Mellanox Spectrum™

Linux Switch Configuration Guide

Rev 1.0

NOTE:
THIS HARDWARE, SOFTWARE OR TEST SUITE PRODUCT ("PRODUCT(S)") AND ITS RELATED DOCUMENTATION ARE PROVIDED BY MELLANOX TECHNOLOGIES "AS-IS" WITH ALL FAULTS OF ANY KIND AND SOLELY FOR THE PURPOSE OF AIDING THE CUSTOMER IN TESTING APPLICATIONS THAT USE THE PRODUCTS IN DESIGNATED SOLUTIONS. THE CUSTOMER'S MANUFACTURING TEST ENVIRONMENT HAS NOT MET THE STANDARDS SET BY MELLANOX TECHNOLOGIES TO FULLY QUALIFY THE PRODUCT(S) AND/OR THE SYSTEM USING IT. THEREFORE, MELLANOX TECHNOLOGIES CANNOT AND DOES NOT GUARANTEE OR WARRANT THAT THE PRODUCTS WILL OPERATE WITH THE HIGHEST QUALITY. ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT ARE DISCLAIMED. IN NO EVENT SHALL MELLANOX BE LIABLE TO CUSTOMER OR ANY THIRD PARTIES FOR ANY DIRECT, INDIRECT, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES OF ANY KIND (INCLUDING, BUT NOT LIMITED TO, PAYMENT FOR PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY FROM THE USE OF THE PRODUCT(S) AND RELATED DOCUMENTATION EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Mellanox Technologies
350 Oakmead Parkway Suite 100
Sunnyvale, CA 94085
U.S.A.
www.mellanox.com
Tel: (408) 970-3400
Fax: (408) 970-3403

© Copyright 2019. Mellanox Technologies Ltd. All Rights Reserved.

Mellanox®, Mellanox logo, Mellanox Open Ethernet®, LinkX®, Mellanox Spectrum®, Mellanox Virtual Modular Switch®, MetroDX®, MetroX®, MLNX-OS®, ONE SWITCH. A WORLD OF OPTIONS®, Open Ethernet logo, Spectrum logo, Switch-IB®, SwitchX®, UFM®, and Virtual Protocol Interconnect® are registered trademarks of Mellanox Technologies, Ltd.

For the complete and most updated list of Mellanox trademarks, visit http://www.mellanox.com/page/trademarks.

All other trademarks are property of their respective owners.

Table of Contents

1 Linux Switch Overview 4

2 Suggested Preconditions For Fabric Configuration 5

2.1 Physical Port Identification using udev rules 5

3 Fabric Types 6

3.1 L3 Fabric (BGP unnumbered) 6

3.1.1 Fabric Troubleshooting 8

3.2 EVPN Based VxLAN L2 Extension 10

3.2.1 Fabric Troubleshooting 12

3.3 EVPN Based VxLAN with Asymmetric Routing 14

3.3.1 Fabric Troubleshooting 18

3.4 EVPN Based VxLAN with Symmetric Routing 20

3.4.1 Fabric Troubleshooting 24

3.5 EVPN Based VxLAN with Symmetric Routing (Type5) 26

3.5.1 Fabric Troubleshooting 30

3.6 Routing from the Host 32

3.6.1 Fabric Troubleshooting 34

3.7 RoCE over VxLAN (PFC/ECN) 36

3.7.1 Fabric Troubleshooting 43

1 Linux Switch Overview

Mellanox Spectrum™ Linux Switch enables users to natively install and use any standard Linux distribution as the switch operating system on the Open Ethernet Mellanox Spectrum™ switch platforms and ASIC.

The Spectrum Linux switch is enabled by a Linux kernel driver model for Ethernet switches. It breaks the dependency of using vendor-specific, closed-source software development kits (SDK).

The open-source Linux driver is developed and maintained in the Linux kernel, replacing proprietary APIs with standard Linux kernel interfaces to control the switch hardware. This allows off-the-shelf Linux-based networking applications to operate on the Spectrum switch, including L2 switching, L3 routing, and IP tables (ACLs) at hardware-accelerated speeds.

For more information, check out Linux Switch Blog.

This document describes the configuration of common data center network fabrics

- L3 fabric with BGP
- EVPN Based VxLAN L2 Extension
- Symmetric Routing
- Asymmetric Routing
- Type5 route insertion with Symmetric routing
- Routing from the Host
- RoCE over VxLAN (PFC/ECN)

The following scenarios were done on the Mellanox SN2XXX switch family using Fedora 28 with kernel version 5.0.5 and FRR.

2 Suggested Preconditions For Fabric Configuration

2.1 Physical Port Identification using udev rules

It is possible to create udev rules which rename the software interfaces (port netdevs) corresponding to the front panel ports according to the front panel numbering.

To do so, create the following rule in **/etc/udev/rules.d/10-local.rules**:

```
SUBSYSTEM=="net", ACTION=="add", DRIVERS=="mlxsw_spectrum*", \
    NAME="sw$attr{phys_port_name}"
```

The configuration commands used in the following examples like IP link/IP address/bridge should also work on other Linux distribution OS, however these commands are not persistent.

To make the configuration persistent, the commands should be included in a bash script and should be run automatically after startup.

For example, one way to do it, is to use **systemd service**:

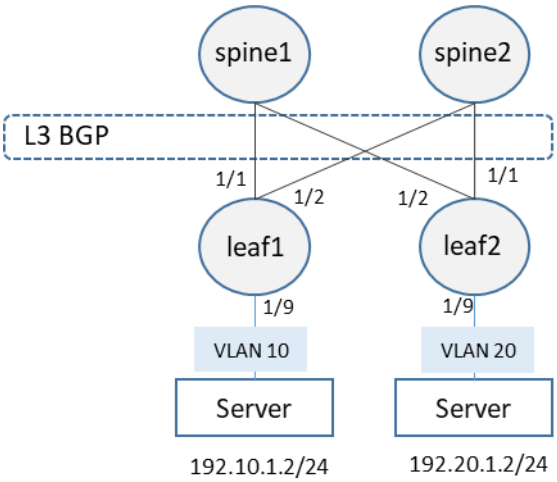
```
#cat scenario1.sh
#!/bin/sh -
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216


#cat /etc/systemd/system/startup.service
[Unit]
Description=startup configuration
After=multi-user.target

[Service]
ExecStart=/root/scenario1.sh

[Install]
WantedBy=default.target

#systemctl enable startup.service
#systemctl start startup.service
```

For FRR installation on Fedora 28, the following procedure can be used:

```
#dnf copr enable nucleo/frr-git
#dnf install frr
```

3 Fabric Types

3.1 L3 Fabric (BGP unnumbered)



Leaf 1 Configuration

1. Configure the physical links

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip link set dev swp9 up mtu 9216
```

2. Configure the loopback interface

```
ip addr add 1.1.1.1/32 dev lo
```

3. Configure a vlan-aware bridge and assign host facing port as member of the bridge

```
ip link add name br0 type bridge vlan_filtering 1 vlan_default_pvid 0 mcast_snooping
0
ip link set dev swp9 master br0
```

4. Add VLAN 10 as access VLAN to the host facing port

```
bridge vlan add vid 10 dev swp9 untagged pvid
```

5. Configure SVI for VLAN 10 (to act as a gateway for the host

```
ip link add link br0 name vlan10 up mtu 9216 type vlan id 10
bridge vlan add vid 10 dev br0 self
ip addr add 192.10.1.1/24 dev vlan10
ip link set dev br0 up
```

6. Under FRR, configure unnumbered BGP neighbors and advertise the local prefixes

```
router bgp 65011
 bgp router-id 1.1.1.1
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal Fabric Network
 neighbor fabric capability extended-nexthop
 neighbor swp1 interface peer-group fabric
 neighbor swp2 interface peer-group fabric
 !
 address-family ipv4 unicast
  network 1.1.1.1/32
  redistribute connected
 exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
```

Leaf 2 Configuration

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip link set dev swp9 up mtu 9216
ip addr add 2.2.2.2/32 dev lo

ip link add name br0 type bridge vlan_filtering 1 vlan_default_pvid 0 mcast_snooping
0
ip link set dev swp9 master br0
bridge vlan add vid 10 dev swp9 untagged pvid
ip link add link br0 name vlan10 up mtu 9216 type vlan id 10
bridge vlan add vid 10 dev br0 self
ip addr add 192.20.1.1/24 dev vlan10
ip link set dev br0 up
```

FRR configuration:

```
router bgp 65012
 bgp router-id 2.2.2.2
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal Fabric Network
 neighbor fabric capability extended-nexthop
 neighbor swp1 interface peer-group fabric
 neighbor swp2 interface peer-group fabric
 !
 address-family ipv4 unicast
  network 2.2.2.2/32
  redistribute connected
 exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
```

Spine 1 Configuration

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip addr add 3.3.3.3/32 dev lo
```

FRR configuration

```
router bgp 65020
 bgp router-id 3.3.3.3
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal
Fabric Network
 neighbor fabric capability extended-
nexthop
 neighbor swp1 interface peer-group
fabric
 neighbor swp2 interface peer-group
fabric
 !
 address-family ipv4 unicast
  network 3.3.3.3/32
 exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
 !
```

Spine 2 Configuration

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip addr add 4.4.4.4/32 dev lo
```

FRR configuration

```
router bgp 65020
 bgp router-id 4.4.4.4
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal
Fabric Network
 neighbor fabric capability extended-
nexthop
 neighbor swp1 interface peer-group
fabric
 neighbor swp2 interface peer-group
fabric
 !
 address-family ipv4 unicast
  network 4.4.4.4/32
 exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
 !
```

Host 1 Configuration

```
ip route add 192.20.1.0/24 via
192.10.1.1
```

Host 2 Configuration

```
ip route add 192.10.1.0/24 via
192.20.1.1
```

3.1.1 Fabric Troubleshooting

1. Verify all IP links are up and configured

```
#ip link
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN mode DEFAULT
group default qlen 1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
14: swp1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc fq_codel state UP mode
DEFAULT group default qlen 1000
    link/ether 50:6b:4b:88:0f:19 brd ff:ff:ff:ff:ff:ff
15: swp2: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc fq_codel state UP mode
DEFAULT group default qlen 1000
    link/ether 50:6b:4b:88:0f:1d brd ff:ff:ff:ff:ff:ff
17: swp9: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc fq_codel master br0 state
UP mode DEFAULT group default qlen 1000  link/ether 50:6b:4b:88:0f:25 brd
ff:ff:ff:ff:ff:ff
24: br0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue state UP mode
DEFAULT group default qlen 1000
    link/ether 50:6b:4b:88:0f:25 brd ff:ff:ff:ff:ff:ff
25: vlan10@br0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue state UP
mode DEFAULT group default qlen 1000 link/ether 50:6b:4b:88:0f:25 brd
ff:ff:ff:ff:ff:ff
```

2. Verify the bridge VLAN setting

```
#bridge vlan show
port    vlan ids
swp9
         1 untagged      10 PVID untagged
br0
         1 PVID untagged         10
```

3. Verify the MAC address table

```
#bridge fdb show vlan 10
00:02:c9:1c:69:30 dev swp9 vlan 10 extern_learn offload br0
50:6b:4b:88:0f:25 dev swp9 vlan 10 br0 permanent
```

4. Verify ARP table

```
#ip neigh | grep 192.10.1.2
192.10.1.2 dev vlan10 lladdr 00:02:c9:1c:69:30 REACHABLE
```

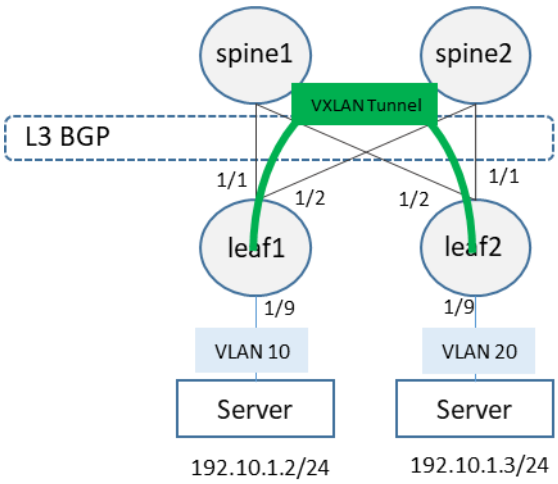
5. Verify BGP neighbors are ESTABLISHED

```
#vtysh -c 'show ip bgp summary'

IPv4 Unicast Summary:
BGP router identifier 1.1.1.1, local AS number 65011 vrf-id 0
BGP table version 9
RIB entries 13, using 2080 bytes of memory
Peers 2, using 41 KiB of memory
Peer groups 1, using 64 bytes of memory

Neighbor        V         AS MsgRcvd MsgSent   TblVer  InQ OutQ  Up/Down State/PfxRcd
swp1            4      65020     228     233        0    0    0 03:39:23            4
swp2            4      65020     229     233        0    0    0 03:39:24            4

Total number of neighbors 2
```

6. Verify the routing table

```
#ip route
2.2.2.2 proto bgp metric 20
        nexthop via 169.254.0.1 dev swp2 weight 1 onlink offload
        nexthop via 169.254.0.1 dev swp1 weight 1 onlink offload
3.3.3.3 via 169.254.0.1 dev swp1 proto bgp metric 20 onlink offload
4.4.4.4 via 169.254.0.1 dev swp2 proto bgp metric 20 onlink offload
192.10.1.0/24 dev vlan10 proto kernel scope link src 192.10.1.1 offload
192.20.1.0/24 proto bgp metric 20
        nexthop via 169.254.0.1 dev swp2 weight 1 onlink offload
        nexthop via 169.254.0.1 dev swp1 weight 1 onlink offload
```

3.2 EVPN Based VxLAN L2 Extension



Leaf 1 Configuration

1. Configure the physical links

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip link set dev swp9 up mtu 9216
```

2. Configure the loopback interface

```
ip addr add 1.1.1.1/32 dev lo
```

3. Create VxLAN interface

```
ip link add name vx10010 up mtu 9216 type vxlan id 10 local 1.1.1.1 dstport 4789
nolearning noudpcsum tos inherit ttl lo0
```

4. Configure a vlan-aware bridge and assign host facing port as member of the bridge

```
ip link add name br0 type bridge vlan_filtering 1 vlan_default_pvid 0 mcast_snooping
0
ip link set dev swp9 master br0
ip link set dev vx10010 master br0
```

4. Add VLAN 10 as access VLAN to the host facing port and VxLAN interface

```
bridge vlan add vid 10 dev swp9 untagged pvid
bridge vlan add vid 10 dev vx10010 untagged pvid
ip link set dev br0 up
```

5. Configure SVI for VLAN 10 (to act as a gateway for the host

```
ip link add link br0 name vlan10 up mtu 9216 type vlan id 10
bridge vlan add vid 10 dev br0 self
ip addr add 192.10.1.1/24 dev vlan10
ip link set dev br0 up
```

6. Under FRR, configure unnumbered BGP neighbors and advertise the loopback and all EVPN VNI

```
router bgp 65011
 bgp router-id 1.1.1.1
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal Fabric Network
 neighbor fabric capability extended-nexthop
 neighbor swp1 interface peer-group fabric
 neighbor swp2 interface peer-group fabric
 !
 address-family ipv4 unicast
  network 1.1.1.1/32
exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
 !
 address-family l2vpn evpn
  neighbor fabric activate
  advertise-all-vni
 exit-address-family
```

Leaf 2 Configuration

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip link set dev swp9 up mtu 9216
ip addr add 2.2.2.2/32 dev lo
ip link add name vx10010 up mtu 9216 type vxlan id 10 local 2.2.2.2 dstport 4789
nolearning noudpcsum tos inherit ttl lo0
ip link add name br0 type bridge vlan_filtering 1 vlan_default_pvid 0 mcast_snooping
0
ip link set dev swp9 master br0
ip link set dev vx10010 master br0
bridge vlan add vid 10 dev swp9 untagged pvid
bridge vlan add vid 10 dev vx10010 untagged pvid
ip link set dev br0 up
```

FRR configuration:

```
router bgp 65012
 bgp router-id 2.2.2.2
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal Fabric Network
 neighbor fabric capability extended-nexthop
 neighbor swp1 interface peer-group fabric
 neighbor swp2 interface peer-group fabric
 !
 address-family ipv4 unicast
  network 2.2.2.2/32
exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
 !
 address-family l2vpn evpn
  neighbor fabric activate
  advertise-all-vni
 exit-address-family
```

Spine 1 Configuration

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip addr add 3.3.3.3/32 dev lo
```

FRR configuration

```
router bgp 65020
 bgp router-id 3.3.3.3
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal
Fabric Network
 neighbor fabric capability extended-
nexthop
 neighbor swp1 interface peer-group
fabric
 neighbor swp2 interface peer-group
fabric
 !
 address-family ipv4 unicast
  network 3.3.3.3/32
 exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
!
address-family l2vpn evpn
  neighbor fabric activate
 exit-address-family
!
```

Spine 2 Configuration

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip addr add 4.4.4.4/32 dev lo
```

FRR configuration

```
router bgp 65020
 bgp router-id 4.4.4.4
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal
Fabric Network
 neighbor fabric capability extended-
nexthop
 neighbor swp1 interface peer-group
fabric
 neighbor swp2 interface peer-group
fabric
!
 address-family ipv4 unicast
  network 4.4.4.4/32
 exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
!
address-family l2vpn evpn
  neighbor fabric activate
 exit-address-family
!
```

3.2.1 Fabric Troubleshooting

1. Verify all IP links are up and configured

```
#ip link
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN mode DEFAULT
group default qlen 1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
10: swp1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc fq_codel state UP mode
DEFAULT group default qlen 1000
    link/ether 50:6b:4b:88:0f:19 brd ff:ff:ff:ff:ff:ff
11: swp2: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc fq_codel state UP mode
DEFAULT group default qlen 1000
    link/ether 50:6b:4b:88:0f:1d brd ff:ff:ff:ff:ff:ff
17: swp9: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc fq_codel master br0 state
UP mode DEFAULT group default qlen 1000  link/ether 50:6b:4b:88:0f:25 brd
ff:ff:ff:ff:ff:ff
20: vx10010: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue master br0
state UNKNOWN mode DEFAULT group default qlen 1000
    link/ether ae:93:4d:6a:7b:8b brd ff:ff:ff:ff:ff:ff
21: br0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue state UP mode
DEFAULT group default qlen 1000
    link/ether 50:6b:4b:88:0f:25 brd ff:ff:ff:ff:ff:ff
```

2. Verify the bridge VLAN setting

```
#bridge vlan show
port    vlan ids
swp9
         1 untagged      10 PVID untagged
vx10010
         1 untagged      10 PVID untagged
br0
         1 PVID untagged
```

3. Verify BGP neighbors are ESTABLISHED

```
IPv4 Unicast Summary:
BGP router identifier 1.1.1.1, local AS number 65011 vrf-id 0
BGP table version 9
RIB entries 13, using 2080 bytes of memory
Peers 2, using 41 KiB of memory
Peer groups 1, using 64 bytes of memory

Neighbor        V         AS MsgRcvd MsgSent   TblVer  InQ OutQ  Up/Down State/PfxRcd
swp1            4      65020     228     233        0    0    0 03:39:23            4
swp2            4      65020     229     233        0    0    0 03:39:24            4

Total number of neighbors 2
```

4. Verify the routing table

```
#ip route
2.2.2.2 proto bgp metric 20
   nexthop via 169.254.0.1 dev swp1 weight 1 onlink offload
   nexthop via 169.254.0.1 dev swp2 weight 1 onlink offload
3.3.3.3 via 169.254.0.1 dev swp1 proto bgp metric 20 onlink offload
4.4.4.4 via 169.254.0.1 dev swp2 proto bgp metric 20 onlink offload
```

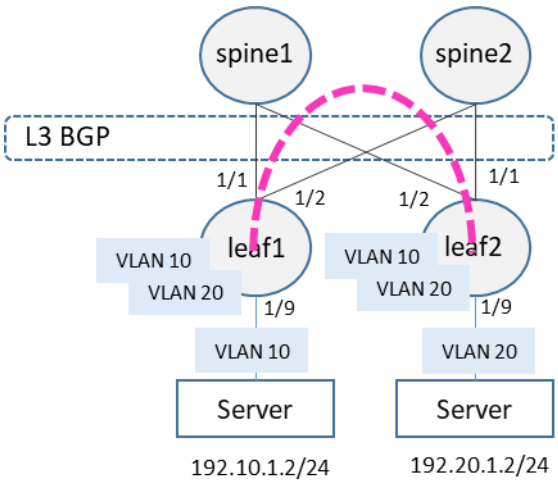
5. Verify the EVPN table

```
#vtysh -c 'show bgp l2vpn evpn'
Route Distinguisher: ip 1.1.1.1:2

*> [2]:[0]:[48]:[00:02:c9:1c:69:30]
                    1.1.1.1                           32768 i
*> [3]:[0]:[32]:[1.1.1.1]
                    1.1.1.1                           32768 i
Route Distinguisher: ip 2.2.2.2:2

*  [2]:[0]:[48]:[00:02:c9:1c:62:40]
                    2.2.2.2                               0 65020 65012 i
*> [2]:[0]:[48]:[00:02:c9:1c:62:40]
                    2.2.2.2                               0 65020 65012 i
*  [3]:[0]:[32]:[2.2.2.2]
                    2.2.2.2                               0 65020 65012 i
*> [3]:[0]:[32]:[2.2.2.2]
                    2.2.2.2                               0 65020 65012 i

Displayed 6 out of 6 total prefixes
```

6. Verify the MAC address table

```
#bridge fdb show vlan 10
00:02:c9:1c:69:30 dev swp9 vlan 10 extern_learn offload br0
50:6b:4b:88:0f:25 dev swp9 vlan 10 br0 permanent
00:02:c9:1c:62:40 dev vx10010 vlan 10 extern_learn offload br0
ae:93:4d:6a:7b:8b dev vx10010 vlan 10 br0 permanent
```

3.3 EVPN Based VxLAN with Asymmetric Routing



Leaf 1 Configuration

1. Configure the physical links

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip link set dev swp9 up mtu 9216
```

2. Configure the loopback interface

```
ip addr add 1.1.1.1/32 dev lo
```

3. Create VxLAN interface for VLAN 10 and VLAN 20

```
ip link add name vx10010 up mtu 9216 type vxlan id 10 local 1.1.1.1 dstport 4789
nolearning noudpcsum tos inherit ttl lo0
ip link add name vx10020 up mtu 9216 type vxlan id 20 local 1.1.1.1 dstport 4789
nolearning noudpcsum tos inherit ttl lo0
```

4. Configure a vlan-aware bridge and set its MAC address to be the same as one of the switch ports

```
ip link add name br0 type bridge vlan_filtering 1 vlan_default_pvid 0 mcast_snooping
0
ip link set dev br0 address <swp1's MAC address>
```

5. Assign the host facing port and VxLAN interfaces as members of the bridge

```
ip link set dev swp9 master br0
ip link set dev vx10010 master br0
ip link set dev vx10020 master br0
```

6. Add VLAN 10 and 20 as access VLAN of the host facing port and VxLAN interface

```
bridge vlan add vid 10 dev swp9 untagged pvid
bridge vlan add vid 10 dev vx10010 untagged pvid
bridge vlan add vid 20 dev vx10020 untagged pvid
ip link set dev br0 up
```

7. Configure SVI for VLAN 10 and 20 with Anycast-GW address

```
ip link add link br0 name vlan10 up mtu 9216 type vlan id 10
ip address add 192.10.1.11/24 dev vlan10
ip link add link vlan10 name vlan10-v up mtu 9216 address 00:00:5e:00:00:10 type
macvlan mode private
```

```
ip address add 192.10.1.1/24 dev vlan10-v metric 1024
ip link add link br0 name vlan20 up mtu 9216 type vlan id 20
ip address add 192.20.1.11/24 dev vlan20
ip link add link vlan20 name vlan20-v up mtu 9216 address 00:00:5e:00:00:20 type
macvlan mode private
ip address add 192.20.1.1/24 dev vlan20-v metric 1024
bridge vlan add vid 10 dev br0 self
bridge vlan add vid 20 dev br0 self
```

8. Add the Anycast MAC address to the bridge

```
bridge fdb add 00:00:5e:00:00:10 dev br0 self local vlan 10
bridge fdb add 00:00:5e:00:00:20 dev br0 self local vlan 20
```

9. Disable rp_filter and enable arp_ignore to make sure ARPs for the anycast IP are answered with the anycast MAC

```
sysctl -w net.ipv4.conf.all.rp_filter=0
sysctl -w net.ipv4.conf.vlan10-v.rp_filter=0
sysctl -w net.ipv4.conf.vlan20-v.rp_filter=0
sysctl -w net.ipv4.conf.all.arp_ignore=1
```

10. Configure ARP suppression

```
ip link set dev vx10010 type bridge_slave neigh_suppress on
ip link set dev vx10020 type bridge_slave neigh_suppress on
```

11. Under FRR, configure unnumbered BGP neighbors, advertise the loopback and all EVPN VNI

```
router bgp 65011
 bgp router-id 1.1.1.1
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal Fabric Network
 neighbor fabric capability extended-nexthop
 neighbor swp1 interface peer-group fabric
 neighbor swp2 interface peer-group fabric
 !
address-family ipv4 unicast
  network 1.1.1.1/32
exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
 !
address-family l2vpn evpn
  neighbor fabric activate
  advertise-all-vni
 exit-address-family
 !
```

Leaf 2 Configuration

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip link set dev swp9 up mtu 9216
ip addr add 2.2.2.2/32 dev lo
ip link add name vx10010 up mtu 9216 type vxlan id 10 local 2.2.2.2 dstport 4789
nolearning noudpcsum tos inherit ttl 100
ip link add name vx10020 up mtu 9216 type vxlan id 20 local 2.2.2.2 dstport 4789
nolearning noudpcsum tos inherit ttl 100
ip link add name br0 type bridge vlan_filtering 1 vlan_default_pvid 0 mcast_snooping
0
ip link set dev br0 address <swp1's MAC address>
ip link set dev swp9 master br0
ip link set dev vx10010 master br0
ip link set dev vx10020 master br0
bridge vlan add vid 20 dev swp9 untagged pvid
bridge vlan add vid 10 dev vx10010 untagged pvid
bridge vlan add vid 20 dev vx10020 untagged pvid
ip link set dev br0 up
ip link add link br0 name vlan10 up mtu 9216 type vlan id 10
ip address add 192.10.1.12/24 dev vlan10
ip link add link vlan10 name vlan10-v up mtu 9216 address 00:00:5e:00:00:10 type
macvlan mode private
ip address add 192.10.1.1/24 dev vlan10-v metric 1024
ip link add link br0 name vlan20 up mtu 9216 type vlan id 20
ip address add 192.20.1.12/24 dev vlan20
ip link add link vlan20 name vlan20-v up mtu 9216 address 00:00:5e:00:00:20 type
macvlan mode private
ip address add 192.20.1.1/24 dev vlan20-v metric 1024
bridge vlan add vid 10 dev br0 self
bridge vlan add vid 20 dev br0 self
bridge fdb add 00:00:5e:00:00:10 dev br0 self local vlan 10
bridge fdb add 00:00:5e:00:00:20 dev br0 self local vlan 20
sysctl -w net.ipv4.conf.all.rp_filter=0
sysctl -w net.ipv4.conf.vlan10-v.rp_filter=0
sysctl -w net.ipv4.conf.vlan20-v.rp_filter=0
sysctl -w net.ipv4.conf.all.arp_ignore=1
ip link set dev vx10010 type bridge_slave neigh_suppress on
ip link set dev vx10020 type bridge_slave neigh_suppress on
```

FRR configuration

```
router bgp 65012
 bgp router-id 2.2.2.2
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal Fabric Network
 neighbor fabric capability extended-nexthop
 neighbor swp1 interface peer-group fabric
 neighbor swp2 interface peer-group fabric
 !
 address-family ipv4 unicast
  network 2.2.2.2/32
exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
 !
 address-family l2vpn evpn
  neighbor fabric activate
  advertise-all-vni
 exit-address-family
 !
```

Spine 1 Configuration

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip addr add 3.3.3.3/32 dev lo
```

FRR configuration

```
router bgp 65020
 bgp router-id 3.3.3.3
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal
Fabric Network
 neighbor fabric capability extended-
nexthop
 neighbor swp1 interface peer-group
fabric
 neighbor swp2 interface peer-group
fabric
 !
 address-family ipv4 unicast
  network 3.3.3.3/32
 exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
 !
 address-family l2vpn evpn
  neighbor fabric activate
 exit-address-family
 !
```

Spine 2 Configuration

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip addr add 4.4.4.4/32 dev lo
```

FRR configuration

```
router bgp 65020
 bgp router-id 4.4.4.4
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal
Fabric Network
 neighbor fabric capability extended-
nexthop
 neighbor swp1 interface peer-group
fabric
 neighbor swp2 interface peer-group
fabric
 !
 address-family ipv4 unicast
  network 4.4.4.4/32
 exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
 !
 address-family l2vpn evpn
  neighbor fabric activate
 exit-address-family
 !
```

Host 1 Configuration

```
ip route add 192.20.1.0/24 via
192.10.1.1
```

Host 2 Configuration

```
ip route add 192.10.1.0/24 via
192.20.1.1
```

3.3.1 Fabric Troubleshooting

1. Verify all IP links are up and configured:

```
#ip link
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN mode DEFAULT
group default qlen 1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
10: swp1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc fq_codel state UP mode
DEFAULT group default qlen 1000
    link/ether 50:6b:4b:88:0f:19 brd ff:ff:ff:ff:ff:ff
11: swp2: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc fq_codel state UP mode
DEFAULT group default qlen 1000
    link/ether 50:6b:4b:88:0f:1d brd ff:ff:ff:ff:ff:ff
13: swp9: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc fq_codel master br0 state
UP mode DEFAULT group default qlen 1000  link/ether 50:6b:4b:88:0f:25 brd
ff:ff:ff:ff:ff:ff
20: vx10010: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue master br0
state UNKNOWN mode DEFAULT group default qlen 1000 link/ether 3a:2a:68:09:4f:ad brd
ff:ff:ff:ff:ff:ff
21: vx10020: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue master br0
state UNKNOWN mode DEFAULT group default qlen 1000 link/ether 16:8a:3e:d5:0f:90 brd
ff:ff:ff:ff:ff:ff
22: br0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue state UP mode
DEFAULT group default qlen 1000
    link/ether 50:6b:4b:88:0f:19 brd ff:ff:ff:ff:ff:ff
23: vlan10@br0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue state UP
mode DEFAULT group default qlen 1000 link/ether 50:6b:4b:88:0f:19 brd
ff:ff:ff:ff:ff:ff
24: vlan10-v@vlan10: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue state
UP mode DEFAULT group default qlen 1000 link/ether 00:00:5e:00:00:10 brd
ff:ff:ff:ff:ff:ff
25: vlan20@br0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue state UP
mode DEFAULT group default qlen 1000 link/ether 50:6b:4b:88:0f:19 brd
ff:ff:ff:ff:ff:ff
26: vlan20-v@vlan20: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue state
UP mode DEFAULT group default qlen 1000 link/ether 00:00:5e:00:00:20 brd
ff:ff:ff:ff:ff:ff
```

2. Verify the bridge Vlan setting:

```
#bridge vlan show
port     vlan ids
swp9
            10 PVID untagged
vx10010
            10 PVID untagged
vx10020
            20 PVID untagged
br0
            10      20
```

3. Verify BGP neighbors are up:

```
#vtysh -c 'show ip bgp summary'

IPv4 Unicast Summary:
BGP router identifier 1.1.1.1, local AS number 65011 vrf-id 0
BGP table version 9
RIB entries 13, using 2080 bytes of memory
Peers 2, using 41 KiB of memory
Peer groups 1, using 64 bytes of memory

Neighbor        V         AS MsgRcvd MsgSent   TblVer  InQ OutQ  Up/Down State/PfxRcd
swp1            4      65020     228     233        0    0    0 03:39:23            4
swp2            4      65020     229     233        0    0    0 03:39:24            4
```

```
Total number of neighbors 2
```

4. Verify routing table

```
#ip route
2.2.2.2 proto bgp metric 20
   nexthop via 169.254.0.1 dev swp1 weight 1 onlink offload
   nexthop via 169.254.0.1 dev swp2 weight 1 onlink offload
3.3.3.3 via 169.254.0.1 dev swp1 proto bgp metric 20 onlink offload
4.4.4.4 via 169.254.0.1 dev swp2 proto bgp metric 20 onlink offload
192.10.1.0/24 dev vlan10 proto kernel scope link src 192.10.1.11 offload
192.10.1.0/24 dev vlan10-v proto kernel scope link src 192.10.1.1 metric 1024
192.20.1.0/24 dev vlan20 proto kernel scope link src 192.20.1.11 offload
192.20.1.0/24 dev vlan20-v proto kernel scope link src 192.20.1.1 metric 1024
```

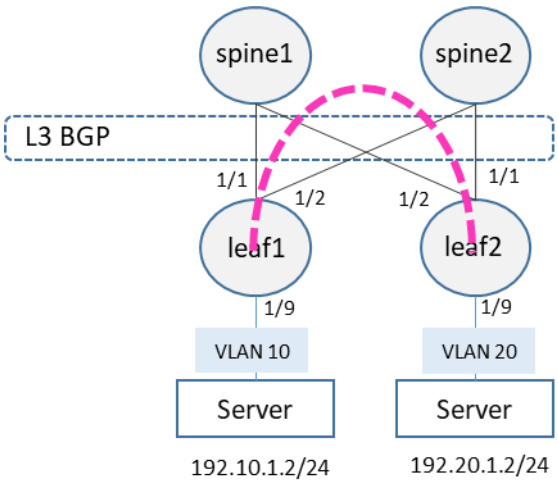
5. Verify the EVPN table

```
#vtysh -c 'show bgp l2vpn evpn'
Route Distinguisher: ip 1.1.1.1:2

*> [2]:[0]:[48]:[00:02:c9:1c:69:30]
                    1.1.1.1                             32768 i
*> [2]:[0]:[48]:[00:02:c9:1c:69:30]:[32]:[192.10.1.2]
                    1.1.1.1                             32768 i
*> [3]:[0]:[32]:[1.1.1.1]
                    1.1.1.1                             32768 i
Route Distinguisher: ip 1.1.1.1:3

*> [3]:[0]:[32]:[1.1.1.1]
                    1.1.1.1                             32768 i
Route Distinguisher: ip 2.2.2.2:2

*  [3]:[0]:[32]:[2.2.2.2]
                    2.2.2.2                                 0 65020 65012 i
*> [3]:[0]:[32]:[2.2.2.2]
                    2.2.2.2                                 0 65020 65012 i
Route Distinguisher: ip 2.2.2.2:3

*  [2]:[0]:[48]:[00:02:c9:1c:62:40]
                    2.2.2.2                                 0 65020 65012 i
*> [2]:[0]:[48]:[00:02:c9:1c:62:40]
                    2.2.2.2                                 0 65020 65012 i
*  [2]:[0]:[48]:[00:02:c9:1c:62:40]:[32]:[192.20.1.2]
                    2.2.2.2                                 0 65020 65012 i
*> [2]:[0]:[48]:[00:02:c9:1c:62:40]:[32]:[192.20.1.2]
                    2.2.2.2                                 0 65020 65012 i
*  [3]:[0]:[32]:[2.2.2.2]
                    2.2.2.2                                 0 65020 65012 i
*> [3]:[0]:[32]:[2.2.2.2]
                    2.2.2.2                                 0 65020 65012 i

Displayed 12 out of 12 total prefixes
```

6. Verify the MAC address table:

```
#bridge fdb show vlan 10
00:02:c9:1c:69:30 dev swp9 vlan 10 extern_learn offload br0
50:6b:4b:88:0f:25 dev swp9 vlan 10 br0 permanent
3a:2a:68:09:4f:ad dev vx10010 vlan 10 br0 permanent
50:6b:4b:88:0f:19 dev br0 vlan 10 br0 permanent

# bridge fdb show vlan 20
00:02:c9:1c:62:40 dev vx10020 vlan 20 extern_learn offload br0
16:8a:3e:d5:0f:90 dev vx10020 vlan 20 br0 permanent
50:6b:4b:88:0f:19 dev br0 vlan 20 br0 permanent
```

3.4 EVPN Based VxLAN with Symmetric Routing



Leaf 1 Configuration

1. Configure the physical links

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip link set dev swp9 up mtu 9216
```

2. Configure the loopback interface

```
ip addr add 1.1.1.1/32 dev lo
```

3. Create VxLAN interface

```
ip link add name vx10010 up mtu 9216 type vxlan id 10 local 1.1.1.1 dstport 4789
nolearning noudpcsum tos inherit ttl lo0
```

4. Create VxLAN interface for L3VNI

```
ip link add name vx4001 up mtu 9216 type vxlan id 104001 local 1.1.1.1 dstport 4789
nolearning noudpcsum tos inherit ttl lo0
```

5. Configure a vlan-aware bridge and set its MAC address to be the same as one of the switch ports

```
ip link add name br0 type bridge vlan_filtering 1 vlan_default_pvid 0 mcast_snooping
0
ip link set dev br0 address <swp1's MAC address>
ip link set dev br0 up
```

6. Assign the host facing port and VxLAN interfaces as members of the bridge

```
ip link set dev swp9 master br0
ip link set dev vx10010 master br0
ip link set dev vx4001 master br0
```

7. Create a tenant VRF

```
ip link add dev vrf1 up type vrf table 10
ip -4 route add table 10 unreachable default metric 4278198272
ip -6 route add table 10 unreachable default metric 4278198272
ip -4 rule add pref 32765 table local
ip -4 rule del pref 0
ip -6 rule add pref 32765 table local
ip -6 rule del pref 0
```

8. Add VLAN 10 as access VLAN of the host facing port and VxLAN interface

```
bridge vlan add vid 10 dev swp9 untagged pvid
bridge vlan add vid 10 dev vx10010 untagged pvid
```

9. Configure SVI for VLAN 10 with Anycast-GW address

```
ip link add link br0 name vlan10 up master vrf1 mtu 9216 type vlan id 10
ip address add 192.10.1.11/24 dev vlan10
ip link add link vlan10 name vlan10-v up master vrf1 mtu 9216 address
00:00:5e:00:00:10 type macvlan mode private
ip address add 192.10.1.1/24 dev vlan10-v metric 1024
bridge vlan add vid 10 dev br0 self
```

10. Add the Anycast MAC address to the bridge

```
bridge fdb add 00:00:5e:00:00:10 dev br0 self local vlan 10
```

11. Add VLAN 4001 as access VLAN of the L3VNI interface

```
bridge vlan add vid 4001 dev vx4001 untagged pvid
```

12. Configure SVI for VLAN 4001

```
ip link add link br0 name vlan4001 up master vrf1 mtu 9216 type vlan id 4001
bridge vlan add vid 4001 dev br0 self
```

13. Disable rp_filter and enable arp_ignore to make sure ARPs for the anycast IP are answered with the anycast MAC

```
sysctl -w net.ipv4.conf.all.rp_filter=0
sysctl -w net.ipv4.conf.vlan10-v.rp_filter=0
sysctl -w net.ipv4.conf.all.arp_ignore=1
```

14. Configure ARP suppression

```
ip link set dev vx10010 type bridge_slave neigh_suppress on
```

15. Under FRR, configure unnumbered BGP neighbors and advertise the loopback and all EVPN VNI

```
router bgp 65011
 bgp router-id 1.1.1.1
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal Fabric Network
 neighbor fabric capability extended-nexthop
 neighbor swp1 interface peer-group fabric
 neighbor swp2 interface peer-group fabric
 !
 address-family ipv4 unicast
  network 1.1.1.1/32
exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
 !
 address-family l2vpn evpn
  neighbor fabric activate
  advertise-all-vni
 exit-address-family
```

16. Under FRR, map the L3VNI to the tenant VRF

```
vrf vrf1
 vni 104001
 exit-vrf
!
```

Leaf 2 Configuration

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip link set dev swp9 up mtu 9216
ip addr add 2.2.2.2/32 dev lo
ip link add name vx10020 up mtu 9216 type vxlan id 20 local 2.2.2.2 dstport 4789
nolearning noudpcsum tos inherit ttl 100
ip link add name vx4001 up mtu 9216 type vxlan id 104001 local 2.2.2.2 dstport 4789
nolearning noudpcsum tos inherit ttl 100
ip link add name br0 type bridge vlan_filtering 1 vlan_default_pvid 0 mcast_snooping
0
ip link set dev br0 address <swp1's MAC address>
ip link set dev br0 up
ip link set dev swp9 master br0
ip link set dev vx10020 master br0
ip link set dev vx4001 master br0
ip link add dev vrf1 up type vrf table 10
ip -4 route add table 10 unreachable default metric 4278198272
ip -6 route add table 10 unreachable default metric 4278198272
ip -4 rule add pref 32765 table local
ip -4 rule del pref 0
ip -6 rule add pref 32765 table local
ip -6 rule del pref 0
bridge vlan add vid 20 dev swp9 untagged pvid
bridge vlan add vid 20 dev vx10020 untagged pvid
ip link add link br0 name vlan20 up master vrf1mtu 9216 type vlan id 20
ip address add 192.20.1.12/24 dev vlan20
ip link add link vlan20 name vlan20-v up master vrf1 mtu 9216 address
00:00:5e:00:00:20 type macvlan mode private
ip address add 192.20.1.1/24 dev vlan20-v metric 1024
bridge vlan add vid 20 dev br0 self
bridge fdb add 00:00:5e:00:00:20 dev br0 self local vlan 20
bridge vlan add vid 4001 dev vx4001 untagged pvid
ip link add link br0 name vlan4001 up master vrf1 mtu 9216 type vlan id 4001
bridge vlan add vid 4001 dev br0 self
sysctl -w net.ipv4.conf.all.rp_filter=0
sysctl -w net.ipv4.conf.vlan20-v.rp_filter=0
sysctl -w net.ipv4.conf.all.arp_ignore=1
ip link set dev vx10020 type bridge_slave neigh_suppress on
```

FRR configuration

```
vrf vrf1
 vni 104001
 exit-vrf
!
router bgp 65012
 bgp router-id 2.2.2.2
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal Fabric Network
 neighbor fabric capability extended-nexthop
 neighbor swp1 interface peer-group fabric
 neighbor swp2 interface peer-group fabric
 !
address-family ipv4 unicast
  network 2.2.2.2/32
```

```
exit-address-family
 !
address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
 !
 address-family l2vpn evpn
  neighbor fabric activate
  advertise-all-vni
 exit-address-family
```

Spine 1 Configuration

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip addr add 3.3.3.3/32 dev lo
```

FRR configuration

```
router bgp 65020
 bgp router-id 3.3.3.3
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal
Fabric Network
 neighbor fabric capability extended-
nexthop
 neighbor swp1 interface peer-group
fabric
 neighbor swp2 interface peer-group
fabric
 !
 address-family ipv4 unicast
  network 3.3.3.3/32
 exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
 !
 address-family l2vpn evpn
  neighbor fabric activate
 exit-address-family
 !
```

Spine 2 Configuration

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip addr add 4.4.4.4/32 dev lo
```

FRR configuration

```
router bgp 65020
 bgp router-id 4.4.4.4
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal
Fabric Network
 neighbor fabric capability extended-
nexthop
 neighbor swp1 interface peer-group
fabric
 neighbor swp2 interface peer-group
fabric
 !
 address-family ipv4 unicast
  network 4.4.4.4/32
 exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
 !
 address-family l2vpn evpn
  neighbor fabric activate
 exit-address-family
 !
```

Host 1 Configuration

```
ip route add 192.20.1.0/24 via
192.10.1.1
```

Host 2 Configuration

```
ip route add 192.10.1.0/24 via
192.20.1.1
```

3.4.1 Fabric Troubleshooting

1. Verify all IP links are up and configured:

```
#ip link
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN mode DEFAULT
group default qlen 1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
10: swp1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc fq_codel state UP mode
DEFAULT group default qlen 1000
    link/ether 50:6b:4b:88:0f:19 brd ff:ff:ff:ff:ff:ff
11: swp2: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc fq_codel state UP mode
DEFAULT group default qlen 1000
    link/ether 50:6b:4b:88:0f:1d brd ff:ff:ff:ff:ff:ff
13: swp9: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc fq_codel master br0 state
UP mode DEFAULT group default qlen 1000  link/ether 50:6b:4b:88:0f:25 brd
ff:ff:ff:ff:ff:ff
20: vx10010: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue master br0
state UNKNOWN mode DEFAULT group default qlen 1000 link/ether 82:d4:17:dd:f5:e6 brd
ff:ff:ff:ff:ff:ff
21: vx4001: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue master br0
state UNKNOWN mode DEFAULT group default qlen 1000 link/ether 0e:f4:16:c8:41:05 brd
ff:ff:ff:ff:ff:ff
22: br0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue state UP mode
DEFAULT group default qlen 1000
    link/ether 50:6b:4b:88:0f:19 brd ff:ff:ff:ff:ff:ff
23: vrf1: <NOARP,MASTER,UP,LOWER_UP> mtu 65536 qdisc noqueue state UP mode DEFAULT
group default qlen 1000
    link/ether 7a:6c:31:41:2a:7a brd ff:ff:ff:ff:ff:ff
24: vlan10@br0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue master vrf1
state UP mode DEFAULT group default qlen 1000 link/ether 50:6b:4b:88:0f:19 brd
ff:ff:ff:ff:ff:ff
25: vlan10-v@vlan10: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue master
vrf1 state UP mode DEFAULT group default qlen 1000 link/ether 00:00:5e:00:00:10 brd
ff:ff:ff:ff:ff:ff
26: vlan4001@br0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue master
vrf1 state UP mode DEFAULT group default qlen 1000 link/ether 50:6b:4b:88:0f:19 brd
ff:ff:ff:ff:ff:ff
```

2. Verify the bridge VLAN setting:

```
#bridge vlan show
port    vlan ids
swp9
         10 PVID untagged
vx10010
         10 PVID untagged
vx4001
         4001 PVID untagged
br0
         10      4001
```

3. Verify BGP neighbors are up:

```
#vtysh -c 'show ip bgp summary'

IPv4 Unicast Summary:
BGP router identifier 1.1.1.1, local AS number 65011 vrf-id 0
BGP table version 9
RIB entries 13, using 2080 bytes of memory
Peers 2, using 41 KiB of memory
Peer groups 1, using 64 bytes of memory

Neighbor        V         AS MsgRcvd MsgSent   TblVer  InQ OutQ  Up/Down State/PfxRcd
swp1            4      65020     228     233        0    0    0 03:39:23            4
swp2            4      65020     229     233        0    0    0 03:39:24            4
```

```
Total number of neighbors 2
```

4. Verify routing table:

```
#ip route
2.2.2.2 proto bgp metric 20
   nexthop via 169.254.0.1 dev swp1 weight 1 onlink offload
   nexthop via 169.254.0.1 dev swp2 weight 1 onlink offload
3.3.3.3 via 169.254.0.1 dev swp1 proto bgp metric 20 onlink offload
4.4.4.4 via 169.254.0.1 dev swp2 proto bgp metric 20 onlink offload

# ip route list vrf vrf1
unreachable default metric 4278198272
192.10.1.0/24 dev vlan10 proto kernel scope link src 192.10.1.11 offload
192.10.1.0/24 dev vlan10-v proto kernel scope link src 192.10.1.1 metric 1024
192.20.1.2 via 2.2.2.2 dev vlan4001 proto bgp metric 20 onlink offload
```

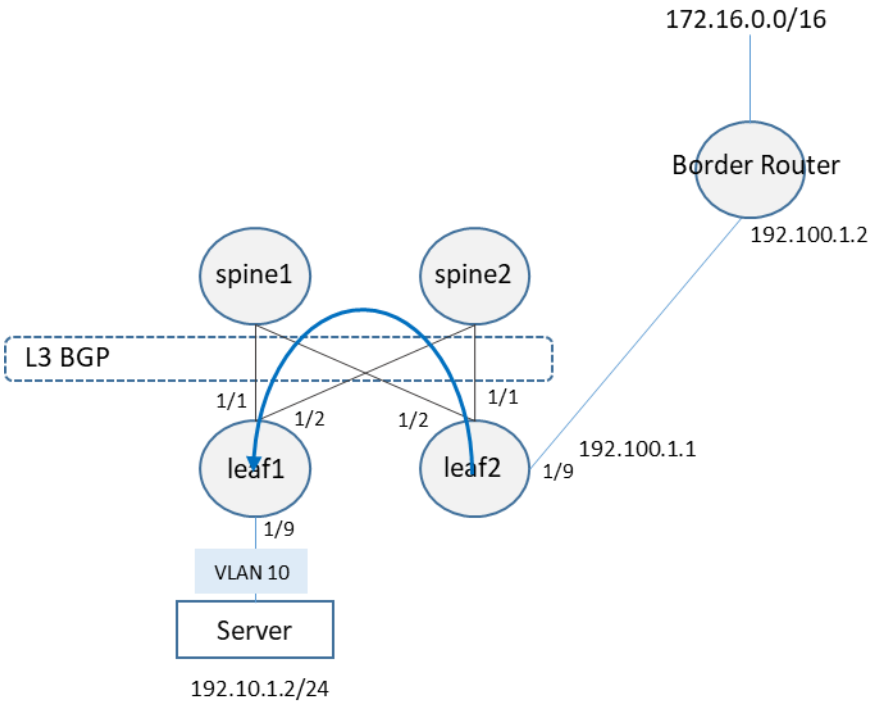
5. Verify the EVPN table

```
#vtysh -c 'show bgp l2vpn evpn'
Route Distinguisher: ip 1.1.1.1:2

*> [2]:[0]:[48]:[00:02:c9:1c:69:30]
                    1.1.1.1                            32768 i
*> [2]:[0]:[48]:[00:02:c9:1c:69:30]:[32]:[192.10.1.2]
                    1.1.1.1                            32768 i
*> [3]:[0]:[32]:[1.1.1.1]
                    1.1.1.1                            32768 i
Route Distinguisher: ip 2.2.2.2:2

*  [2]:[0]:[48]:[00:02:c9:1c:62:40]
                    2.2.2.2                                0 65020 65012 i
*> [2]:[0]:[48]:[00:02:c9:1c:62:40]
                    2.2.2.2                                0 65020 65012 i
*  [2]:[0]:[48]:[00:02:c9:1c:62:40]:[32]:[192.20.1.2]
                    2.2.2.2                                0 65020 65012 i
*> [2]:[0]:[48]:[00:02:c9:1c:62:40]:[32]:[192.20.1.2]
                    2.2.2.2                                0 65020 65012 i
*  [3]:[0]:[32]:[2.2.2.2]
                    2.2.2.2                                0 65020 65012 i
*> [3]:[0]:[32]:[2.2.2.2]
                    2.2.2.2                                0 65020 65012 i

Displayed 9 out of 9 total prefixes
```

6. Verify the MAC address table:

```
# bridge fdb | grep extern_learn
00:02:c9:1c:69:30 dev swp9 vlan 10 extern_learn offload br0
ec:0d:9a:5d:eb:d9 dev vx4001 vlan 4001 extern_learn offload br0
ec:0d:9a:5d:eb:d9 dev vx4001 2.2.2.2 self extern_learn offload
```

3.5 EVPN Based VxLAN with Symmetric Routing (Type5)



Leaf 1 Configuration

1. Configure the physical links

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip link set dev swp9 up mtu 9216
```

2. Configure the loopback interface

```
ip addr add 1.1.1.1/32 dev lo
```

3. Create VxLAN interface for VLAN 10

```
ip link add name vx10010 up mtu 9216 type vxlan id 10 local 1.1.1.1 dstport 4789
nolearning noudpcsum tos inherit ttl lo0
```

4. Create VxLAN interface for L3VNI

```
ip link add name vx4001 up mtu 9216 type vxlan id 104001 local 1.1.1.1 dstport 4789
nolearning noudpcsum tos inherit ttl lo0
```

5. Configure a vlan-aware bridge and set its MAC address to be the same as one of the switch ports

```
ip link add name br0 type bridge vlan_filtering 1 vlan_default_pvid 0 mcast_snooping
0
ip link set dev br0 address <swp1's MAC address>
ip link set dev br0 up
```

6. Assign the host facing port and VxLAN interfaces as members of the bridge

```
ip link set dev swp9 master br0
ip link set dev vx10010 master br0
ip link set dev vx4001 master br0
```

7. Create a tenant VRF

```
ip link add dev vrf1 up type vrf table 10
ip -4 route add table 10 unreachable default metric 4278198272
ip -6 route add table 10 unreachable default metric 4278198272
ip -4 rule add pref 32765 table local
ip -4 rule del pref 0
ip -6 rule add pref 32765 table local
ip -6 rule del pref 0
```

8. Add VLAN 10 as access VLAN of host facing port and VxLAN interface

```
bridge vlan add vid 10 dev swp9 untagged pvid
bridge vlan add vid 10 dev vx10010 untagged pvid
```

9. Configure SVI for VLAN 10 with Anycast-GW address

```
ip link add link br0 name vlan10 up master vrf1 mtu 9216 type vlan id 10
ip address add 192.10.1.11/24 dev vlan10
ip link add link vlan10 name vlan10-v up master vrf1 mtu 9216 address
00:00:5e:00:00:10 type macvlan mode private
ip address add 192.10.1.1/24 dev vlan10-v metric 1024
bridge vlan add vid 10 dev br0 self
```

10. Add the Anycast MAC address to the bridge

```
bridge fdb add 00:00:5e:00:00:10 dev br0 self local vlan 10
```

11. Add VLAN 4001 as access VLAN of L3VNI interface

```
bridge vlan add vid 4001 dev vx4001 untagged pvid
```

12. Configure SVI for VLAN 4001

```
ip link add link br0 name vlan4001 up master vrf1 mtu 9216 type vlan id 4001
bridge vlan add vid 4001 dev br0 self
```

13. Disable rp_filter and enable arp_ignore to make sure ARPs for the anycast IP are answered with the anycast MAC

```
sysctl -w net.ipv4.conf.all.rp_filter=0
sysctl -w net.ipv4.conf.vlan10-v.rp_filter=0
sysctl -w net.ipv4.conf.all.arp_ignore=1
```

14. Configure ARP suppression

```
ip link set dev vx10010 type bridge_slave neigh_suppress on
```

15. Under FRR, configure unnumbered BGP neighbors, advertise the loopback and all EVPN VNI

```
router bgp 65011
 bgp router-id 1.1.1.1
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal Fabric Network
 neighbor fabric capability extended-nexthop
 neighbor swp1 interface peer-group fabric
 neighbor swp2 interface peer-group fabric
 !
 address-family ipv4 unicast
  network 1.1.1.1/32
exit-address-family
 !
 address-family ipv6 unicast
```

```
  neighbor fabric activate
 exit-address-family
 !
 address-family l2vpn evpn
  neighbor fabric activate
  advertise-all-vni
 exit-address-family
```

16. Under FRR, map L3VNI to tenant VRF

```
vrf vrf1
 vni 104001
 exit-vrf
!
```

Leaf 2 Configuration

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip link set dev swp9 up mtu 9216
ip address add 192.100.1.1/24 dev swp9
ip addr add 2.2.2.2/32 dev lo
ip link add name vx4001 up mtu 9216 type vxlan id 104001 local 2.2.2.2 dstport 4789
nolearning noudpcsum tos inherit ttl 100
ip link add name br0 type bridge vlan_filtering 1 vlan_default_pvid 0 mcast_snooping
0
ip link set dev br0 address <swp1's MAC address>
ip link set dev br0 up
ip link set dev vx4001 master br0
ip link add dev vrf1 up type vrf table 10
ip -4 route add table 10 unreachable default metric 4278198272
ip -6 route add table 10 unreachable default metric 4278198272
ip -4 rule add pref 32765 table local
ip -4 rule del pref 0
ip -6 rule add pref 32765 table local
ip -6 rule del pref 0
ip link set dev swp9 master vrf1
bridge vlan add vid 4001 dev vx4001 untagged pvid
ip link add link br0 name vlan4001 up master vrf1 mtu 9216 type vlan id 4001
bridge vlan add vid 4001 dev br0 self
```

FRR configuration

```
vrf vrf1
 ip route 172.16.0.0/16 192.100.1.2
 vni 104001
 exit-vrf
!
router bgp 65012
bgp router-id 2.2.2.2
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal Fabric Network
 neighbor fabric capability extended-nexthop
 neighbor swp1 interface peer-group fabric
 neighbor swp2 interface peer-group fabric
!
 address-family ipv4 unicast
  network 2.2.2.2/32
exit-address-family
 !
address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
 !
```

```
address-family l2vpn evpn
  neighbor fabric activate
  advertise-all-vni
exit-address-family
!
router bgp 65012 vrf vrf1
 bgp router-id 2.2.2.2
 !
address-family ipv4 unicast
  redistribute static
 exit-address-family
 !
address-family l2vpn evpn
  advertise ipv4 unicast
 exit-address-family
!
```

Spine 1 Configuration

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip addr add 3.3.3.3/32 dev lo
```

FRR configuration

```
router bgp 65020
 bgp router-id 3.3.3.3
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal
Fabric Network
 neighbor fabric capability extended-
nexthop
 neighbor swp1 interface peer-group
fabric
 neighbor swp2 interface peer-group
fabric
 !
 address-family ipv4 unicast
  network 3.3.3.3/32
 exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
 !
 address-family l2vpn evpn
  neighbor fabric activate
 exit-address-family
 !
```

Spine 2 Configuration

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip addr add 4.4.4.4/32 dev lo
```

FRR configuration

```
router bgp 65020
 bgp router-id 4.4.4.4
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal
Fabric Network
 neighbor fabric capability extended-
nexthop
 neighbor swp1 interface peer-group
fabric
 neighbor swp2 interface peer-group
fabric
 !
 address-family ipv4 unicast
  network 4.4.4.4/32
 exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
 !
 address-family l2vpn evpn
  neighbor fabric activate
 exit-address-family
 !
```

Host 1 Configuration

```
ip route add 192.20.1.0/24 via
192.10.1.1
```

3.5.1 Fabric Troubleshooting

1. Verify all IP links are up and configured:

```
#ip link
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN mode DEFAULT
group default qlen 1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
10: swp1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc fq_codel state UP mode
DEFAULT group default qlen 1000
    link/ether 50:6b:4b:88:0f:19 brd ff:ff:ff:ff:ff:ff
11: swp2: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc fq_codel state UP mode
DEFAULT group default qlen 1000
    link/ether 50:6b:4b:88:0f:1d brd ff:ff:ff:ff:ff:ff
13: swp9: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc fq_codel master br0 state
UP mode DEFAULT group default qlen 1000  link/ether 50:6b:4b:88:0f:25 brd
ff:ff:ff:ff:ff:ff
20: vx10010: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue master br0
state UNKNOWN mode DEFAULT group default qlen 1000 link/ether 82:d4:17:dd:f5:e6 brd
ff:ff:ff:ff:ff:ff
21: vx4001: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue master br0
state UNKNOWN mode DEFAULT group default qlen 1000 link/ether 0e:f4:16:c8:41:05 brd
ff:ff:ff:ff:ff:ff
22: br0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue state UP mode
DEFAULT group default qlen 1000
    link/ether 50:6b:4b:88:0f:19 brd ff:ff:ff:ff:ff:ff
23: vrf1: <NOARP,MASTER,UP,LOWER_UP> mtu 65536 qdisc noqueue state UP mode DEFAULT
group default qlen 1000
    link/ether 7a:6c:31:41:2a:7a brd ff:ff:ff:ff:ff:ff
24: vlan10@br0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue master vrf1
state UP mode DEFAULT group default qlen 1000 link/ether 50:6b:4b:88:0f:19 brd
ff:ff:ff:ff:ff:ff
25: vlan10-v@vlan10: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue master
vrf1 state UP mode DEFAULT group default qlen 1000 link/ether 00:00:5e:00:00:10 brd
ff:ff:ff:ff:ff:ff
26: vlan4001@br0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue master
vrf1 state UP mode DEFAULT group default qlen 1000 link/ether 50:6b:4b:88:0f:19 brd
ff:ff:ff:ff:ff:ff
```

2. Verify the bridge VLAN setting:

```
#bridge vlan show
port    vlan ids
swp9
         10 PVID untagged
vx10010
         10 PVID untagged
vx4001
         4001 PVID untagged
br0
         10      4001
```

3. Verify BGP neighbors are up:

```
#vtysh -c 'show ip bgp summary'

IPv4 Unicast Summary:
BGP router identifier 1.1.1.1, local AS number 65011 vrf-id 0
BGP table version 9
RIB entries 13, using 2080 bytes of memory
Peers 2, using 41 KiB of memory
Peer groups 1, using 64 bytes of memory

Neighbor        V         AS MsgRcvd MsgSent   TblVer  InQ OutQ  Up/Down State/PfxRcd
swp1            4      65020     228     233        0    0    0 03:39:23            4
swp2            4      65020     229     233        0    0    0 03:39:24            4
```

```
Total number of neighbors 2
```

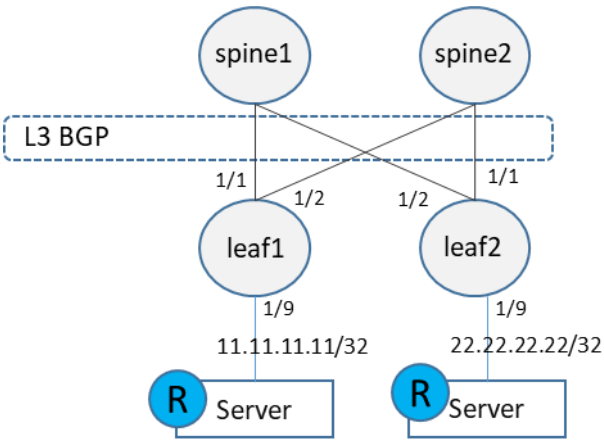
4. Verify routing table:

```
#ip route
2.2.2.2 proto bgp metric 20
   nexthop via 169.254.0.1 dev swp1 weight 1 onlink offload
   nexthop via 169.254.0.1 dev swp2 weight 1 onlink offload
3.3.3.3 via 169.254.0.1 dev swp1 proto bgp metric 20 onlink offload
4.4.4.4 via 169.254.0.1 dev swp2 proto bgp metric 20 onlink offload

# ip route list vrf vrf1
unreachable default metric 4278198272
172.16.0.0/16 via 2.2.2.2 dev vlan4001 proto bgp metric 20 onlink offload
192.10.1.0/24 dev vlan10 proto kernel scope link src 192.10.1.11 offload
192.10.1.0/24 dev vlan10-v proto kernel scope link src 192.10.1.1 metric 1024
```

5. Verify the EVPN table

```
#vtysh -c 'show bgp l2vpn evpn'
Route Distinguisher: ip 1.1.1.1:2
*> [2]:[0]:[48]:[00:02:c9:1c:69:30]
                    1.1.1.1                           32768 i
*> [2]:[0]:[48]:[00:02:c9:1c:69:30]:[32]:[192.10.1.2]
                    1.1.1.1                           32768 i
*> [3]:[0]:[32]:[1.1.1.1]
                    1.1.1.1                           32768 i
Route Distinguisher: ip 2.2.2.2:2
*  [5]:[0]:[16]:[172.16.0.0]
                    2.2.2.2                               0 65020 65012 ?
*> [5]:[0]:[16]:[172.16.0.0]
                    2.2.2.2                               0 65020 65012 ?
Displayed 5 out of 5 total prefixes
```

3.6 Routing from the Host



Leaf 1 Configuration

1. Configure the physical links:

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip link set dev swp9 up mtu 9216
```

2. Configure the loopback interface:

```
ip addr add 1.1.1.1/32 dev lo
```

3. Under FRR, configure unnumbered BGP neighbors and advertise the loopback interface:

```
router bgp 65011
 bgp router-id 1.1.1.1
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal Fabric Network
 neighbor fabric capability extended-nexthop
 neighbor swp1 interface peer-group fabric
 neighbor swp2 interface peer-group fabric
 neighbor swp9 interface peer-group fabric
 !
 address-family ipv4 unicast
  network 1.1.1.1/32
exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
```

Leaf 2 Configuration

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip link set dev swp9 up mtu 9216
ip addr add 2.2.2.2/32 dev lo
```

FRR configuration

```
router bgp 65012
 bgp router-id 2.2.2.2
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal Fabric Network
 neighbor fabric capability extended-nexthop
 neighbor swp1 interface peer-group fabric
 neighbor swp2 interface peer-group fabric
 neighbor swp9 interface peer-group fabric
 !
 address-family ipv4 unicast
  network 2.2.2.2/32
exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
```

Spine 1 Configuration

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip addr add 3.3.3.3/32 dev lo
```

FRR configuration

```
router bgp 65020
 bgp router-id 3.3.3.3
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal
Fabric Network
 neighbor fabric capability extended-
nexthop
 neighbor swp1 interface peer-group
fabric
 neighbor swp2 interface peer-group
fabric
 !
 address-family ipv4 unicast
  network 3.3.3.3/32
 exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
 !
```

Spine 2 Configuration

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip addr add 4.4.4.4/32 dev lo
```

FRR configuration

```
router bgp 65020
 bgp router-id 4.4.4.4
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal
Fabric Network
 neighbor fabric capability extended-
nexthop
 neighbor swp1 interface peer-group
fabric
 neighbor swp2 interface peer-group
fabric
 !
 address-family ipv4 unicast
  network 4.4.4.4/32
 exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
 !
```

Host 1 Configuration

```
ip addr add 11.11.11.11/32 dev lo
```

FRR configuration

```
router bgp 65001
 bgp router-id 11.11.11.11
 bgp bestpath as-path multipath-relax
 neighbor enp5s0 interface remote-as
external
 !
 address-family ipv4 unicast
  network 11.11.11.11/32
 exit-address-family
 !
 address-family ipv6 unicast
  neighbor enp5s0 activate
 exit-address-family
!
```

Host 2 Configuration

```
ip addr add 22.22.22.22/32 dev lo
```

FRR configuration

```
router bgp 65002
 bgp router-id 22.22.22.22
 bgp bestpath as-path multipath-relax
 neighbor enp5s0 interface remote-as
external
 !
 address-family ipv4 unicast
  network 22.22.22.22/32
 exit-address-family
 !
 address-family ipv6 unicast
  neighbor enp5s0 activate
 exit-address-family
!
```

3.6.1 Fabric Troubleshooting

1. Verify all IP links are up and configured

```
#ip link
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN mode DEFAULT
group default qlen 1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
14: swp1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc fq_codel state UP mode
DEFAULT group default qlen 1000
    link/ether 50:6b:4b:88:0f:19 brd ff:ff:ff:ff:ff:ff
15: swp2: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc fq_codel state UP mode
DEFAULT group default qlen 1000
    link/ether 50:6b:4b:88:0f:1d brd ff:ff:ff:ff:ff:ff
17: swp9: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc fq_codel master br0 state
UP mode DEFAULT group default qlen 1000  link/ether 50:6b:4b:88:0f:25 brd
ff:ff:ff:ff:ff:ff
```

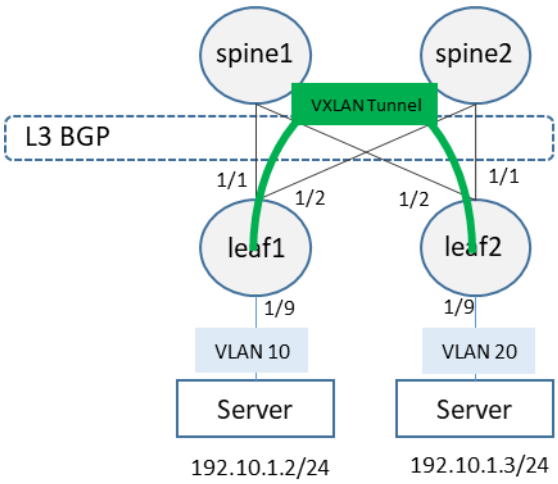
2. Verify BGP neighbors are up:

```
#vtysh -c 'show ip bgp summary'

IPv4 Unicast Summary:
BGP router identifier 1.1.1.1, local AS number 65011 vrf-id 0
BGP table version 15
RIB entries 11, using 1760 bytes of memory
Peers 3, using 62 KiB of memory
Peer groups 1, using 64 bytes of memory

Neighbor        V         AS MsgRcvd MsgSent   TblVer  InQ OutQ  Up/Down State/PfxRcd
swp1            4      65020      31      35        0    0    0 00:18:34            3
swp2            4      65020      31      35        0    0    0 00:18:33            3
swp9            4      65001     381     382        0    0    0 00:18:27            1

Total number of neighbors 3
```

3. Verify routing table:

```
#ip route
2.2.2.2 proto bgp metric 20
        nexthop via 169.254.0.1 dev swp2 weight 1 onlink offload
        nexthop via 169.254.0.1 dev swp1 weight 1 onlink offload
3.3.3.3 via 169.254.0.1 dev swp1 proto bgp metric 20 onlink offload
4.4.4.4 via 169.254.0.1 dev swp2 proto bgp metric 20 onlink offload
11.11.11.11 via 169.254.0.1 dev swp9 proto bgp metric 20 onlink offload
22.22.22.22 proto bgp metric 20
        nexthop via 169.254.0.1 dev swp1 weight 1 onlink offload
        nexthop via 169.254.0.1 dev swp2 weight 1 onlink offload
```

3.7 RoCE over VxLAN (PFC/ECN)



Leaf 1 Configuration

1. Configure the physical links

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip link set dev swp9 up mtu 9216
```

2. Configure the loopback interface

```
ip addr add 1.1.1.1/32 dev lo
```

3. Create VxLAN interface for VLAN 10 and for VLAN 20

```
ip link add name vx10010 up mtu 9216 type vxlan id 10 local 1.1.1.1 dstport 4789
nolearning noudpcsum tos inherit ttl 100
ip link add name vx10020 up mtu 9216 type vxlan id 20 local 1.1.1.1 dstport 4789
nolearning noudpcsum tos inherit ttl 100
```

4. Configure a vlan-aware bridge and set its MAC address to be the same as one of the switch ports

```
ip link add name br0 type bridge vlan_filtering 1 vlan_default_pvid 0 mcast_snooping
0
ip link set dev br0 address <swp1's MAC address>
```

5. Assign the host facing port and VxLAN interfaces as members of the bridge

```
ip link set dev swp9 master br0
ip link set dev vx10010 master br0
ip link set dev vx10020 master br0
```

6. Add VLAN 10 and 20 as access VLAN of host facing port and VxLAN interface

```
bridge vlan add vid 10 dev swp9 untagged pvid
bridge vlan add vid 10 dev vx10010 untagged pvid
bridge vlan add vid 20 dev vx10020 untagged pvid
ip link set dev br0 up
```

7. Configure SVI for VLAN 10 and 20 with anycast gw address

```
ip link add link br0 name vlan10 up mtu 9216 type vlan id 10
ip address add 192.10.1.11/24 dev vlan10
```

```
ip link add link vlan10 name vlan10-v up mtu 9216 address 00:00:5e:00:00:10 type
macvlan mode private
ip address add 192.10.1.1/24 dev vlan10-v metric 1024
ip link add link br0 name vlan20 up mtu 9216 type vlan id 20
ip address add 192.20.1.11/24 dev vlan20
ip link add link vlan20 name vlan20-v up mtu 9216 address 00:00:5e:00:00:20 type
macvlan mode private
ip address add 192.20.1.1/24 dev vlan20-v metric 1024
bridge vlan add vid 10 dev br0 self
bridge vlan add vid 20 dev br0 self
```

8. Add the anycast MAC address to the bridge

```
bridge fdb add 00:00:5e:00:00:10 dev br0 self local vlan 10
bridge fdb add 00:00:5e:00:00:20 dev br0 self local vlan 20
```

9. Disable rp_filter and enable arp_ignore to make sure ARPs for the anycast IP are answered with the anycast MAC

```
sysctl -w net.ipv4.conf.all.rp_filter=0
sysctl -w net.ipv4.conf.vlan10-v.rp_filter=0
sysctl -w net.ipv4.conf.vlan20-v.rp_filter=0
sysctl -w net.ipv4.conf.all.arp_ignore=1
```

10. Configure arp suppression

```
ip link set dev vx10010 type bridge_slave neigh_suppress on
ip link set dev vx10020 type bridge_slave neigh_suppress on
```

11. Configure lossy ingress and egress buffer pools

```
devlink sb pool set pci/0000:01:00.0 pool 0 size 6000000 thtype dynamic
devlink sb pool set pci/0000:01:00.0 pool 4 size 6000000 thtype dynamic
```

12. Bind the lossy ingress and egress pools to the relevant interfaces

```
devlink sb port pool set swp1 pool 0 th 13
devlink sb port pool set swp2 pool 0 th 13
devlink sb port pool set swp9 pool 0 th 13
devlink sb tc bind set swp1 tc 0 type ingress pool 0 th 13
devlink sb tc bind set swp2 tc 0 type ingress pool 0 th 13
devlink sb tc bind set swp9 tc 0 type ingress pool 0 th 13
devlink sb port pool set swp1 pool 4 th 13
devlink sb port pool set swp2 pool 4 th 13
devlink sb port pool set swp9 pool 4 th 13
devlink sb tc bind set swp1 tc 0 type egress pool 4 th 13
devlink sb tc bind set swp2 tc 0 type egress pool 4 th 13
devlink sb tc bind set swp9 tc 0 type egress pool 4 th 13
```

13. Configure ingress and egress buffer pools for lossless traffic (RoCE)

```
devlink sb pool set pci/0000:01:00.0 pool 1 size 6000000 thtype dynamic
devlink sb pool set pci/0000:01:00.0 pool 5 size 15990500 thtype dynamic
```

14. Bind the lossy ingress and egress pools, PG and TC to the relevant interfaces

```
devlink sb port pool set swp1 pool 1 th 10
devlink sb port pool set swp2 pool 1 th 10
devlink sb port pool set swp9 pool 1 th 10
devlink sb tc bind set swp1 tc 3 type ingress pool 1 th 10
devlink sb tc bind set swp2 tc 3 type ingress pool 1 th 10
devlink sb tc bind set swp9 tc 3 type ingress pool 1 th 10
devlink sb port pool set swp1 pool 5 th 16
devlink sb port pool set swp2 pool 5 th 16
devlink sb port pool set swp9 pool 5 th 16
```

```
devlink sb tc bind set swp1 tc 3 type egress pool 5 th 16
devlink sb tc bind set swp2 tc 3 type egress pool 5 th 16
devlink sb tc bind set swp9 tc 3 type egress pool 5 th 16
```

15. Classify dscp 26 (RoCE traffic) into switch priority 3 and dscp 48 (RoCE CNP) into switch priority 6

```
systemctl start lldpad
lldptool -T -i swp1 -V APP app=3,5,26
lldptool -T -i swp2 -V APP app=3,5,26
lldptool -T -i swp9 -V APP app=3,5,26
lldptool -T -i swp1 -V APP app=6,5,48
lldptool -T -i swp2 -V APP app=6,5,48
lldptool -T -i swp9 -V APP app=6,5,48
```

16. Disable priority update after routing

```
sysctl -w net.ipv4.ip_forward_update_priority=0
```

17. Map each switch priority to different traffic class

```
lldptool -T -i swp1 -V ETS-CFG up2tc=0:0,1:1,2:2,3:3,4:4,5:5,6:6,7:7
lldptool -T -i swp2 -V ETS-CFG up2tc=0:0,1:1,2:2,3:3,4:4,5:5,6:6,7:7
lldptool -T -i swp9 -V ETS-CFG up2tc=0:0,1:1,2:2,3:3,4:4,5:5,6:6,7:7
```

18. Configure WRR and strict priority schedulers

```
lldptool -T -i swp1 -V ETS-CFG
tsa=0:ets,1:ets,2:ets,3:ets,4:ets,5:ets,6:strict,7:ets tcbw=12,12,12,25,13,13,0,13
lldptool -T -i swp2 -V ETS-CFG
tsa=0:ets,1:ets,2:ets,3:ets,4:ets,5:ets,6:strict,7:ets tcbw=12,12,12,25,13,13,0,13
lldptool -T -i swp9 -V ETS-CFG
tsa=0:ets,1:ets,2:ets,3:ets,4:ets,5:ets,6:strict,7:ets tcbw=12,12,12,25,13,13,0,13
```

19. Enable PFC for priority 3

```
lldptool -T -i swp1 -V PFC enabled=3
lldptool -T -i swp2 -V PFC enabled=3
lldptool -T -i swp3 -V PFC enabled=3
lldptool -T -i swp1 -V PFC delay=32768
lldptool -T -i swp2 -V PFC delay=32768
lldptool -T -i swp9 -V PFC delay=32768
```

20. Under FRR, configure unnumbered BGP neighbors, advertise the loopback and all EVPN VNI

```
router bgp 65011
 bgp router-id 1.1.1.1
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal Fabric Network
 neighbor fabric capability extended-nexthop
 neighbor swp1 interface peer-group fabric
 neighbor swp2 interface peer-group fabric
 !
 address-family ipv4 unicast
  network 1.1.1.1/32
exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
 !
 address-family l2vpn evpn
  neighbor fabric activate
  advertise-all-vni
```

```
 exit-address-family
```

Leaf 2 Configuration

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip link set dev swp9 up mtu 9216
ip addr add 2.2.2.2/32 dev lo
ip link add name vx10010 up mtu 9216 type vxlan id 10 local 2.2.2.2 dstport 4789
nolearning noudpcsum tos inherit ttl 100
ip link add name vx10020 up mtu 9216 type vxlan id 20 local 2.2.2.2 dstport 4789
nolearning noudpcsum tos inherit ttl 100
ip link add name br0 type bridge vlan_filtering 1 vlan_default_pvid 0 mcast_snooping
0
ip link set dev br0 address <swp1's MAC address>
ip link set dev swp9 master br0
ip link set dev vx10010 master br0
ip link set dev vx10020 master br0
bridge vlan add vid 20 dev swp9 untagged pvid
bridge vlan add vid 10 dev vx10010 untagged pvid
bridge vlan add vid 20 dev vx10020 untagged pvid
ip link set dev br0 up
ip link add link br0 name vlan10 up mtu 9216 type vlan id 10
ip address add 192.10.1.12/24 dev vlan10
ip link add link vlan10 name vlan10-v up mtu 9216 address 00:00:5e:00:00:10 type
macvlan mode private
ip address add 192.10.1.1/24 dev vlan10-v metric 1024
ip link add link br0 name vlan20 up mtu 9216 type vlan id 20
ip address add 192.20.1.12/24 dev vlan20
ip link add link vlan20 name vlan20-v up mtu 9216 address 00:00:5e:00:00:20 type
macvlan mode private
ip address add 192.20.1.1/24 dev vlan20-v metric 1024
bridge vlan add vid 10 dev br0 self
bridge vlan add vid 20 dev br0 self
bridge fdb add 00:00:5e:00:00:10 dev br0 self local vlan 10
bridge fdb add 00:00:5e:00:00:20 dev br0 self local vlan 20
sysctl -w net.ipv4.conf.all.rp_filter=0
sysctl -w net.ipv4.conf.vlan10-v.rp_filter=0
sysctl -w net.ipv4.conf.vlan20-v.rp_filter=0
sysctl -w net.ipv4.conf.all.arp_ignore=1
ip link set dev vx10010 type bridge_slave neigh_suppress on
ip link set dev vx10020 type bridge_slave neigh_suppress on
devlink sb pool set pci/0000:01:00.0 pool 0 size 6000000 thtype dynamic
devlink sb pool set pci/0000:01:00.0 pool 4 size 6000000 thtype dynamic
devlink sb pool set pci/0000:01:00.0 pool 1 size 6000000 thtype dynamic
devlink sb pool set pci/0000:01:00.0 pool 5 size 15990500 thtype dynamic
devlink sb port pool set swp1 pool 0 th 13
devlink sb port pool set swp2 pool 0 th 13
devlink sb port pool set swp9 pool 0 th 13
devlink sb tc bind set swp1 tc 0 type ingress pool 0 th 13
devlink sb tc bind set swp2 tc 0 type ingress pool 0 th 13
devlink sb tc bind set swp9 tc 0 type ingress pool 0 th 13
devlink sb port pool set swp1 pool 4 th 13
devlink sb port pool set swp2 pool 4 th 13
devlink sb port pool set swp9 pool 4 th 13
devlink sb tc bind set swp1 tc 0 type egress pool 4 th 13
devlink sb tc bind set swp2 tc 0 type egress pool 4 th 13
devlink sb tc bind set swp9 tc 0 type egress pool 4 th 13
devlink sb port pool set swp1 pool 1 th 10
devlink sb port pool set swp2 pool 1 th 10
devlink sb port pool set swp9 pool 1 th 10
devlink sb tc bind set swp1 tc 3 type ingress pool 1 th 10
devlink sb tc bind set swp2 tc 3 type ingress pool 1 th 10
devlink sb tc bind set swp9 tc 3 type ingress pool 1 th 10
devlink sb port pool set swp1 pool 5 th 16
devlink sb port pool set swp2 pool 5 th 16
```

```
devlink sb port pool set swp9 pool 5 th 16
devlink sb tc bind set swp1 tc 3 type egress pool 5 th 16
devlink sb tc bind set swp2 tc 3 type egress pool 5 th 16
devlink sb tc bind set swp9 tc 3 type egress pool 5 th 16
systemctl start lldpad
lldptool -T -i swp1 -V APP app=3,5,26
lldptool -T -i swp2 -V APP app=3,5,26
lldptool -T -i swp9 -V APP app=3,5,26
lldptool -T -i swp1 -V APP app=6,5,48
lldptool -T -i swp2 -V APP app=6,5,48
lldptool -T -i swp9 -V APP app=6,5,48
sysctl -w net.ipv4.ip_forward_update_priority=0
lldptool -T -i swp1 -V ETS-CFG up2tc=0:0,1:1,2:2,3:3,4:4,5:5,6:6,7:7
lldptool -T -i swp2 -V ETS-CFG up2tc=0:0,1:1,2:2,3:3,4:4,5:5,6:6,7:7
lldptool -T -i swp9 -V ETS-CFG up2tc=0:0,1:1,2:2,3:3,4:4,5:5,6:6,7:7
lldptool -T -i swp1 -V ETS-CFG
tsa=0:ets,1:ets,2:ets,3:ets,4:ets,5:ets,6:strict,7:ets tcbw=12,12,12,25,13,13,0,13
lldptool -T -i swp2 -V ETS-CFG
tsa=0:ets,1:ets,2:ets,3:ets,4:ets,5:ets,6:strict,7:ets tcbw=12,12,12,25,13,13,0,13
lldptool -T -i swp9 -V ETS-CFG
tsa=0:ets,1:ets,2:ets,3:ets,4:ets,5:ets,6:strict,7:ets tcbw=12,12,12,25,13,13,0,13
lldptool -T -i swp1 -V PFC enabled=3
lldptool -T -i swp2 -V PFC enabled=3
lldptool -T -i swp3 -V PFC enabled=3
lldptool -T -i swp1 -V PFC delay=32768
lldptool -T -i swp2 -V PFC delay=32768
lldptool -T -i swp9 -V PFC delay=32768
```

FRR Configuration

```
router bgp 65012
 bgp router-id 2.2.2.2
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal Fabric Network
 neighbor fabric capability extended-nexthop
 neighbor swp1 interface peer-group fabric
 neighbor swp2 interface peer-group fabric
 !
 address-family ipv4 unicast
  network 2.2.2.2/32
exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
 !
 address-family l2vpn evpn
  neighbor fabric activate
  advertise-all-vni
 exit-address-family
```

Spine 1 Configuration

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip addr add 3.3.3.3/32 dev lo
devlink sb pool set pci/0000:01:00.0 pool 0 size 6000000 thtype dynamic
devlink sb pool set pci/0000:01:00.0 pool 4 size 6000000 thtype dynamic
devlink sb pool set pci/0000:01:00.0 pool 1 size 6000000 thtype dynamic
devlink sb pool set pci/0000:01:00.0 pool 5 size 15990500 thtype dynamic
devlink sb port pool set swp1 pool 0 th 13
devlink sb port pool set swp2 pool 0 th 13
devlink sb tc bind set swp1 tc 0 type ingress pool 0 th 13
devlink sb tc bind set swp2 tc 0 type ingress pool 0 th 13
devlink sb port pool set swp1 pool 4 th 13
devlink sb port pool set swp2 pool 4 th 13
devlink sb tc bind set swp1 tc 0 type egress pool 4 th 13
devlink sb tc bind set swp2 tc 0 type egress pool 4 th 13
devlink sb port pool set swp1 pool 1 th 10
devlink sb port pool set swp2 pool 1 th 10
devlink sb tc bind set swp1 tc 3 type ingress pool 1 th 10
devlink sb tc bind set swp2 tc 3 type ingress pool 1 th 10
devlink sb port pool set swp1 pool 5 th 16
devlink sb port pool set swp2 pool 5 th 16
devlink sb tc bind set swp1 tc 3 type egress pool 5 th 16
devlink sb tc bind set swp2 tc 3 type egress pool 5 th 16
systemctl start lldpad
lldptool -T -i swp1 -V APP app=3,5,26
lldptool -T -i swp2 -V APP app=3,5,26
lldptool -T -i swp1 -V APP app=6,5,48
lldptool -T -i swp2 -V APP app=6,5,48
sysctl -w net.ipv4.ip_forward_update_priority=0
lldptool -T -i swp1 -V ETS-CFG up2tc=0:0,1:1,2:2,3:3,4:4,5:5,6:6,7:7
lldptool -T -i swp2 -V ETS-CFG up2tc=0:0,1:1,2:2,3:3,4:4,5:5,6:6,7:7
lldptool -T -i swp1 -V ETS-CFG
tsa=0:ets,1:ets,2:ets,3:ets,4:ets,5:ets,6:strict,7:ets tcbw=12,12,12,25,13,13,0,13
lldptool -T -i swp2 -V ETS-CFG
tsa=0:ets,1:ets,2:ets,3:ets,4:ets,5:ets,6:strict,7:ets tcbw=12,12,12,25,13,13,0,13
lldptool -T -i swp1 -V PFC enabled=3
lldptool -T -i swp2 -V PFC enabled=3
lldptool -T -i swp1 -V PFC delay=32768
lldptool -T -i swp2 -V PFC delay=32768
```

FRR configuration

```
router bgp 65020
 bgp router-id 3.3.3.3
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal Fabric Network
 neighbor fabric capability extended nexthop
 neighbor swp1 interface peer-group fabric
 neighbor swp2 interface peer-group fabric
 !
 address-family ipv4 unicast
  network 3.3.3.3/32
 exit-address-family
 !
 address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
!
```

Spine 2 Configuration

```
ip link set dev swp1 up mtu 9216
ip link set dev swp2 up mtu 9216
ip addr add 4.4.4.4/32 dev lo
devlink sb pool set pci/0000:01:00.0 pool 0 size 6000000 thtype dynamic
devlink sb pool set pci/0000:01:00.0 pool 4 size 6000000 thtype dynamic
devlink sb pool set pci/0000:01:00.0 pool 1 size 6000000 thtype dynamic
devlink sb pool set pci/0000:01:00.0 pool 5 size 15990500 thtype dynamic
devlink sb port pool set swp1 pool 0 th 13
devlink sb port pool set swp2 pool 0 th 13
devlink sb tc bind set swp1 tc 0 type ingress pool 0 th 13
devlink sb tc bind set swp2 tc 0 type ingress pool 0 th 13
devlink sb port pool set swp1 pool 4 th 13
devlink sb port pool set swp2 pool 4 th 13
devlink sb tc bind set swp1 tc 0 type egress pool 4 th 13
devlink sb tc bind set swp2 tc 0 type egress pool 4 th 13
devlink sb port pool set swp1 pool 1 th 10
devlink sb port pool set swp2 pool 1 th 10
devlink sb tc bind set swp1 tc 3 type ingress pool 1 th 10
devlink sb tc bind set swp2 tc 3 type ingress pool 1 th 10
devlink sb port pool set swp1 pool 5 th 16
devlink sb port pool set swp2 pool 5 th 16
devlink sb tc bind set swp1 tc 3 type egress pool 5 th 16
devlink sb tc bind set swp2 tc 3 type egress pool 5 th 16
systemctl start lldpad
lldptool -T -i swp1 -V APP app=3,5,26
lldptool -T -i swp2 -V APP app=3,5,26
lldptool -T -i swp1 -V APP app=6,5,48
lldptool -T -i swp2 -V APP app=6,5,48
sysctl -w net.ipv4.ip_forward_update_priority=0
lldptool -T -i swp1 -V ETS-CFG up2tc=0:0,1:1,2:2,3:3,4:4,5:5,6:6,7:7
lldptool -T -i swp2 -V ETS-CFG up2tc=0:0,1:1,2:2,3:3,4:4,5:5,6:6,7:7
lldptool -T -i swp1 -V ETS-CFG
tsa=0:ets,1:ets,2:ets,3:ets,4:ets,5:ets,6:strict,7:ets tcbw=12,12,12,25,13,13,0,13
lldptool -T -i swp2 -V ETS-CFG
tsa=0:ets,1:ets,2:ets,3:ets,4:ets,5:ets,6:strict,7:ets tcbw=12,12,12,25,13,13,0,13
lldptool -T -i swp1 -V PFC enabled=3
lldptool -T -i swp2 -V PFC enabled=3
lldptool -T -i swp1 -V PFC delay=32768
lldptool -T -i swp2 -V PFC delay=32768
```

FRR configuration

```
router bgp 65020
 bgp router-id 4.4.4.4
 bgp bestpath as-path multipath-relax
 neighbor fabric peer-group
 neighbor fabric remote-as external
 neighbor fabric description Internal Fabric Network
 neighbor fabric capability extended-nexthop
 neighbor swp1 interface peer-group fabric
 neighbor swp2 interface peer-group fabric
 !
address-family ipv4 unicast
  network 4.4.4.4/32
 exit-address-family
 !
address-family ipv6 unicast
  neighbor fabric activate
 exit-address-family
!
address-family l2vpn evpn
  neighbor fabric activate
 exit-address-family
```

```
!
address-family l2vpn evpn
  neighbor fabric activate
 exit-address-family
!
```

Host 1 Configuration

```
ip route add 192.20.1.0/24 via 192.10.1.1
mlnx_qos -i <interface> --trust dscp
echo 106 > /sys/class/infiniband/<mlx-device>/tc/1/traffic_class
cma_roce_tos -d <mlx-device> -t 106
mlnx_qos -i <interface> --pfc 0,0,0,1,0,0,0,0
```

Host 2 Configuration

```
ip route add 192.10.1.0/24 via 192.20.1.1
mlnx_qos -i <interface> --trust dscp
echo 106 > /sys/class/infiniband/<mlx-device>/tc/1/traffic_class
cma_roce_tos -d <mlx-device> -t 106
mlnx_qos -i <interface> --pfc 0,0,0,1,0,0,0,0
```

3.7.1 Fabric Troubleshooting

1. Verify all IP links are up and configured

```
#ip link
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN mode DEFAULT
group default qlen 1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
10: swp1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc fq_codel state UP mode
DEFAULT group default qlen 1000
    link/ether 50:6b:4b:88:0f:19 brd ff:ff:ff:ff:ff:ff
11: swp2: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc fq_codel state UP mode
DEFAULT group default qlen 1000
    link/ether 50:6b:4b:88:0f:1d brd ff:ff:ff:ff:ff:ff
13: swp9: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc fq_codel master br0 state
UP mode DEFAULT group default qlen 1000  link/ether 50:6b:4b:88:0f:25 brd
ff:ff:ff:ff:ff:ff
20: vx10010: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue master br0
state UNKNOWN mode DEFAULT group default qlen 1000 link/ether 3a:2a:68:09:4f:ad brd
ff:ff:ff:ff:ff:ff
21: vx10020: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue master br0
state UNKNOWN mode DEFAULT group default qlen 1000 link/ether 16:8a:3e:d5:0f:90 brd
ff:ff:ff:ff:ff:ff
22: br0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue state UP mode
DEFAULT group default qlen 1000
    link/ether 50:6b:4b:88:0f:19 brd ff:ff:ff:ff:ff:ff
23: vlan10@br0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue state UP
mode DEFAULT group default qlen 1000 link/ether 50:6b:4b:88:0f:19 brd
ff:ff:ff:ff:ff:ff
24: vlan10-v@vlan10: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue state
UP mode DEFAULT group default qlen 1000 link/ether 00:00:5e:00:00:10 brd
ff:ff:ff:ff:ff:ff
25: vlan20@br0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue state UP
mode DEFAULT group default qlen 1000 link/ether 50:6b:4b:88:0f:19 brd
ff:ff:ff:ff:ff:ff
26: vlan20-v@vlan20: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9216 qdisc noqueue state
UP mode DEFAULT group default qlen 1000 link/ether 00:00:5e:00:00:20 brd
ff:ff:ff:ff:ff:ff
```

2. Verify the bridge VLAN setting

```
#bridge vlan show
port     vlan ids
swp9
            10 PVID untagged
vx10010
            10 PVID untagged
vx10020
            20 PVID untagged
br0
            10      20
```

3. Verify bgp neighbors are up

```
#vtysh -c 'show ip bgp summary'

IPv4 Unicast Summary:
BGP router identifier 1.1.1.1, local AS number 65011 vrf-id 0
BGP table version 9
RIB entries 13, using 2080 bytes of memory
Peers 2, using 41 KiB of memory
Peer groups 1, using 64 bytes of memory

Neighbor        V         AS MsgRcvd MsgSent   TblVer  InQ OutQ  Up/Down State/PfxRcd
swp1            4      65020     228     233        0    0    0 03:39:23            4
swp2            4      65020     229     233        0    0    0 03:39:24            4

Total number of neighbors 2
```

4. Verify routing table:

```
#ip route
2.2.2.2 proto bgp metric 20
   nexthop via 169.254.0.1 dev swp1 weight 1 onlink offload
   nexthop via 169.254.0.1 dev swp2 weight 1 onlink offload
3.3.3.3 via 169.254.0.1 dev swp1 proto bgp metric 20 onlink offload
4.4.4.4 via 169.254.0.1 dev swp2 proto bgp metric 20 onlink offload
192.10.1.0/24 dev vlan10 proto kernel scope link src 192.10.1.11 offload
192.10.1.0/24 dev vlan10-v proto kernel scope link src 192.10.1.1 metric 1024
192.20.1.0/24 dev vlan20 proto kernel scope link src 192.20.1.11 offload
192.20.1.0/24 dev vlan20-v proto kernel scope link src 192.20.1.1 metric 1024
```

5. Verify the EVPN table

```
#vtysh -c 'show bgp l2vpn evpn'
Route Distinguisher: ip 1.1.1.1:2

*> [2]:[0]:[48]:[00:02:c9:1c:69:30]
                    1.1.1.1                            32768 i
*> [2]:[0]:[48]:[00:02:c9:1c:69:30]:[32]:[192.10.1.2]
                    1.1.1.1                            32768 i
*> [3]:[0]:[32]:[1.1.1.1]
                    1.1.1.1                            32768 i
Route Distinguisher: ip 1.1.1.1:3

*> [3]:[0]:[32]:[1.1.1.1]
                    1.1.1.1                            32768 i
Route Distinguisher: ip 2.2.2.2:2

*  [3]:[0]:[32]:[2.2.2.2]
                    2.2.2.2                                0 65020 65012 i
*> [3]:[0]:[32]:[2.2.2.2]
                    2.2.2.2                                0 65020 65012 i
Route Distinguisher: ip 2.2.2.2:3
```

```
*   [2]:[0]:[48]:[00:02:c9:1c:62:40]
                    2.2.2.2                                  0 65020 65012 i
*> [2]:[0]:[48]:[00:02:c9:1c:62:40]
                    2.2.2.2                                  0 65020 65012 i
*   [2]:[0]:[48]:[00:02:c9:1c:62:40]:[32]:[192.20.1.2]
                    2.2.2.2                                  0 65020 65012 i
*> [2]:[0]:[48]:[00:02:c9:1c:62:40]:[32]:[192.20.1.2]
                    2.2.2.2                                  0 65020 65012 i
*   [3]:[0]:[32]:[2.2.2.2]
                    2.2.2.2                                  0 65020 65012 i
*> [3]:[0]:[32]:[2.2.2.2]
                    2.2.2.2                                  0 65020 65012 i

Displayed 12 out of 12 total prefixes
```

6. Verify the MAC address table:

```
#bridge fdb show vlan 10
001:02:c9:1c:69:30 dev swp9 vlan 10 extern_learn offload br0
50:6b:4b:88:0f:25 dev swp9 vlan 10 br0 permanent
3a:2a:68:09:4f:ad dev vx10010 vlan 10 br0 permanent
50:6b:4b:88:0f:19 dev br0 vlan 10 br0 permanent

# bridge fdb show vlan 20
00:02:c9:1c:62:40 dev vx10020 vlan 20 extern_learn offload br0
16:8a:3e:d5:0f:90 dev vx10020 vlan 20 br0 permanent
50:6b:4b:88:0f:19 dev br0 vlan 20 br0 permanent
```

7. Verify buffer pools, PG and TC are configured and binded to ports:

```
# devlink sb pool show
pci/0000:01:00.0:
  sb 0 pool 0 type ingress size 6000000 thtype dynamic
  sb 0 pool 1 type ingress size 6000000 thtype dynamic
  sb 0 pool 2 type ingress size 0 thtype dynamic
  sb 0 pool 3 type ingress size 200064 thtype dynamic
  sb 0 pool 4 type egress size 6000000 thtype dynamic
  sb 0 pool 5 type egress size 15990528 thtype dynamic
  sb 0 pool 6 type egress size 0 thtype dynamic
  sb 0 pool 7 type egress size 0 thtype dynamic
  sb 0 pool 8 type egress size 15990816 thtype static

# devlink sb tc bind show swp9 tc 3 type ingress
swp9: sb 0 tc 3 type ingress pool 1 threshold 10

# devlink sb tc bind show swp9 tc 6 type ingress
swp9: sb 0 tc 6 type ingress pool 0 threshold 3

# devlink sb tc bind show swp9 tc 3 type egress
swp9: sb 0 tc 3 type egress pool 5 threshold 16

# devlink sb tc bind show swp9 tc 6 type egress
swp9: sb 0 tc 6 type egress pool 4 threshold 11
```

8. Verify dscp classification is configured:

```
#lldptool -t -i swp9 -V APP -c app
APP=(prio,sel,proto)
0:(3,5,26) local hw (set)
1:(6,5,48) local hw (set)
```

9. Verify Schedulers are configured:

```
#lldptool -t -i swp9 -V ETS-CFG -c
enableTx=no
willing=
```

```
numtcs=
up2tc=0:0,1:1,2:2,3:3,4:4,5:5,6:6,7:7
tcbw=12,12,12,25,13,13,0,13
tsa=0:ets,1:ets,2:ets,3:ets,4:ets,5:ets,6:strict,7:ets
```

10. Verify RoCE traffic is flowing on priority 3 and check PFC counters:

```
# ethtool -S swp1 |grep prio_3
     rx_octets_prio_3: 0
     rx_frames_prio_3: 0
     tx_octets_prio_3: 13114043008160
     tx_frames_prio_3: 3115714661
     rx_pause_prio_3: 2880193651
     rx_pause_duration_prio_3: 4586569657
     tx_pause_prio_3: 0
     tx_pause_duration_prio_3: 0
```

11. Check current/maximum buffer usage on each port, take a snapshot of current buffer usage:

```
# devlink sb occupancy snapshot pci/0000:01:00.0
```

Check current/maximum buffer usage in this snapshot:

```
# devlink sb occupancy show swp1
swp1:
  pool:     0:0/480              1:0/0       2:0/0        3:0/288          4:0/480
            5:3024384/3037056    6:0/0       7:0/0        8:0/0
  itc:   0(0):0/480           1(0):0/0    2(0):0/0     3(1):0/0
         4(0):0/0             5(0):0/0    6(0):0/480   7(0):0/0
  etc:   0(4):0/288           1(4):0/0    2(4):0/0     3(5):3024384/3037056
         4(4):0/0             5(4):0/0    6(4):0/0     7(4):0/0
         8(8):0/0             9(8):0/0   10(8):0/0    11(8):0/0
        12(8):0/0            13(8):0/0   14(8):0/0    15(8):0/0
```

maximum usage can be cleared with the following command:

```
#devlink sb occupancy clearmax pci/0000:01:00.0
```