5634.5600.0001]:[0xffffffff]/152
```

```
  AS-Path: NONE, path sourced internal to AS
    192.0.2.3 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Extcommunity: RT:1:2001001 RT:1:2001002 ENCAP:8 ESI:0:000000
      Originator: 192.0.2.10 Cluster list: 192.0.2.21

  Path type: internal, path is valid, not best reason: Router Id, multipath, no
labeled nexthop, in rib
             Imported from
192.0.16.1:7817:[1]:[0300.0034.5634.5600.0001]:[0xffffffff]/152
  AS-Path: NONE, path sourced internal to AS
    192.0.2.4 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Extcommunity: RT:1:2001001 RT:1:2001002 ENCAP:8 ESI:0:000000
      Originator: 192.0.16.1 Cluster list: 192.0.2.21

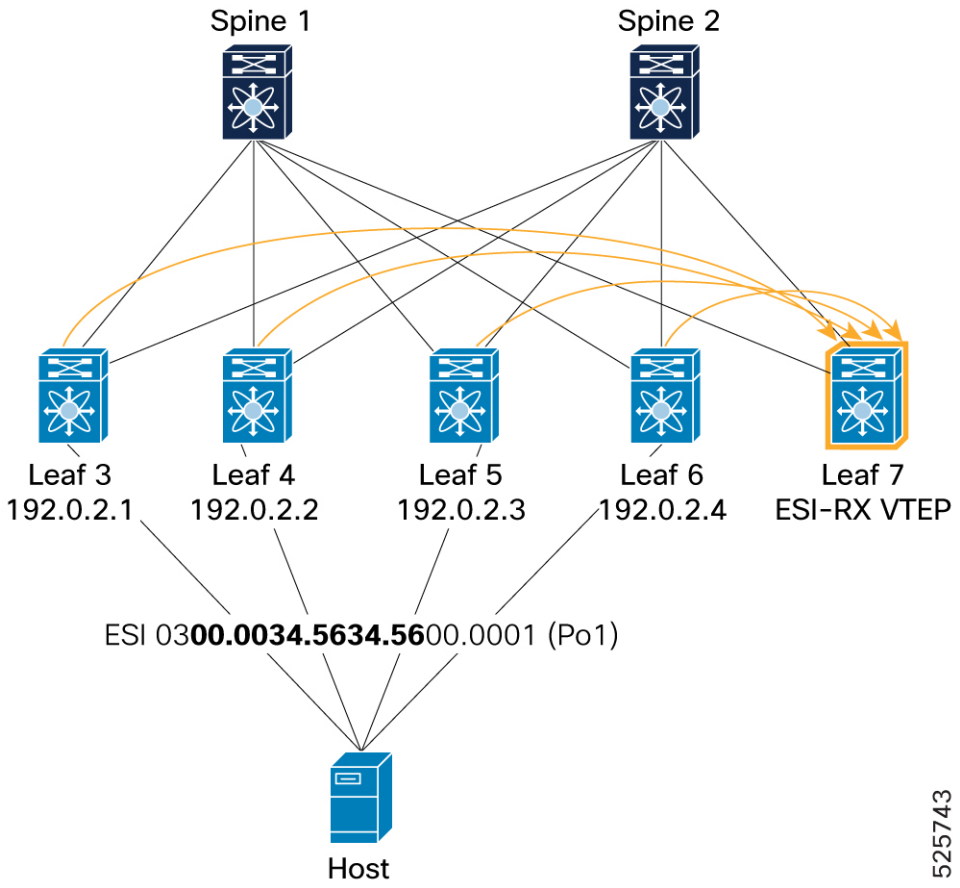
  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop, in rib
             Imported from
192.0.14.1:7817:[1]:[0300.0034.5634.5600.0001]:[0xffffffff]/152
  AS-Path: NONE, path sourced internal to AS
    192.0.2.2 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Extcommunity: RT:1:2001001 RT:1:2001002 ENCAP:8 ESI:0:000000
      Originator: 192.0.14.1 Cluster list: 192.0.2.21
```

Example for L2RIB component:

```
Leaf3# show l2route evpn ead es
Topology ID   Prod   ESI                        Sent To    Num PLs  Flags
------------- ------ -------------------------- ---------- -------- ----------
4294967294    BGP    0300.0034.5634.5600.0001   -          0        A
              Next-Hops: 192.0.2.2
                         192.0.2.3
                         192.0.2.4
```

- **Remote VTEP (ESI-RX, Leaf 7)**: Verifies that a remote VTEP (Leaf 7), which is an ESI-RX VTEP, correctly receives Type-1 EAD-ES routes from all members of the 4-way ESI MH cluster (Leaf 3, Leaf4, Leaf 5, and Leaf 6). The command `show bgp l2vpn evpn route-type 1` on Leaf 7 shows four available paths for the ESI, each originating from a different Leaf in the ESI cluster (192.0.2.1, 192.0.2.2, 192.0.2.3, 192.0.2.4). The commands `show l2route evpn ead es` and `show l2route evpn path-list all detail` on Leaf 7 confirm the multiple next-hops for the ESI, indicating proper ECMP path installation.

Figure 104: Verification of PIP mode – Remote VTEP (ESI-RX)



Example for BGP component:

```
Leaf7# show bgp l2vpn evpn route-type 1

Route Distinguisher: 192.0.17.1:65534    (L2VNI 0)
BGP routing table entry for [1]:[0300.0034.5634.5600.0001]:[0xffffffff]/152, version
 2645
Paths: (4 available, best #3)
Flags: (0x000012) (high32 00000000) on xmit-list, is in l2rib/evpn, is not in HW
Multipath: eBGP iBGP

  Path type: internal, path is valid, not best reason: Router Id, multipath, no
labeled nexthop, in rib
             Imported from
192.0.2.10:7817:[1]:[0300.0034.5634.5600.0001]:[0xffffffff]/152
  AS-Path: NONE, path sourced internal to AS
    192.0.2.3 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Extcommunity: RT:1:2001001 RT:1:2001002 ENCAP:8 ESI:0:000000
      Originator: 192.0.2.10 Cluster list: 192.0.2.21

  Path type: internal, path is valid, not best reason: Router Id, multipath, no
labeled nexthop, in rib
             Imported from
192.0.16.1:7817:[1]:[0300.0034.5634.5600.0001]:[0xffffffff]/152
```

```
  AS-Path: NONE, path sourced internal to AS
    192.0.2.4 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Extcommunity: RT:1:2001001 RT:1:2001002 ENCAP:8 ESI:0:000000
      Originator: 192.0.16.1 Cluster list: 192.0.2.21

  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop, in rib
             Imported from
192.0.13.1:7817:[1]:[0300.0034.5634.5600.0001]:[0xffffffff]/152
  AS-Path: NONE, path sourced internal to AS
    192.0.2.1 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Extcommunity: RT:1:2001001 RT:1:2001002 ENCAP:8 ESI:0:000000
      Originator: 192.0.13.1 Cluster list: 192.0.2.21

  Path type: internal, path is valid, not best reason: Router Id, multipath, no
labeled nexthop, in rib
             Imported from
192.0.14.1:7817:[1]:[0300.0034.5634.5600.0001]:[0xffffffff]/152
  AS-Path: NONE, path sourced internal to AS
    192.0.2.2 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 0
      Extcommunity: RT:1:2001001 RT:1:2001002 ENCAP:8 ESI:0:000000
      Originator: 192.0.14.1 Cluster list: 192.0.2.21
```

Example for L2RIB component:

```
Leaf7# show l2route evpn ead es

Topology ID   Prod   ESI                        Sent To    Num PLs  Flags
------------- ------ -------------------------- ---------- -------- ----------
4294967294    BGP    0300.0034.5634.5600.0001   -          2        A
                     Next-Hops: 192.0.2.1
                                192.0.2.2
                                192.0.2.3
                                192.0.2.4



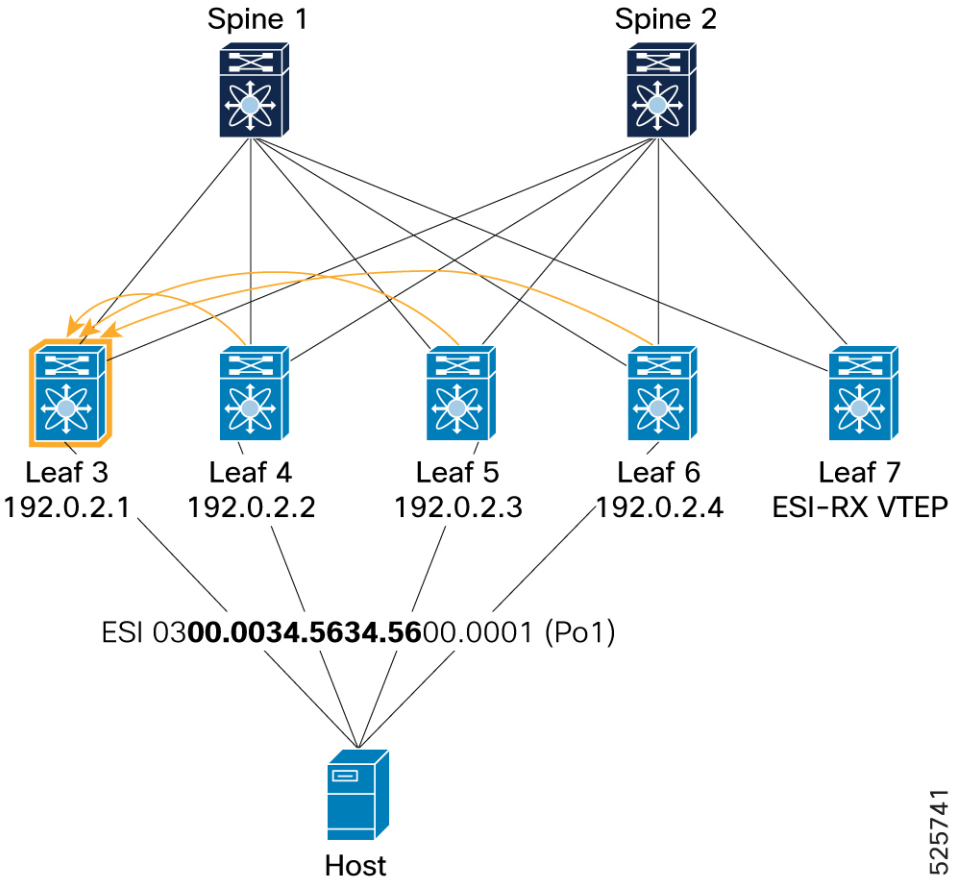
Leaf7# show l2route evpn path-list all detail

Topology ID  Prod   ESI                       ECMP Label Flags  Client Ctx  MACs
    Sent To
------------ ------ ------------------------- ---------- ------ ----------- ----------
 ----------
1001         None   0300.0034.5634.5600.0001  3          A      0           15
    UFDM
                    CP Next-Hops: 192.0.2.1 , 192.0.2.2 , 192.0.2.3 , 192.0.2.4
                    Gbl EAD Next-Hops: 192.0.2.1 (1,R)
                                       192.0.2.2 (3,R)
                                       192.0.2.3 (2,R)
                                       192.0.2.4 (5,R)
                    Res Next-Hops: 192.0.2.1
                                   192.0.2.2
                                   192.0.2.3
                                   192.0.2.4
                    Bkp Next-Hops:
                    Res Next-Hops from UFDM: 192.0.2.1
                                             192.0.2.2
                                             192.0.2.3
```

```
                                                    192.0.2.4
                Bkp Next-Hops from UFDM:
```

2. Verification of NVE Ethernet Segment

Figure 105: PIP mode - NVE Ethernet Segment verification



- NVE Ethernet Segment verification: Validates the state and configuration of the NVE Ethernet Segment on Leaf3. show nve ethernet-segment summary and show nve ethernet-segment on Leaf3 confirm the ESI 0300.0034.5634.5600.0001 is "Up," associated with port-channel1, and lists details like active VLANs/VNIs, DF (Designated Forwarder) list (192.0.2.1, 192.0.2.2, 192.0.2.3, 192.0.2.4), and ESI type.

```
Leaf3# show nve ethernet-segment summary
ESI                             Parent interface   ES State
------------------------------  ------------------ ----------
0300.0034.5634.5600.0001        port-channel1      Up


Leaf3# show nve ethernet-segment

ESI: 0300.0034.5634.5600.0001
   Parent interface: port-channel1
  ES State: Up
  Port-channel state: Up
  NVE Interface: nve1
```

```
 NVE State: Up
 Host Learning Mode: control-plane
Active Vlans: 1001-1002
 DF Vlans:
 Active VNIs: 2001001-2001002
DF BDs: N/A
 DF VNIs: N/A
 Number of ES members: 4
My ordinal: 0
DF timer start time: 00:00:00
DF timer expiry: 17:11:42
Config State: config-applied
DF List: 192.0.2.1 192.0.2.2 192.0.2.3 192.0.2.4
Bounce peer : 192.0.2.2
ES route added to L2RIB: True
EAD/ES routes added to L2RIB: True
EAD/EVI route timer age: not running [Disabled]
EAD/EVI timer duration: 00:05:00
ESI type: Ether-segment
ESI DF election mode: Per-flow
```

- NVE Interface Detail: Provides detailed information about the NVE interface (nve1) on Leaf3, including its state, encapsulation (VXLAN), host learning mode (control-plane), source interface (loopback1 with IP 192.0.2.1), and ESI multihoming delay-restore times.

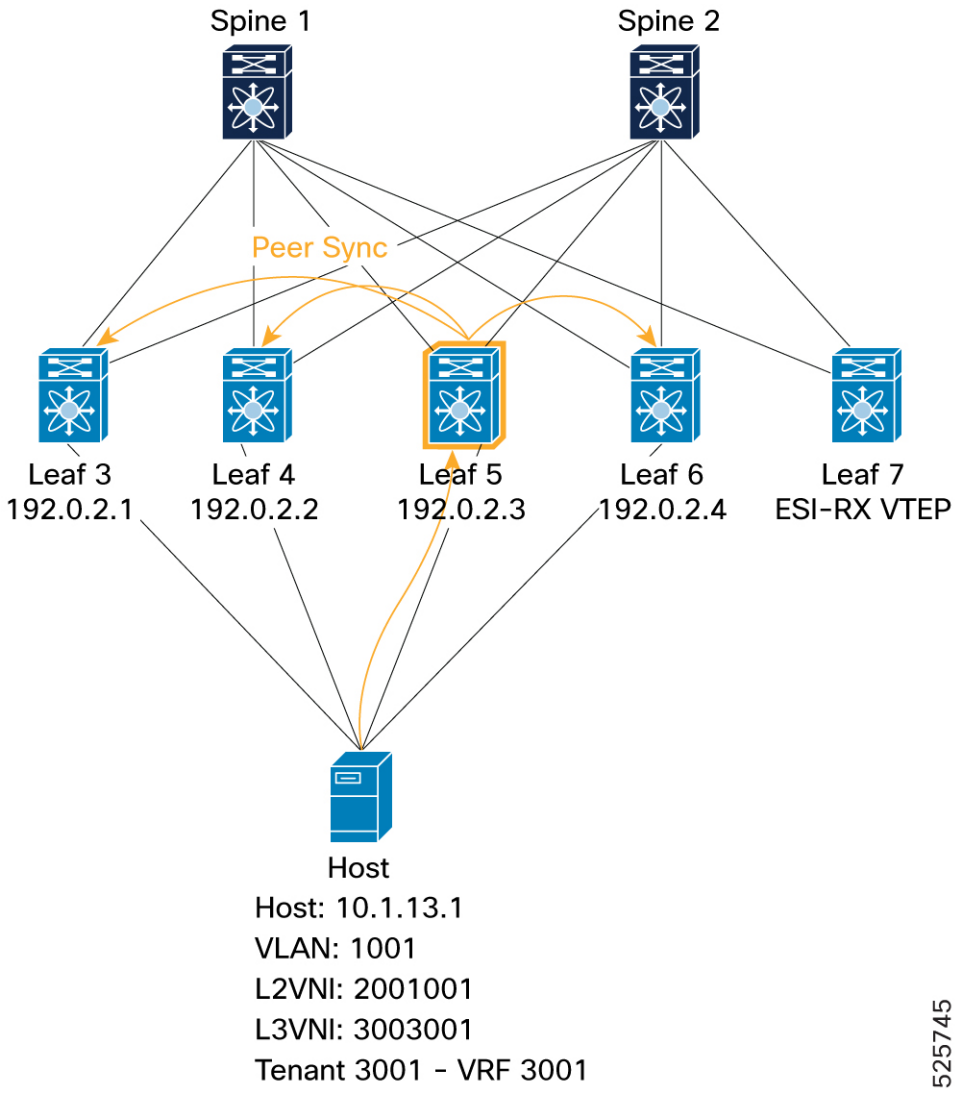
```
Leaf3# show nve interface nve 1 detail

Interface: nve1, State: Up, encapsulation: VXLAN
 VPC Capability: VPC-VIP-Only [not-notified]
 Local Router MAC: 4880.0290.0727
 Host Learning Mode: Control-Plane
 Source-Interface: loopback1 (primary: 192.0.2.1, secondary: 0.0.0.0)
 Source Interface State: Up
 Virtual RMAC Advertisement: No
 NVE Flags:
 Interface Handle: 0x49000001
 Source Interface hold-down-time: 180
 Source Interface hold-up-time: 30
 Remaining hold-down time: 0 seconds
 Virtual Router MAC: N/A
 Interface state: nve-intf-add-complete
 ESI multihoming delay-restore time: 180 seconds
 ESI multihoming delay-restore time left: 0 seconds
 ESI multihoming FRR anycast source IP: 203.0.113.1
 Fabric convergence time: 135 seconds
 Fabric convergence time left: 0 seconds
```

3. EVPN Type-2 route advertisement (MAC-IP)

   - **Local VTEP to ESI peers (peer sync)**: Shows how a local VTEP (Leaf5) advertises a Type-2 MAC-IP route for a host (MAC 0010.0100.1301, IP 10.1.13.1) connected to the ESI. The commands `show mac address-table` and `show ip arp` on Leaf 5 confirm the host's MAC and IP. The command `show l2route evpn mac-ip evi 1001 detail` on Leaf 5 shows the route as locally learned (Prod: HMM, Flags: L) and sent to BGP, with the associated ESI. The command `show bgp l2vpn evpn 10.1.13.1` on Leaf 5 confirms the locally originated Type-2 route, including its labels (L2VNI 2001001, L3VNI 3003001) and ESI.

Figure 106: Verification of PIP mode - Local VTEP to ESI peers (peer sync)



L2FM:

```
Leaf5# show mac address-table vlan 1001
VLAN     MAC Address      Type      age     Secure NTFY Ports
---------+-----------------+--------+---------+------+----+------------------
C 1001     0010.0100.1301   dynamic  NA         F      F    Po1
```

ARP:

```
Leaf5# show ip arp vrf 3001

Address         Age       MAC Address     Interface       Flags
10.1.13.1       00:18:43  0010.0100.1301  Vlan1001
```

L2RIB:

```
Leaf5# show l2route evpn mac-ip evi 1001 detail

Topology    Mac Address    Host IP                                   Prod   Flags
       Seq No     Next-Hops
```

```
----------- -------------- ------------------------------------ ------
----------------- ---------- --------------------------
1001        0010.0100.1301 10.1.13.1                            HMM    L,
       0          Local
            L3-Info: 3003001
            Sent To: BGP
            ESI : 0300.0034.5634.5600.0001
```

BGP:

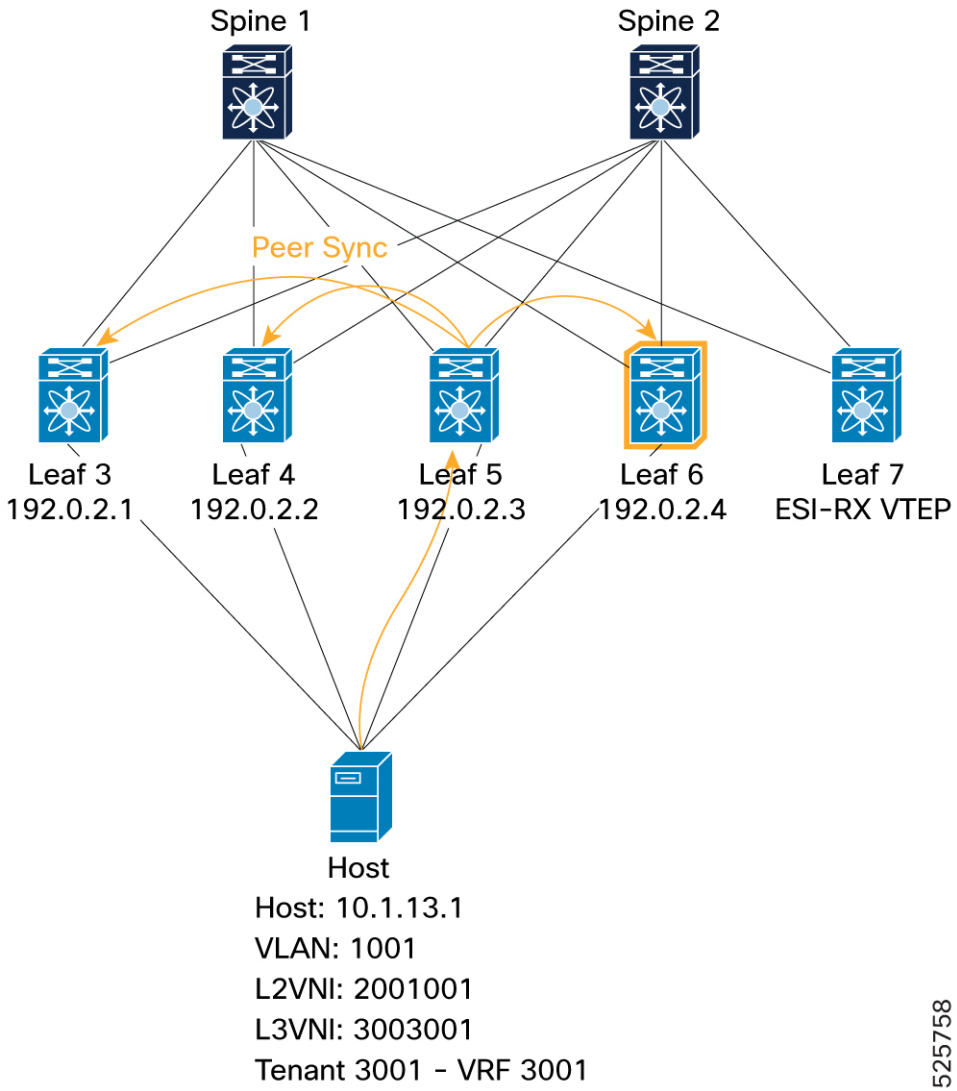
```
Leaf5# show bgp l2vpn evpn 10.1.13.1

Route Distinguisher: 192.0.2.10:33768    (L2VNI 2001001)
<SNIP>
  Advertised path-id 1
  Path type: local, path is valid, is best path, no labeled nexthop
  AS-Path: NONE, path locally originated
    192.0.2.3 (metric 0) from 0.0.0.0 (192.0.2.10)
      Origin IGP, MED not set, localpref 100, weight 32768
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 ENCAP:8 Router MAC:4880.0290.01af
      ESI: 0300.0034.5634.5600.0001

  Path-id 1 advertised to peers:
    192.0.2.21         192.0.2.22
```

- **ESI-MH Peer VTEP (Peer Sync)**: Shows how a peer VTEP (Leaf 6) receives the Type-2 route for the same host via peer synchronization. The command `show bgp l2vpn evpn 10.1.13.1` on Leaf 6 shows the route imported from the peer (192.0.2.10, which is Leaf5's IP), with mh_peer_synced flag. The command `show l2route evpn mac-ip evi 1001 detail` on Leaf 6 shows the route received via BGP (Prod: BGP, Flags: PS) with the next-hop pointing to Leaf5's VTEP IP.

Figure 107: Verification of PIP mode - ESI-MH Peer VTEP (Peer Sync)



BGP:

```
Leaf6# show bgp l2vpn evpn 10.1.13.1

Route Distinguisher: 192.0.16.1:33768    (L2VNI 2001001)
<SNIP>
  Path type: internal, path is valid, not best reason: Local ESI, mh_peer_synced,
no labeled nexthop, in rib
            Imported from
192.0.2.10:33768:[2]:[0]:[0]:[48]:[0010.0100.1301]:[32]:[10.1.13.1]/272
  AS-Path: NONE, path sourced internal to AS
    192.0.2.3 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 ENCAP:8 Router MAC:4880.0290.01af
      Originator: 192.0.2.10 Cluster list: 192.0.2.21
      ESI: 0300.0034.5634.5600.0001
```

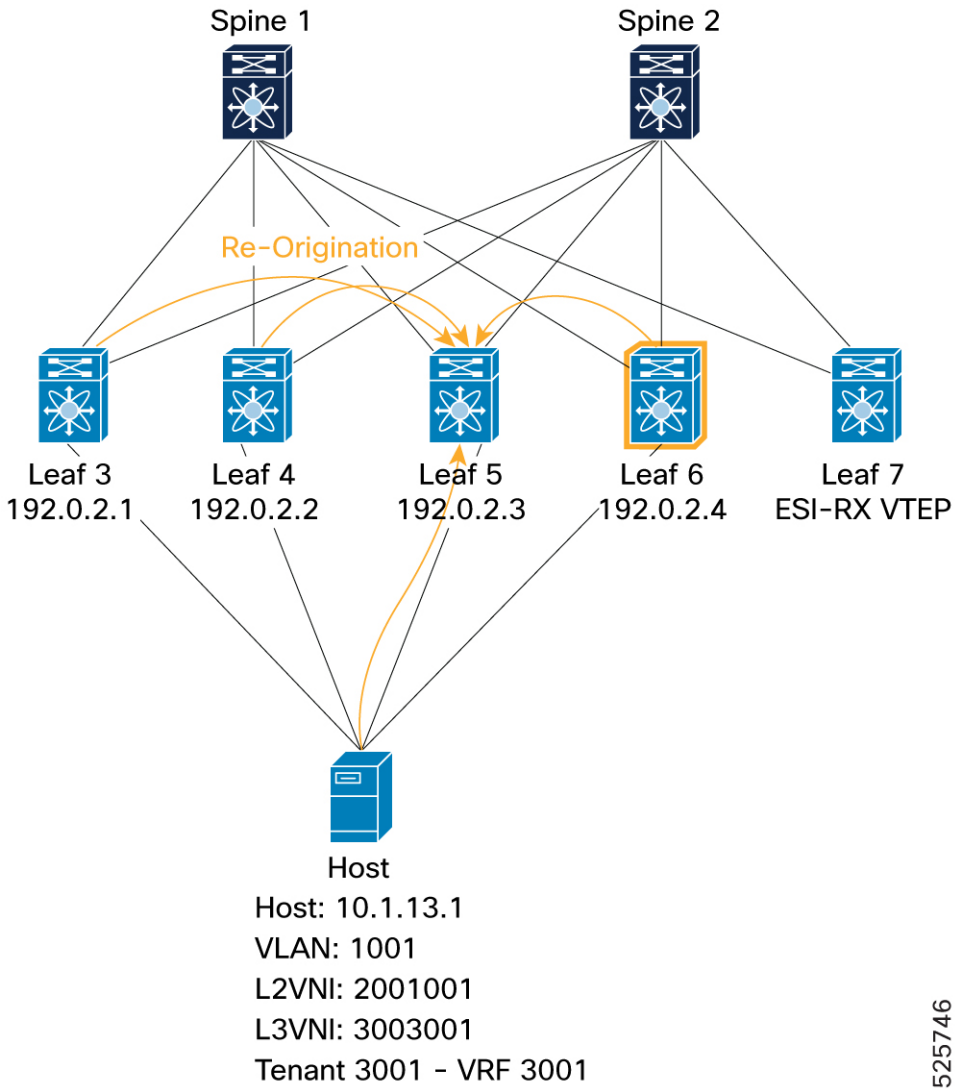
L2RIB:

```
Leaf6# show l2route evpn mac-ip evi 1001 detail

Topology    Mac Address    Host IP                                 Prod   Flags
       Seq No     Next-Hops
----------- -------------- ------------------------------------ ------
----------------- ---------- --------------------------
1001        0010.0100.1301 10.1.13.1                               BGP    PS,
       0          192.0.2.3 (Label: 2001001)
            Sent To: ARP
            ESI : 0300.0034.5634.5600.0001
            Port-Channel Info: Po1
            Encap-type:1
```

- **ESI-MH Remote VTEP (Re-Origination)**: Validates that a VTEP (Leaf6) within the ESI cluster re-originates the Type-2 route to remote VTEPs. The command `show bgp l2vpn evpn 10.1.13.1` on Leaf 6 shows the locally originated path for 10.1.13.1 with its own VTEP IP (192.0.2.4) as the next-hop, indicating re-origination. The command `show l2route evpn mac-ip evi 1001 detail` on Leaf 6 shows both the peer-synced (PS) route and the re-originated (RO) route.

Figure 108: Verification of PIP mode - ESI-MH Remote VTEP (Re-Origination)



BGP:

```
Leaf6# show bgp l2vpn evpn 10.1.13.1
Route Distinguisher: 192.0.16.1:33768    (L2VNI 2001001)
<SNIP>
 Advertised path-id 1
 Path type: local, path is valid, is best path, no labeled nexthop
  AS-Path: NONE, path locally originated
    192.0.2.4 (metric 0) from 0.0.0.0 (192.0.16.1)
      Origin IGP, MED not set, localpref 100, weight 32768
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 SOO:192.0.16.1:256 ENCAP:8
          Router MAC:f839.184d.48c7
      ESI: 0300.0034.5634.5600.0001
 Path-id 1 advertised to peers:
    192.0.2.21         192.0.2.22
```

L2RIB:

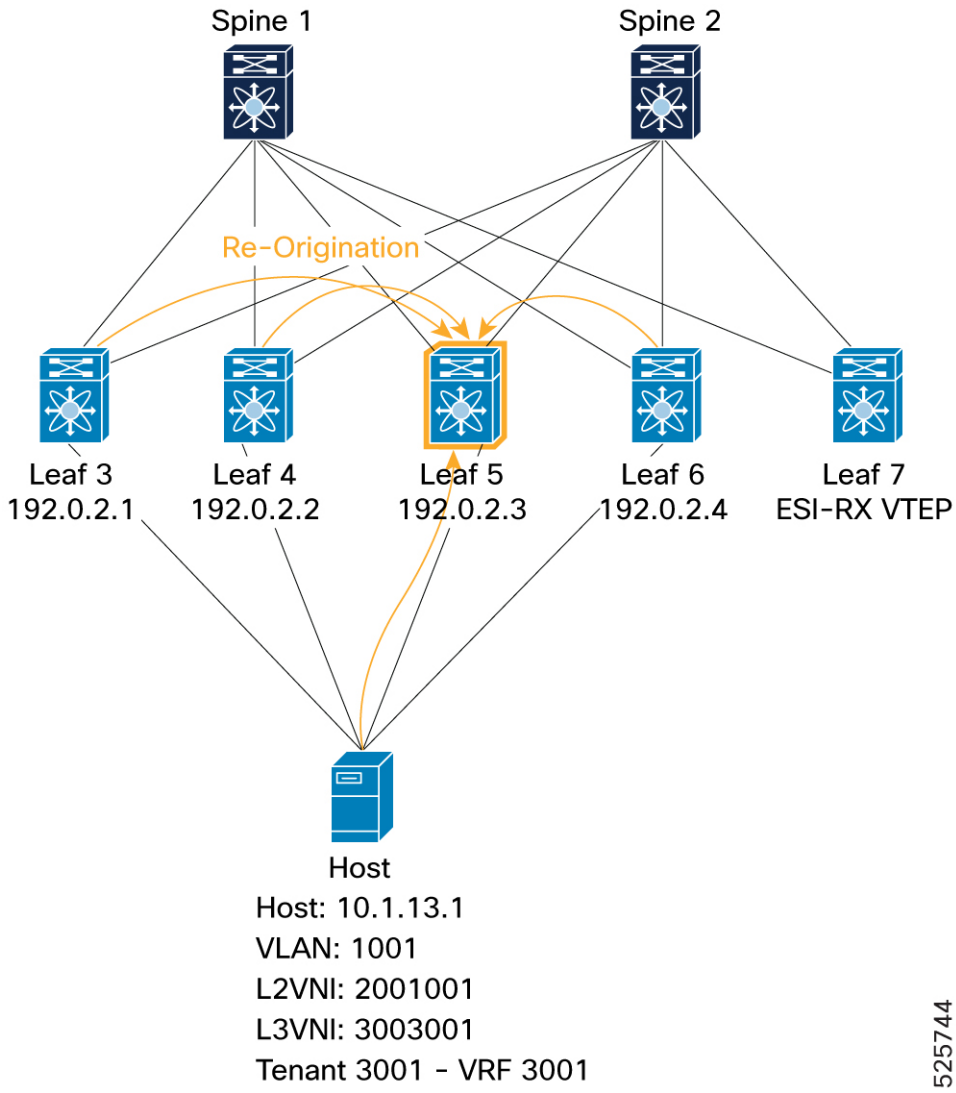
```
Leaf6# show l2route evpn mac-ip evi 1001 detail

Topology    Mac Address    Host IP                                 Prod   Flags
       Seq No     Next-Hops
----------- -------------- ------------------------------------ ------
----------------- ---------- --------------------------
1001        0010.0100.1301 10.1.13.1                               BGP    PS,
       0          192.0.2.4 (Label: 2001001)
            Sent To: ARP
            ESI : 0300.0034.5634.5600.0001
            Port-Channel Info: Po1
            Encap-type:1

1001        0010.0100.1301 10.1.13.1                               HMM    RO,
       0         Local
            L3-Info: 3003001
            Sent To: BGP
            ESI : 0300.0034.5634.5600.0001
```

- **Re-Origination Verification (Local VTEP)**: Shows how a local VTEP (Leaf 5) receives re-originated Type-2 routes from other ESI peers (Leaf 3, Leaf 4, and Leaf6). The command `show bgp l2vpn evpn 10.1.13.1` on Leaf 5 displays multiple paths for the host, imported from various ESI peers, each flagged as mh_peer_reoriginated.

Figure 109: Verification of PIP mode - ESI-MH Local VTEP (Re-Origination)




```
Leaf5# show bgp l2vpn evpn 10.1.13.1

Route Distinguisher: 192.0.2.10:33768    (L2VNI 2001001)
<SNIP>
  Path type: internal, path is valid, not best reason: Router Id,
mh_peer_reoriginated, no labeled nexthop
             Imported from
192.0.14.1:33768:[2]:[0]:[0]:[48]:[0010.0100.1301]:[32]:[10.1.13.1]/272
  AS-Path: NONE, path sourced internal to AS
    192.0.2.2 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 SOO:192.0.14.1:256 ENCAP:8
          Router MAC:f839.1867.df6b
      Originator: 192.0.14.1 Cluster list: 192.0.2.21
      ESI: 0300.0034.5634.5600.0001

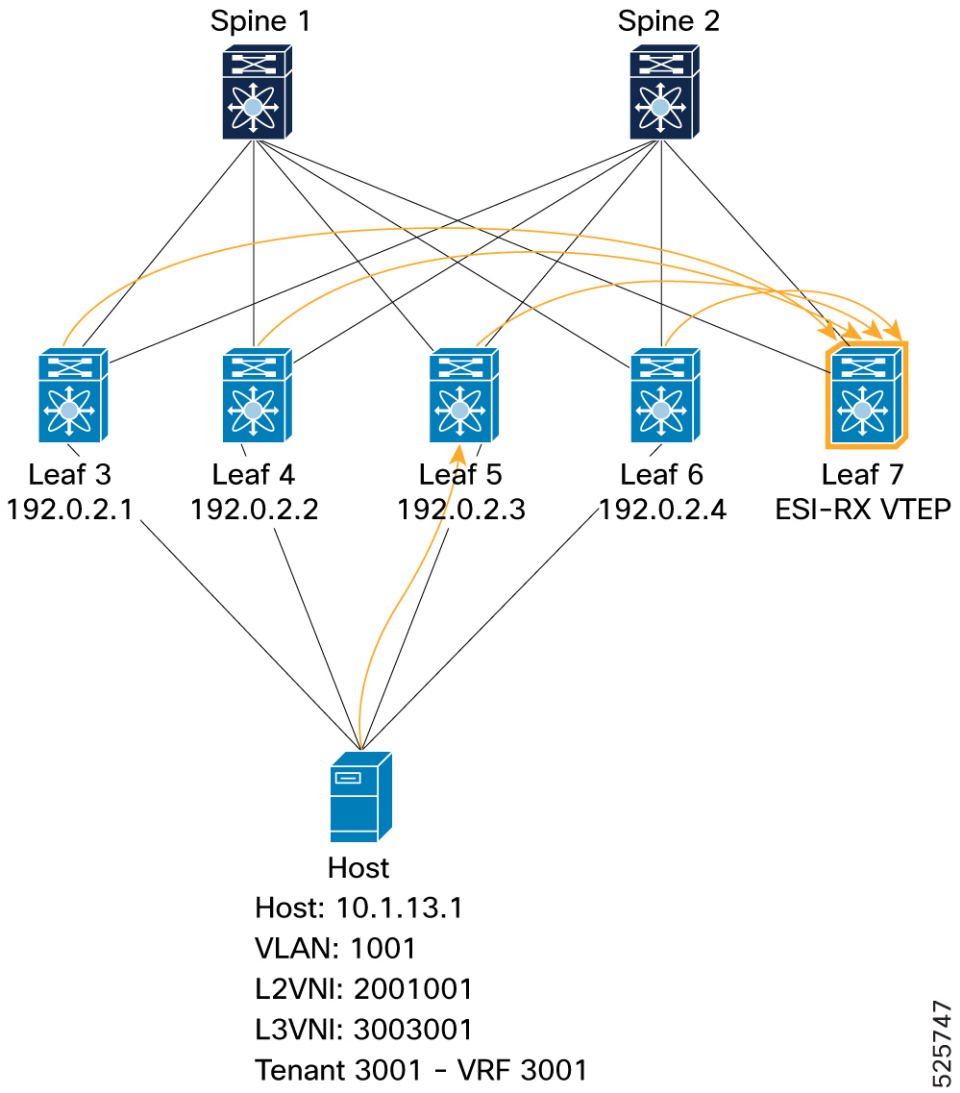
  Path type: internal, path is valid, not best reason: Router Id,
```

```
mh_peer_reoriginated, no labeled nexthop
            Imported from
192.0.16.1:33768:[2]:[0]:[0]:[48]:[0010.0100.1301]:[32]:[10.1.13.1]/272
  AS-Path: NONE, path sourced internal to AS
    192.0.2.4 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 SOO:192.0.16.1:256 ENCAP:8
          Router MAC:f839.184d.48c7
      Originator: 192.0.16.1 Cluster list: 192.0.2.21
      ESI: 0300.0034.5634.5600.0001

  Path type: internal, path is valid, not best reason: Local ESI,
mh_peer_reoriginated, no labeled nexthop
            Imported from
192.0.13.1:33768:[2]:[0]:[0]:[48]:[0010.0100.1301]:[32]:[10.1.13.1]/272
  AS-Path: NONE, path sourced internal to AS
    192.0.2.1 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 SOO:192.0.13.1:256 ENCAP:8
          Router MAC:4880.0290.0727
      Originator: 192.0.13.1 Cluster list: 192.0.2.21
      ESI: 0300.0034.5634.5600.0001
```

- **To Remote VTEP (ESI-RX, Leaf 7)**: Confirms that a remote VTEP (Leaf 7) receives all Type-2 routes for the host from the 4-way ESI MH cluster. The command `show bgp l2vpn evpn 10.1.13.1` on Leaf 7 shows four available paths for the host, each pointing to a different VTEP in the ESI cluster (192.0.2.1, 192.0.2.2, 192.0.2.3, 192.0.2.4).

Figure 110: Verification of PIP mode - Remote VTEP (ESI-RX)



BGP:

```
Leaf7# show bgp l2vpn evpn 10.1.13.1

  Path type: internal, path is valid, not best reason: Router Id, multipath, no
labeled nexthop, in rib
             Imported from
192.0.2.10:33768:[2]:[0]:[0]:[48]:[0010.0100.1301]:[32]:[10.1.13.1]/272
  AS-Path: NONE, path sourced internal to AS
    192.0.2.3 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 SOO:192.0.2.10:256 ENCAP:8
          Router MAC:4880.0290.01af
      Originator: 192.0.2.10 Cluster list: 192.0.2.21
      ESI: 0300.0034.5634.5600.0001

  Path type: internal, path is valid, not best reason: Router Id, multipath, no
labeled nexthop, in rib
```

```
            Imported from
192.0.14.1:33768:[2]:[0]:[0]:[48]:[0010.0100.1301]:[32]:[10.1.13.1]/272
  AS-Path: NONE, path sourced internal to AS
    192.0.2.2 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 SOO:192.0.14.1:256 ENCAP:8
          Router MAC:f839.1867.df6b
      Originator: 192.0.14.1 Cluster list: 192.0.2.21
      ESI: 0300.0034.5634.5600.0001

  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop, in rib
            Imported from
192.0.13.1:33768:[2]:[0]:[0]:[48]:[0010.0100.1301]:[32]:[10.1.13.1]/272
  AS-Path: NONE, path sourced internal to AS
    192.0.2.1 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 SOO:192.0.13.1:256 ENCAP:8
          Router MAC:4880.0290.0727
      Originator: 192.0.13.1 Cluster list: 192.0.2.21
      ESI: 0300.0034.5634.5600.0001

  Path type: internal, path is valid, not best reason: Router Id, multipath, no
labeled nexthop, in rib
            Imported from
192.0.16.1:33768:[2]:[0]:[0]:[48]:[0010.0100.1301]:[32]:[10.1.13.1]/272
  AS-Path: NONE, path sourced internal to AS
    192.0.2.4 (metric 9) from 192.0.2.21 (192.0.2.21)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 2001001 3003001
      Extcommunity: RT:1:2001001 RT:1:3003001 ENCAP:8 Router MAC:f839.184d.48c7
      Originator: 192.0.16.1 Cluster list: 192.0.2.21
      ESI: 0300.0034.5634.5600.0001
```

L2RIB:

```
Leaf7# show l2route evpn mac-ip evi 1001 detail

Topology    Mac Address    Host IP                                 Prod   Flags
       Seq No     Next-Hops
----------- -------------- ------------------------------------ ------
----------------- ---------- --------------------------
1001        0010.0100.1301 10.1.13.1                           BGP    --
       0         192.0.2.1 (Label: 2001001)

                 192.0.2.2 (Label: 2001001)

                 192.0.2.3 (Label: 2001001)

                 192.0.2.4 (Label: 2001001)
            Sent To: ARP
            ESI : 0300.0034.5634.5600.0001
            Encap-type:1
```

L2FM

```
Leaf7# show mac address-table vlan 1001
VLAN     MAC Address      Type      age       Secure   NTFY   Ports
---------+-----------------+--------+---------+------+----+------------------
C 1001     0010.0100.1301   dynamic  NA         F      F    nve1(192.0.2.1 192.0.2.2
 192.0.2.3 192.0.2.4)
```

URIB:

```
Leaf7# show ip route 10.1.13.1 detail vrf 3001

10.1.13.1/32, ubest/mbest: 4/0
    Extended Community: 0x1b 1c 01 03 65 01 0f 01 01 00 00 00 00 00 00 00 00 03
00 00 34 56 34 56 00 00 01
    *via 192.0.2.1%default, [200/0], 1d00h, bgp-1, internal, tag 1, segid: 3003001
tunnelid: 0x66010d01 encap: VXLAN

          BGP-EVPN: VNI=3003001 (EVPN)
         client-specific data: 125
         recursive next hop: 192.0.2.1/32%default
         extended route information: BGP origin AS 1 BGP peer AS 1
    *via 192.0.2.2%default, [200/0], 1d00h, bgp-1, internal, tag 1, segid: 3003001
tunnelid: 0x66010e01 encap: VXLAN

          BGP-EVPN: VNI=3003001 (EVPN)
         client-specific data: 138
         recursive next hop: 192.0.2.2/32%default
         extended route information: BGP origin AS 1 BGP peer AS 1
    *via 192.0.2.3%default, [200/0], 1d00h, bgp-1, internal, tag 1, segid: 3003001
tunnelid: 0x66010f01 encap: VXLAN

          BGP-EVPN: VNI=3003001 (EVPN)
         client-specific data: 1e9
         recursive next hop: 192.0.2.3/32%default
         extended route information: BGP origin AS 1 BGP peer AS 1
    *via 192.0.2.4%default, [200/0], 1d00h, bgp-1, internal, tag 1, segid: 3003001
tunnelid: 0x66011001 encap: VXLAN

          BGP-EVPN: VNI=3003001 (EVPN)
         client-specific data: a2
         recursive next hop: 192.0.2.4/32%default
         extended route information: BGP origin AS 1 BGP peer AS 1
```

FIB:

```
Leaf7# show forwarding ipv4 route 10.1.13.1 vrf 3001

──────────────┼──────────────────────────────┼──────────────┼─────────────┼─────────────
Prefix        | Next-hop                     | Interface
| Labels      | Partial Install
──────────────┼──────────────────────────────┼──────────────┼─────────────┼─────────────
10.1.13.1/32
                  192.0.2.1         nve1
                  192.0.2.2         nve1
                  192.0.2.3         nve1
                  192.0.2.4         nve1
```

IGMP or MLD Snooping with ESI - EVPN Type-7 and Type-8 route

This section describes how to configure the IGMP or MLD Snooping with ESI feature on Cisco NX-OS devices.

IGMP or MLD Snooping with ESI

IGMP or MLD snooping with ESI supports the synchronization of IGMP and MLD membership reports and leave messages between ESI peers, as specified in RFC 9251.

The supported synchronizations include:

- Multicast Membership Report Synch Route (Type-7)
- Multicast Leave Synch Route (Type-8)

EVPN Type-7 and Type-8 route synchronization support 2-way, 3-way, and 4-way ESI multi-homing with PIP or VIP modes.

EVPN route Type-7 (ReportSync route)

EVPN Type-7 routes synchronize membership report information learned on the ESI port-channel between ESI peers, as specified in RFC 9251.

- The local node that receives the membership report originates the Type-7 route.
- ESI peers install the ESI port as the Outgoing Interface (OIF) for the (S,G) learned remotely.
- The originating node is responsible for withdrawing the Type-7 route when the OIF expires on the local node.

EVPN route Type-8 (LeaveSync routes)

EVPN Type-8 LeaveSync routes are used to synchronize the leaves received on a local node to remote nodes as specified in RFC 9251.

- As part of leave processing, the node that receives the leave message sends a group-specific query on the ESI port to solicit a membership report from any other hosts on that port.
- Leave information, including a leave max-response-time, is advertised as an EVPN Type-8 route.
- Local and remote nodes start a leave timer to give hosts time to send a membership report in response to the group-specific query that originated on the ESI port. If a membership report is received from a host before the leave timer expires, the ESI port and report-sync state are maintained for the snooped IGMP route. If not, the ESI port is removed as an OIF for the snooped IGMP route after the leave timer expires.

Note

- The leave response timer on the originating node is adjusted to account for expected delays in sending the BGP message to ESI peers. This ensures that all ESI peers wait for approximately the same amount of time for a membership report response to the group-specific query originated on receipt of the leave while the leave timer is running.
- The original membership report, leave message, and response to the group-specific query can each be received by different ESI peers, as supported by RFC 9251.

- The originating node that advertises the Type-8 route is responsible for withdrawing it.

EVPN Type-7 (Wildcard routes)

- EVPN Type-7 supports advertisement of a (*,*) wildcard route between ESI peers.
- The wildcard route is needed for two cases
  - An external querier is present on the ESI port, and the external querier wins the querier election based on querier IP address that is performed with the VTEP.
  - A PIM router that sends PIM hellos is present on the ESI port.
- A wildcard route is needed for these two cases because membership reports from hosts on the ESI port may only be received by the external querier or PIM router. Therefore, the VTEP might be unaware of some or all (*,G) or (S,G) joins received from hosts on the ESI port. The resolution is to install the (*,*) wildcard route with the ESI port as an OIF for the route. This ensures that all multicast traffic is forwarded to the ESI port.
- Type-7 routes are used to synchronize the wildcard (,) route across ESI peers because only one VTEP in the ESI complex is aware that the ESI port is associated with the (,) route. This occurs because only one VTEP in the ESI complex participates in the querier election with the external querier. Similarly, only one VTEP in the ESI complex receives the PIM hellos from the PIM router. The Type-7 EVPN (*,*) report-sync message ensures that all ESI peers are aware that the ESI port is associated with a (*,*) route, and that remote ESI peers also install the wildcard (*,*) route with the ESI port as an OIF for that route.
- The EVPN Type (*,*) route is withdrawn when the VTEP stops receiving PIM hellos from the PIM router, when the external querier is removed, or when the VTEP wins the querier election due to its IP address over the ESI port.



Note While PIM router on the ESI port is supported, configuring an SVI with PIM sparse-mode enabled on the VTEP is not supported. Configuring an SVI with PIM sparse-mode causes the VTEP to function in L3 TRM mode, which is not currently supported.

Supported L2 multicast topologies and IGMP control flows

The following information describes the supported L2 multicast topologies and IGMP control flows for EVPN route Type-7 and route Type-8.

The EVPN route Type-7 and EVPN route Type-8 support 2-way, 3-way, and 4-way ESI with external querier and external PIM router on ESI port. However, you cannot configure PIM sparse mode on the VTEP ESI port SVI.

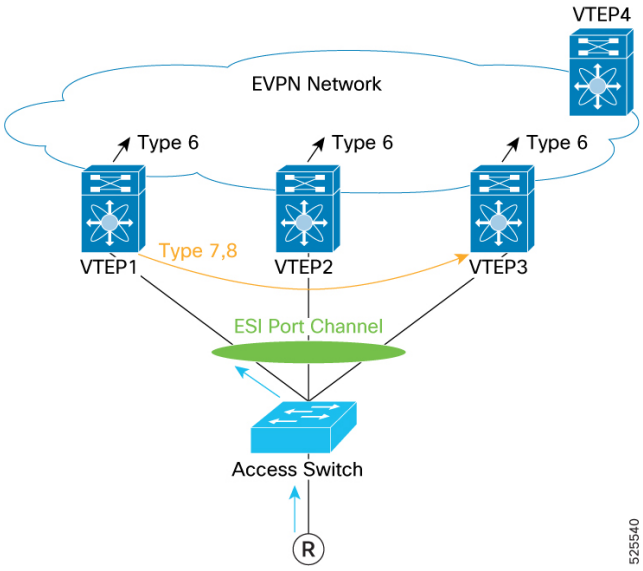
The supported L2 multicast topologies are as follows:

ESI Topology with no external querier or router

This topology contains these key components:

- **Receiver R**: Connected to the access switch or ESI port channel..
- **VTEPs**: VTEP1, VTEP2, VTEP3 that are part of the ESI, and remote VTEP4.

Figure 111: ESI Topology with no external querier or router

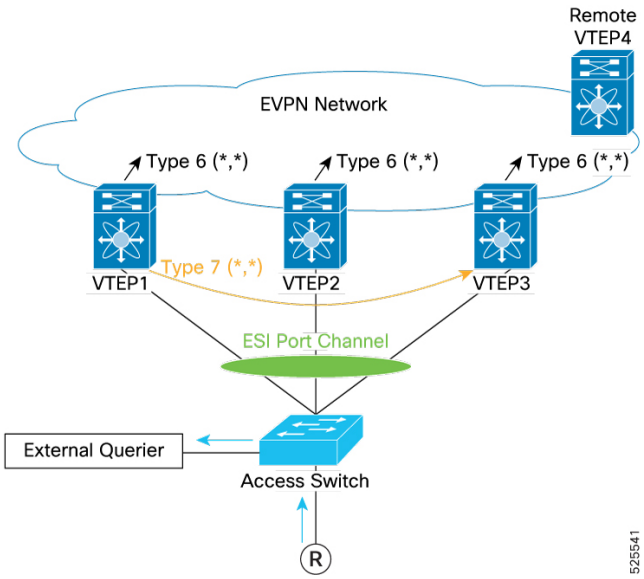


The process involves these steps:

- **Membership report handling**: When Receiver R sends a membership report, it is transmitted via the access switch or ESI port-channel and received by a VTEP, such as VTEP1. VTEP1 processes the report by sending a Type-7 Report Synchronization route to its ESI peers, VTEP2 and VTEP3. It also sends a Type-6 EVPN route to notify the remote VTEP4, as shown in the topology. After receiving the Type-7 route, VTEP2 and VTEP3 also send a Type-6 EVPN route. Currently, the Designated Forwarder (DF) is not the only VTEP that exclusively originates the Type-6 route, as specified in RFC 9251. As a result, each ESI peer originates redundant SMET routes containing the same information.
- **Leave message handling**: When Receiver R sends a leave message, it is received by a VTEP, such as VTEP1. VTEP1 initiates the leave message handling procedure and originates a group-specific query. It then sends a Type-8 route to its ESI peers, including VTEP2 and VTEP3. Subsequently, all involved VTEPs begin the leave synchronization procedure described in RFC 9251. As a result, the leave message is propagated across the network, preparing the system for the withdrawal of routes, if necessary.
- **Post leave synchronization phase**: If no membership report is received during the synchronization procedure, VTEP1, which sent the initial Type-7 report synchronization route, withdraws the Type-7 route from its ESI peers, VTEP2 and VTEP3. Similarly, VTEP1 withdraws the Type-8 leave synchronization route it sent to its ESI peers. As a result, the Type-6 SMET route is also withdrawn by VTEP1, VTEP2, and VTEP3 from the remote VTEP4, completing the route cleanup process and ensuring proper synchronization across the topology.

ESI Topology with external querier on ESI port

Figure 112: ESI Topology with external querier on ESI port

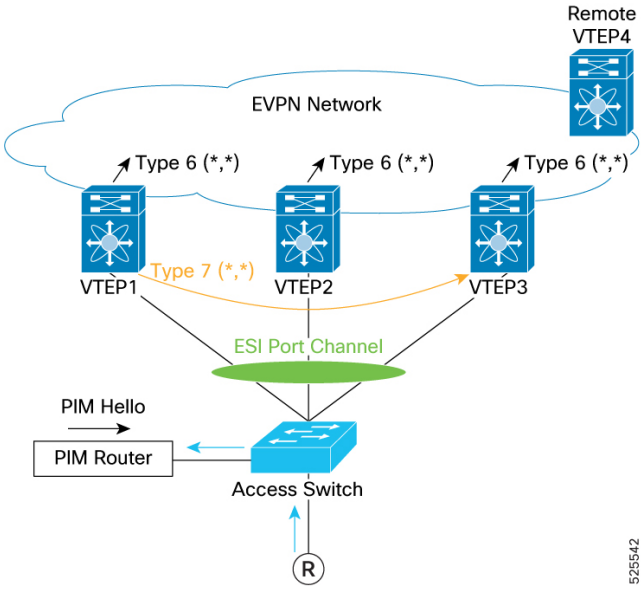


In this topology, an external querier has won the querier election and is the only node sending an IGMP general query. Since membership reports from the receiver R are sent only towards the port on which there is a querier, membership reports are not seen by the VTEPs. However, one of the VTEPs in the ESI complex, for example, VTEP1 that has participated in the querier election is aware that there is an external querier that has won the querier election; this VTEP marks the ESI port as an “mrouter port”. That VTEP (VTEP1) then originates a Type-7 (*,*) report-sync route to other ESI peers, which also correspondingly add the ESI port as an mrouter port. Separately, a Type-6 EVPN (*,*) route is sent by ESI peers to the remote VTEP4, as shown.

When the external querier is removed, or if the VTEPs that are part of the ESI win the querier election, the VTEP that originated the Type-7 (*,*) route withdraws the route from its ESI peers. Note that in this situation, since there is no ‘leave’ associated with the origination of the Type-7 (*,*) route , there will be no Type-8 (*,*) route sent by any of the ESI peers.

ESI topology with PIM router on ESI port

Figure 113: ESI topology with PIM router on ESI port



In this topology, the PIM router is detected when one of the VTEPs in the ESI complex sees a PIM hello, for example, VTEP1, and marks the ESI port as an “mrouter port.” This occurs regardless of whether the VTEPs in the ESI complex or the PIM router has won the querier election. The VTEP that sees the PIM hello will originate a Type-7 (*,*) route to other ESI peers, which will then also consider the ESI port as an mrouter port. Separately, a Type-6 EVPN (*,*) route is sent by ESI peers to the remote VTEP4, as shown.

When the PIM router is removed, the VTEP that originated the Type-7 (*,*) route withdraws the route from its ESI peers. Note that in this situation, since there is no ‘leave’ associated with the Type-7 (*,*) route, there is no Type-8 (*,*) route sent by any of the ESI peers.

Restriction The VTEP must not be configured with PIM sparse-mode for this topology to operate.

Guidelines and limitations for IGMP and MLD snooping with ESI

This section summarizes the supported behaviors, requirements, and restrictions for configuring IGMP and MLD snooping with ESI:

Configuration requirements

Enable both the **advertise evpn multicast** and **evpn multihoming** commands to originate Type 7 (ReportSync) and Type 8 (Leavesync) routes.

Operational behavior

- Multiple query packets, one from each ESI switch peer, reach the host. However, responses typically reach only one ESI node. This behavior does not impact traffic, but future optimization may reduce duplicate queries across ESI nodes.

- The Querier version and Querier IP address must be consistent and configured identically across all ESI nodes and peers.
- During leave processing, the node that receives a leave message from a host sends a group-specific query and advertises a Type 8 Leavesync route to ESI peers. Nodes that receive a leavesync perform leave procedures without sending further leavesync messages.
- During Joinsync withdrawal, remote nodes that receive the withdrawal send a query on the ESI port to check for memberships on that port.

IPv4 and IPv6 multicast support

- IPv4 multicast and IR underlay are supported.
- IPv6 multicast and IR underlay are not supported.
- PIM SVI configuration on an ESI port is not supported for Layer 2 multicast functionality.

IGMP and MLD snooping ESI synchronization command outputs

These commands display synchronization status and events for IGMP and MLD snooping in an ESI multi-homing EVPN environment. The command outputs include these fields:

- **VLAN**: VLAN ID associated with the multicast event.
- **VNI**: VXLAN Network Identifier for the event.
- **ESI**: Ethernet Segment Identifier, indicates multi-homing or single-homing segment.
- **Group**: Multicast group address (IPv4 or IPv6).
- **Source**: Multicast source address.
- **Local/Remote**: Specifies whether the report/event was learned on the local device or from a remote ESI peer.
- **Flags**: Include mode (I), Exclude mode (E), IGMP/MLD version (v2, v3, etc.).
- **MRT,sec**: Maximum Response Time in seconds (for leave events).

IGMP Snooping EVPN report synchronization

The **show ip igmp snooping evpn report-sync** command displays detailed information about the synchronization status of IGMP multicast reports in an ESI multi-homing environment.

The output provides critical insights such as the associated VLAN, multicast group, source IP, and whether the report was learned locally or from a remote ESI peer.

```
switch# show ip igmp snooping evpn report-sync
Flags: I - Include Source, E - Exclude Source
       v2 - IGMPv2, v3 - IGMPv3

VLAN       VNI        ESI                       Group                     Source
   Local/Remote   Flags
500        10500      0300.0a00.0b00.0e00.00ce 232.0.20.1              192.0.2.15
     Remote       I v3
509        10509      0300.0a00.0b00.0e00.00cf 232.0.29.1              192.0.2.44
```

```
     Local        I v3
version (v2 or v3) being used.
```

Summary of IGMP Report Synchronization for VLAN 500 and VLAN 509

For VLAN 500, the IGMP report for multicast group 232.0.20.1 (source 192.0.2.15) was synchronized from a remote EVPN peer. This report is associated with the ESI 0300.0a00.0b00.0e00.00ce and uses IGMPv3 in Include mode.

For VLAN 509, the IGMP report for multicast group 232.0.29.1 (source 192.0.2.44) was synchronized locally on the device. This report is associated with the ESI 0300.0a00.0b00.0e00.00cf and also uses IGMPv3 in Include mode.

MLD Snooping EVPN report synchronization

The **show ipv6 mld snooping evpn report-sync** command displays details about the synchronization status of MLD (IPv6 multicast) reports in an ESI multi-homing environment.

This output helps monitor multicast group membership for IPv6 traffic and indicates whether reports were learned locally or from remote EVPN peers.

```
switch# show ipv6 mld snooping evpn report-sync
Flags: I - Include Source, E - Exclude Source
       v1 - MLDv1, v2 - MLDv2

VLAN       VNI        ESI                       Group                     Source
       Local/Remote Flags
509        10509      0300.0a00.0b00.0e00.00cf 2001:DB8::32                 2001:DB8::9
           Local        I v2
```

The MLD report for VLAN 509 tracks the IPv6 multicast group 2001:DB8::32 with source 2001:DB8::9. It is associated with the ESI 0300.0a00.0b00.0e00.00cf, was learned locally on the device, and uses MLDv2 in Include mode (I v2).

IGMP Snooping EVPN leave synchronization

This command displays details about the synchronization of IGMP leave messages for IPv4 multicast groups.

```
switch# show ip igmp snooping evpn leave-sync
Flags: I - Include Source, E - Exclude Source
       v2 - IGMPv2, v3 - IGMPv3

VLAN       VNI        ESI                       Group                     Source
       Local/Remote Flags        MRT,sec
500        10500      0300.0a00.0b00.0e00.00ce 232.0.20.1                192.0.2.35
     Remote                    0
509        10509      0300.0a00.0b00.0e00.00cf 232.0.29.1                192.0.2.44
     Local        I v3         2
```

VLAN 500: The leave event for multicast group 232.0.20.1 (source 192.0.2.35) was learned from a remote EVPN peer, with an MRT of 0 seconds, indicating the group is about to be cleared.

VLAN 509: The leave event for multicast group 232.0.29.1 (source 192.0.2.44) was learned locally, using IGMPv3 in Include mode (I v3), with an MRT of 2 seconds.

MLD Snooping EVPN leave synchronization

This command provides details about the synchronization of MLD (Multicast Listener Discovery) leave messages for IPv6 multicast groups. Below is an example of the command output:

```
switch# show ipv6 mld snooping evpn leave-sync
Flags: I - Include Source, E - Exclude Source
       v1 - MLDv1, v2 - MLDv2

VLAN       VNI        ESI                       Group                     Source
       Local/Remote Flags        MRT,sec
509        10509      0300.0a00.0b00.0e00.00cf 2001:DB8::32            2001:DB8::9
      Local        I v2         2
```

VLAN 509: The leave event for IPv6 multicast group 2001:DB8::32 (source 2001:DB8::9) was learned locally, using MLDv2 in Include mode (I v2), with an MRT of 2 seconds.

ECMP reuse

This section describes how to configure the ECMP reuse feature on Cisco NX-OS devices.

ECMP reuse

ECMP reuse is a scale optimization feature that:

- provides ECMP and adjacency sharing capability among different L3VNIs or VRFs.
- allows sharing ECMP resources for IPv4 and IPv6 prefixes within same VRF, and
- reduces the adjacency and ECMP resource usage to handle workloads more efficiently.

Use this feature only if you want to optimize overlay ECMP resources when overlay ECMP resources are exhausted in ESI PIP mode. For more information on enabling this feature, see Enable ECMP reuse, on page 860 section.

Guidelines and limitations for ECMP reuse

This section summarizes the supported behaviors, requirements, and restrictions for configuring ECMP reuse:

- Only new style L3VNIs (VNIs configured with “l3” keyword under “vrf context”) are supported. The VN-segment associated with a VLAN is not supported.
- A system reload is required to either enable or disable the ECMP reuse feature.

Enable ECMP reuse

When overlay ECMP resources are exhausted in ESI PIP mode, enable ECMP Reuse to optimize overlay ECMP resources.

Follow these steps to enable or disable the overlay ECMP reuse.

Before you begin

Ensure that Layer 3 VNIs are configured with new L3VNI. For more information on configuration, see Configuring Layer 3 Tenant Routed Multicast, on page 410.

Procedure

Step 1 Enter global configuration mode.

Use the **configure terminal** command.

Example:

```
switch# configure terminal
switch(config)#
```

Step 2 (Optional) Enable ECMP reuse.

Use the **system nve ecmp-reuse** command.

Example:

```
switch(config)# system nve ecmp-reuse
```

To disable ECMP reuse if needed, use the **no system nve ecmp-reuse** command.

Step 3 Copy the running configuration to the startup configuration.

Use the **copy running-config startup-config** command.

Example:

```
switch(config)# copy running-config startup-config
```

Step 4 Reload the switch.

Use the **reload** command.

Example:

```
switch(config)# reload
```

ECMP reuse is enabled and active after the switch reloads.

Layer 2 Gateway Spanning Tree Protocol

This section describes how to configure the Layer 2 Gateway Spanning Tree Protocol (L2G-STP) feature on Cisco NX-OS devices.

Layer 2 Gateway Spanning Tree Protocol

A Layer 2 Gateway Spanning Tree Protocol (L2G-STP) is a network protocol that

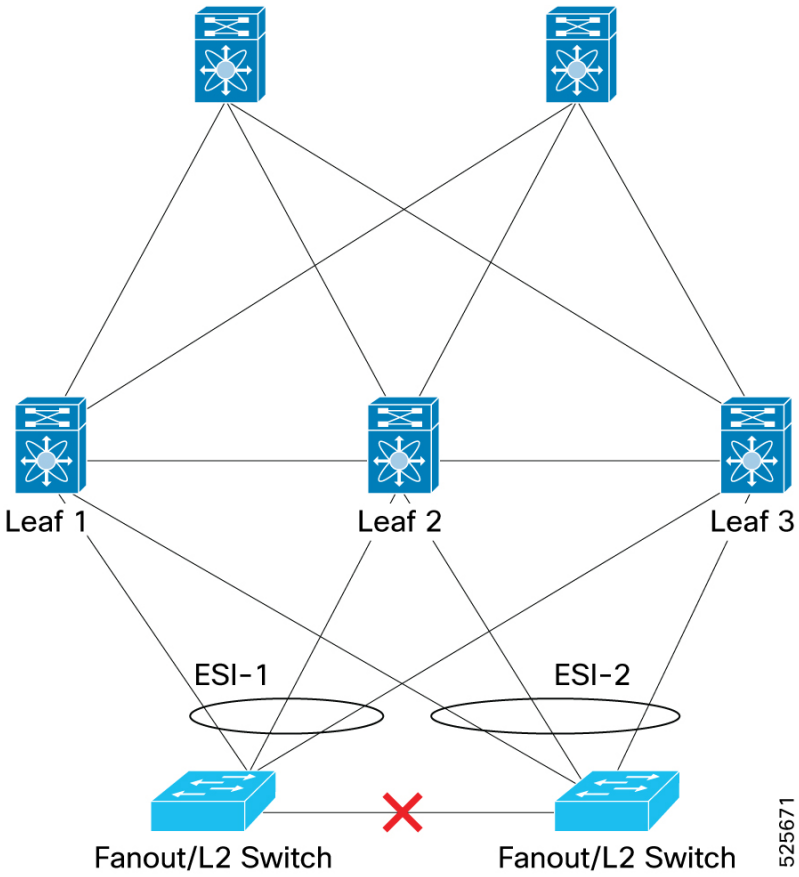
- builds a loop-free topology to prevent broadcast storms in VXLAN EVPN multi-homing environments,
- enables all VTEPs to present as a single logical root bridge in the network, and
- automatically applies root guard and manages bridge priority to maintain central control of spanning tree operations.

L2G-STP is disabled by default. You must enable it using the `spanning-tree domain enable` command for VTEPs to function as a pseudo-root for downstream switches. When enabled, the VXLAN fabric emulates a single bridge, maintaining stability, redundancy, and loop prevention throughout the fabric. Bridge IDs are composed of a consistent MAC address and bridge priority, which allows all participating switches to appear as one logical root. Root guard ensures the overlay remains the persistent root even in the presence of superior spanning tree information from access switches.

Example

Suppose multiple VTEPs participate in a VXLAN EVPN multi-homing scenario. When L2G-STP is enabled, all these VTEPs act as a single root to downstream switches. This approach centralizes control, prevents root conflicts or loops, and allows redundancy and rapid failover.

Figure 114: Topology of L2G-STP



Best practice for moving to Layer 2 Gateway STP

Follow these best practices when moving to Layer 2 Gateway STP:

- Enable root guard by default on all access ports when using Layer 2 Gateway STP.
- Ensure all VTEPs in the VXLAN fabric act as a single pseudo-root switch for southbound Layer 2 switches.

- Configure all access ports from VTEPs that connect to southbound Layer 2 switches so that those ports remain in the **Desg FWD** state by default.
- Make sure all ports on southbound Layer 2 switches connecting to VTEPs are either in the **root-port FWD** or **Altn BLK** state.
- Activate root guard if superior spanning-tree information is received from southbound Layer 2 switches. This action puts ports in the **BLK L2GW_Inc** state to secure the root on the VXLAN fabric and prevent loops.
- Configure an explicit domain ID to enable spanning-tree BPDU tunneling across the fabric.
- Set all VTEPs to the lowest spanning-tree priority among all switches in the spanning-tree domain, making all VTEPs the root bridge so the entire VXLAN fabric appears as one virtual bridge.
- Do not enable ESI interfaces in spanning-tree edge mode if Layer 2 Gateway STP needs to run across the VTEP and access layer.
- You can continue to use ESIs or orphans (single-homed hosts) in spanning-tree edge mode for direct connections to hosts/servers that do not run Spanning Tree Protocol.
- Configure all VTEPs connected by a common customer access layer in the same Layer 2 Gateway STP domain.
- Remember: The Layer 2 Gateway STP domain scope is global, and all ESIs on a given VTEP can participate in only one domain.
- Ensure that mappings between Multiple Spanning Tree (MST) instances and VLANs are consistent across VTEPs within a Layer 2 Gateway STP domain.
- Do not directly connect non-Layer 2 Gateway STP-enabled VTEPs to Layer 2 Gateway STP-enabled VTEPs. This connection may cause conflicts due to unwanted BPDU propagation and may also steer the root outside.
- Ensure that the root of a configured STP domain local to the VXLAN fabric is always a VTEP within the fabric.
- After upgrading to the latest build, retain spanning-tree edge mode and BPDU filter configurations on Cisco Nexus switches and southbound Layer 2 switches.
- Enable Layer 2 Gateway STP on all switches with recommended priority and MST instance mapping as needed. Use the commands `spanning-tree domain enable` and spanning-tree mst instance-id priority 8192.
- First, remove BPDU filter configurations on the switch side. Then, remove BPDU filter configurations and spanning-tree edge mode on the southbound Layer 2 switch.

Enable Layer 2 Gateway STP on a switch

Enable Spanning Tree Protocol (STP) in Layer 2 gateway mode to prevent loops and ensure proper network path selection on a switch.

Follow these steps to enable Layer 2 Gateway STP on a switch.

Procedure

Step 1 Enter global configuration mode.

Use the **configure terminal** command.

Example:

```
switch# configure terminal
switch(config)#
```

Step 2 Enable the Spanning Tree Protocol mode.

Use the **spanning-tree mode [rapid-pvst | mst]** command.

Example:

```
switch(config)# spanning-tree mode mst
```

To disable, use **no spanning-tree mode [rapid-pvst | mst]** command.

Step 3 Enable Layer 2 Gateway STP.

Use the **spanning-tree domain** [**enable** | *domain identifier*] command.

Table 19: Spanning tree domain options

Option	Description	Example
enable	Enable Layer 2 Gateway STP.	`switch(config)# spanning-tree domain enable`
domain identifier	Enable Layer 2 Gateway STP and set explicit domain ID to tunnel encoded BPDUs.	`switch(config)# spanning-tree domain 1`

Note
Before reconfiguring, use the **no spanning-tree domain enable** command to restore to default.

Example:

```
switch(config)# spanning-tree domain 1
switch(config)# no spanning-tree domain enable
switch(config)# spanning-tree domain enable
```

Step 4 Configure Spanning Tree Protocol priority.

Use the **spanning-tree** [**mst** *id* | **vlan** *id*] *priority value* command.

Example:

```
switch(config)# spanning-tree mst 1 priority 8192
switch(config)# spanning-tree vlan 1001 1 priority 8192
```

To remove the priority, use the **no spanning-tree** [**mst** *id* | **vlan** *id*] *priority value* command.

Step 5 Verify the STP configuration.

Use the **show spanning-tree summary**

Example:

```
switch(config)# show spanning-tree summary
```

For sample verification examples, see Troubleshooting Layer 2 Gateway STP configurations, on page 865.

What to do next

Verify STP status. For more information, see Troubleshooting Layer 2 Gateway STP configurations, on page 865.

Troubleshooting Layer 2 Gateway STP configurations

These examples show how Layer 2 Gateway STP configurations affect bridge priority, root guard handling, and port roles.

Spanning-tree priority settings

Each Layer 2 Gateway STP VLAN uses a default spanning-tree priority of 8192, which is lower than the priority used by most customer edge devices. This setting makes the VTEP the spanning-tree root. Priorities increase in steps of 4096.

- For example, a calculated priority may be: 8192 (base) + 0 (instance ID) = 8192.
- You can use the show spanning-tree command to confirm that the VTEP is the root.

```
switch# show spanning-tree summary
Switch is in mst mode (IEEE Standard)
Root bridge for: MST0000
L2 Gateway STP bridge for: MST0000
L2 Gateway Domain ID: 1
Port Type Default                        is disable
Edge Port [PortFast] BPDU Guard Default  is disabled
Edge Port [PortFast] BPDU Filter Default is disabled
Bridge Assurance                         is enabled
Loopguard Default                        is disabled
Pathcost method used                     is long
PVST Simulation                          is enabled
STP-Lite                                 is disabled

Name                   Blocking Listening Learning Forwarding STP Active
---------------------- -------- --------- -------- ---------- ----------
MST0000                       0         0        0         12         12
---------------------- -------- --------- -------- ---------- ----------
1 mst                         0         0        0         12         12

switch# show spanning-tree vlan 1001

MST0000
  Spanning tree enabled protocol mstp

  Root ID    Priority    8192
             Address     c84c.75fa.6001   !L2G-STP reserved mac+ domain id
             This bridge is the root
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec

  Bridge ID  Priority    8192   (priority 8192 sys-id-ext 0)
```

```
                Address     c84c.75fa.6001
                Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec
```

Root guard and inconsistent state

- VTEP access ports use root guard.
- If the port receives a superior BPDU, it enters the 'L2 Gateway inconsistent' state until the condition clears. Log messages show these events, and interface states marked '*L2GW_Inc' in show spanning-tree output indicate inconsistency due to BPDUs.

```
2016 Aug 29 19:14:19 TOR9-leaf4 %$ VDC-1 %$ %STP-2-L2GW_BACKBONE_BLOCK: L2 Gateway
Backbone port inconsistency blocking port Ethernet1/1 on MST0000.
2016 Aug 29 19:14:19 TOR9-leaf4 %$ VDC-1 %$ %STP-2-L2GW_BACKBONE_BLOCK: L2 Gateway
Backbone port inconsistency blocking port port-channel13 on MST0000.
switch# show spanning-tree

MST0000
  Spanning tree enabled protocol mstp
  Root ID    Priority    8192
             Address     c84c.75fa.6001
             This bridge is the root
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec

  Bridge ID  Priority    8192   (priority 8192 sys-id-ext 0)
             Address     c84c.75fa.6001
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec

Interface        Role Sts Cost      Prio.Nbr   Type
---------------- ---- --- --------- -------- --------------------------------
Po1              Desg FWD 20000     128.4096  Edge P2p
Po2              Desg FWD 20000     128.4097  Edge P2p
Po3              Desg FWD 20000     128.4098  Edge P2p
Po12             Desg BKN*2000      128.4107  P2p *L2GW_Inc
Po13             Desg BKN*1000      128.4108  P2p *L2GW_Inc
Eth1/1           Desg BKN*2000      128.1     P2p *L2GW_Inc
```

Disabling Layer 2 Gateway STP

- To disable STP, enter the command: **spanning-tree domain disable**
- After you disable STP, the bridge MAC address resets to the system MAC address. The VTEP may not remain the root, and access switches can become the root.

```
switch(config)# spanning-tree domain disable

switch# show spanning-tree summary
Switch is in mst mode (IEEE Standard)
Root bridge for: none
L2 Gateway STP                          is disabled
Port Type Default                       is disable
Edge Port [PortFast] BPDU Guard Default  is disabled
Edge Port [PortFast] BPDU Filter Default is disabled
Bridge Assurance                        is enabled
Loopguard Default                       is disabled
Pathcost method used                    is long
PVST Simulation                         is enabled
STP-Lite                                is disabled

Name                   Blocking Listening Learning Forwarding STP Active
---------------------- -------- --------- -------- ---------- ----------
```

```
MST0000                     4         0        0          8         12
---------------------- -------- --------- -------- ---------- ----------
1 mst                       4         0        0          8         12

switch# show spanning-tree vlan 1001
MST0000
  Spanning tree enabled protocol mstp
  Root ID    Priority    4096
             Address     00c8.8ba6.5073
             Cost        0
             Port        4108 (port-channel13)
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec

  Bridge ID  Priority    8192   (priority 8192 sys-id-ext 0)
             Address     5897.bd1d.db95
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec
```

When STP is disabled, access ports behave in the following ways:

- With STP disabled, VTEP access ports function as regular spanning-tree ports and receive BPDUs from access switches.
- When you disable STP, ports lose rapid forwarding. Each port must complete a proposal or agreement handshake before it enters the forwarding state.