Advanced VXLAN with vPC: Configuration and Verification of L2VNI and L3VNI

Contents

Introduction

This document describes how to set up a lab with Nexus 9Kv switches using Advanced Virtual eXtensible Local Area Network (VXLAN) with Virtual Port-Channel (vPC).

Prerequisites

Requirements

Cisco recommends that you have knowledge of these topics:

- Understanding of routing and switching, as well as Multiprotocol Label Switching (MPLS) technology
- Experience with multicast routing principles such as Rendezvous Point (RP) and Platform Independent Multicast (PIM)
- Understanding of Border Gateway Protocol (BGP) Address Family Indicator (AFI)/Subsequent Address Family Indicator (SAFI)

Components Used

This document is not restricted to specific software and hardware versions.

The information in this document was created from the devices in a specific lab environment. All of the devices used in this document started with a cleared (default) configuration. If your network is live, ensure that you understand the potential impact of any command.

Background Information

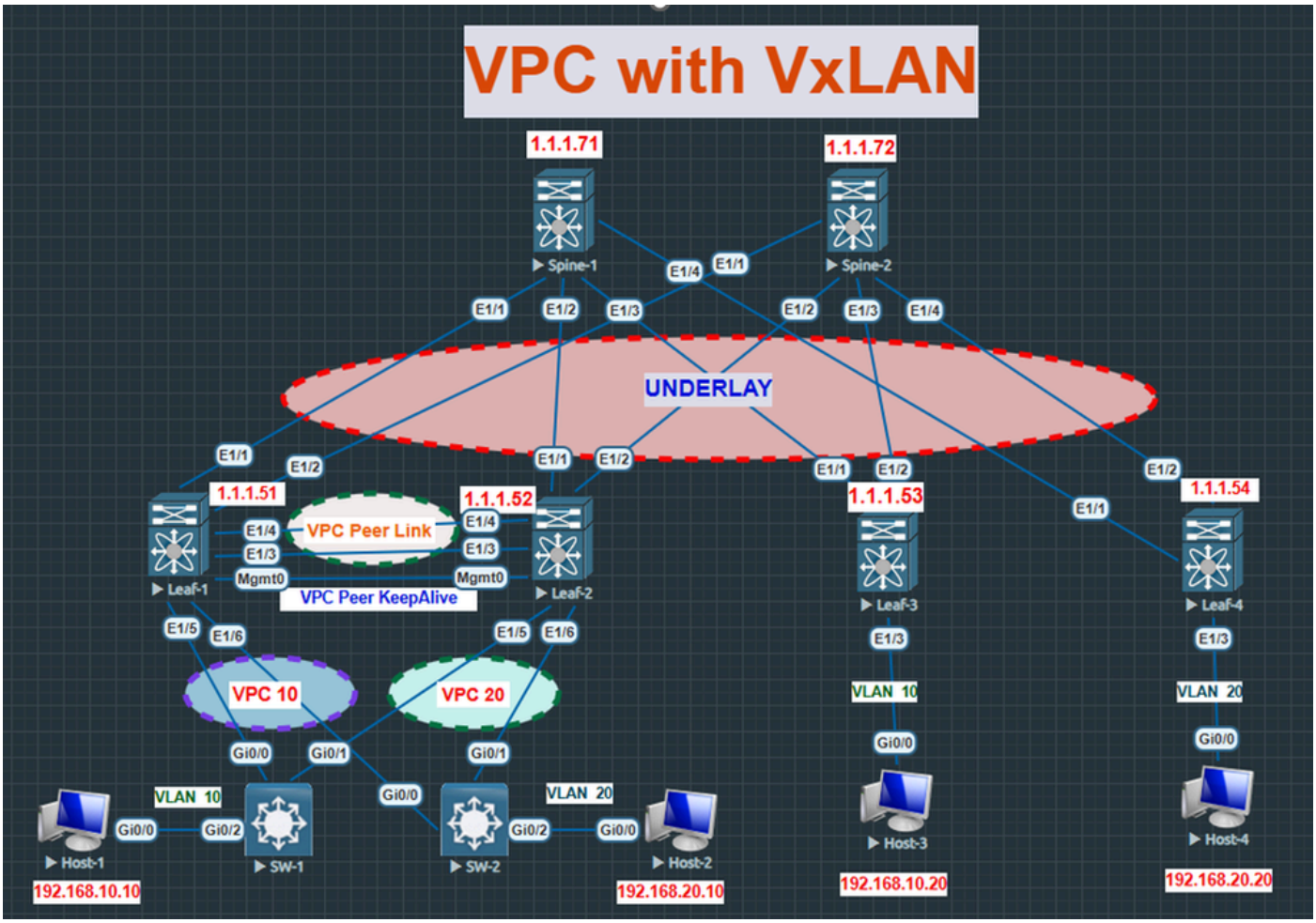
The document also provides guidance on deploying the lab, as well as verifying configurations and operations.

For this lab, the EveNg with Nexus 9000V switches is utilized for both the Leaf and Spine.

Virtual Tunnel Endpoint (VTEP)	LEAF1, LEAF2, LEAF3, LEAF4
vPC	LEAF1 and LEAF2
LEAF1 Primary and Secondary loopback IP	Loopback0 – 1.1.1.51, Loopback1 - 10.1.1.100
LEAF2 Primary and Secondary loopback IP	Loopback0 – 1.1.1.52, Loopback1 - 10.1.1.100
LEAF3 loopback IP	1.1.1.53
LEAF4 loopback IP	1.1.1.54
SPINE1 loopback and Anycast RP	Loopback0 - 1.1.1.71, Loopback1 - 10.1.2.10 (Anycast RP)
SPINE2 loopback and Anycast RP	Loopback0 - 1.1.1.72, Loopback1 - 10.1.2.10 (Anycast RP)
HOST 1	192.168.10.10 (0000. 0000.aaaa) (VLAN 10)
HOST 2	192.168.20.10 (0000. 0000.bbbb) (VLAN 20)
HOST 3	192.168.10.20 (0000. 0000.cccc) (VLAN 10)
HOST 4	192.168.20.20 (0000. 0000.dddd) (VLAN 20)
VLAN 10	L2VNI 100010
VLAN 20	L2VNI 100020
VLAN 500	L3VNI 50000

Configure

Network Diagram



Configurations

- Underlay and PIM neighborships are already established.

LEAF Switch:

```
feature ospf

router ospf UNDERLAY
  log-adjacency-changes

interface loopback0
  ip router ospf UNDERLAY area 0.0.0.0

interface Ethernet1/1
  ip ospf cost 4
  ip ospf network point-to-point
  ip router ospf UNDERLAY area 0.0.0.0

interface Ethernet1/2
  ip ospf cost 4
  ip ospf network point-to-point
  ip router ospf UNDERLAY area 0.0.0.0
```

Enabling Open Shortest Path First (OSPF) on Leaf Switch

```
feature pim

ip pim rp-address 10.1.2.10 group-list 224.0.0.0/4
ip pim ssm range 232.0.0.0/8

vrf context TENANT1
  ip pim ssm range 232.0.0.0/8

interface Vlan10
  ip pim sparse-mode

interface Vlan20
  ip pim sparse-mode

interface loopback0
  ip pim sparse-mode

interface Ethernet1/1
  ip pim sparse-mode

interface Ethernet1/2
  ip pim sparse-mode
```

Enabling PIM on Leaf Switch

```
LEAF-1# show ip ospf neighbors
 OSPF Process ID UNDERLAY VRF default
 Total number of neighbors: 2
 Neighbor ID     Pri State            Up Time  Address         Interface
 1.1.1.71          1 FULL/ -          04:32:03 192.168.11.1    Eth1/1
 1.1.1.72          1 FULL/ -          04:17:47 192.168.21.2    Eth1/2
LEAF-1# sh ip pim neighbor
PIM Neighbor Status for VRF "default"
Neighbor        Interface            Uptime    Expires   DR       Bidir-  BFD    ECMP Redirect
                                                         Priority Capable State     Capable
192.168.11.1    Ethernet1/1          04:32:14  00:01:30  1        yes     n/a     no
192.168.21.2    Ethernet1/2          04:17:58  00:01:44  1        yes     n/a     no
LEAF-1#
```

OSPF Neighbor

Spine Switch:

```
feature pim

ip pim rp-address 10.1.2.10 group-list 224.0.0.0/4
ip pim ssm range 232.0.0.0/8
ip pim anycast-rp 10.1.2.10 1.1.1.71
ip pim anycast-rp 10.1.2.10 1.1.1.72
```

Enabling PIM on Spine Switch

- Underlay and PIM neighborships are already established.
- Both Spine switches will be the identical Anycast RP for the whole multicast group 224.0.0.0/4.
- Maximum Transmission Unit (MTU) is set to 9000/9216 on the interfaces between the Leaf and Spine switches.

First, lets set up a vPC between Leaf1 and Leaf2.

Step 1. vPC feature and domain enablement.

- Enable feature vPC and Link Aggregation Control Protocol (LACP).
- Configure the vPC domain.
- The mgmt 0 interfaces are used as a peer keepalive link and Eth1/3 and Eth1/4 will be the part of vPC peer link (Port-Channel 1).
- Ensure that the peer-switch command is configured to share a common MAC address with descending switches.

```
feature lacp
feature vpc
```

Enabling Feature on Leaf Switch

```
LEAF-1# sh run vpc

!Command: show running-config vpc
!Running configuration last done at: Sat Dec 28 07:17:18 2024
!Time: Sat Dec 28 07:39:48 2024

version 7.0(3)I7(9) Bios:version
feature vpc

vpc domain 1
  peer-switch
  role priority 100
  peer-keepalive destination 192.168.0.52
  peer-gateway

interface port-channel1
  vpc peer-link
```

Enabling vPC on Leaf Switch 1

```
LEAF-2# sh run vpc

!Command: show running-config vpc
!Running configuration last done at: Sat Dec 28 07:17:14 2024
!Time: Sat Dec 28 07:40:20 2024

version 7.0(3)I7(9) Bios:version
feature vpc

vpc domain 1
  peer-switch
  role priority 200
  peer-keepalive destination 192.168.0.51
  peer-gateway

interface port-channel1
  vpc peer-link
```

Enabling vPC on Leaf Switch 2

Step 2. Port member assignment.

- Assign the port member to the channel group and include them in the vPC. In this case, two vPCs are being used. vPC 20 and vPC 10.

```
LEAF-1# sh run int port-channel 10, port-channel 20 membership

!Command: show running-config interface port-channel10, port-channel20 membership
!Running configuration last done at: Sat Dec 28 07:17:18 2024
!Time: Sat Dec 28 07:42:44 2024

version 7.0(3)I7(9) Bios:version

interface port-channel10
  switchport mode trunk
  vpc 10

interface Ethernet1/5

  switchport mode trunk
  channel-group 10 mode active

interface port-channel20
  switchport mode trunk
  vpc 20

interface Ethernet1/6

  switchport mode trunk
  channel-group 20 mode active

LEAF-1#
```

Assigning Port Channel on Leaf Switch 1

```
LEAF-2# sh run int port-channel 10, port-channel 20 membership

!Command: show running-config interface port-channel10, port-channel20 membership
!Running configuration last done at: Sat Dec 28 07:17:14 2024
!Time: Sat Dec 28 07:43:16 2024

version 7.0(3)I7(9) Bios:version

interface port-channel10
  switchport mode trunk
  vpc 10

interface Ethernet1/5

  switchport mode trunk
  channel-group 10 mode active

interface port-channel20
  switchport mode trunk
  vpc 20

interface Ethernet1/6

  switchport mode trunk
  channel-group 20 mode active

LEAF-2#
```

Assigning Port Channel on Leaf Switch 2

- Here, a vPC is created, and peers begin exchanging keepalive messages in order to verify availability.

```
LEAF-1# show 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                          : primary
Number of vPCs configured         : 2
Peer Gateway                      : Enabled
Dual-active excluded VLANs        : -
Graceful Consistency Check        : Enabled
Auto-recovery status              : Disabled
Delay-restore status              : Timer is off.(timeout = 30s)
Delay-restore SVI status          : Timer is off.(timeout = 10s)
Operational Layer3 Peer-router    : Disabled

vPC Peer-link status
---------------------------------------------------------------------
id    Port   Status Active vlans
--    ----   ------ --------------------------------------------------
1     Po1    up     1,10,20,500

vPC status
----------------------------------------------------------------------------
Id    Port          Status Consistency Reason                Active vlans
--    ------------  ------ ----------- ------                ---------------
10    Po10          up     success     success               1,10,20,500

20    Po20          up     success     success               1,10,20,500


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.

LEAF-1#
```

vPC Status on Leaf Switch 1

```
LEAF-2# 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
Number of vPCs configured         : 2
Peer Gateway                      : Enabled
Dual-active excluded VLANs        : -
Graceful Consistency Check        : Enabled
Auto-recovery status              : Disabled
Delay-restore status              : Timer is off.(timeout = 30s)
Delay-restore SVI status          : Timer is off.(timeout = 10s)
Operational Layer3 Peer-router    : Disabled

vPC Peer-link status
---------------------------------------------------------------------
id    Port   Status Active vlans
--    ----   ------ -----------------------------------------------
1     Po1    up     1,10,20,500

vPC status
----------------------------------------------------------------------------
Id    Port          Status Consistency Reason                Active vlans
--    ------------  ------ ----------- ------                ---------------
10    Po10          up     success     success               1,10,20,500

20    Po20          up     success     success               1,10,20,500


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.

LEAF-2#
```

vPC Status on Leaf Switch 2

- VLAN 10, 20, 500 is already configured and passed over the vPC member ports and vPC peer link.

Step 3. Configure the Secondary IP address.

- When vPC is included in the VXLAN fabric, both vPC VTEP peers start using virtual IP (VIP) addresses as source addresses instead of their physical IP addresses (PIP). This also means that when BGP Ethernet VPN (EVPN) advertises Route Types 2 (MAC/IP advertisement) and 5 (IP prefix-route) by default, VIP is used as a next-hop. The Loopback 0 interface in our example is set up with two IP addresses: 10.1.1.100/32 (VIP) as the secondary IP and 1.1.1.51/32 (PIP) as the primary IP.
- Here a common IP address is configured as a secondary one under the loopback 0 interface.

```
LEAF-1# sh run int lo

!Command: show running-config interface loopback0
!Running configuration last done at: Sat Dec 28 07:51:58 2024
!Time: Sat Dec 28 07:55:26 2024

version 7.0(3)I7(9) Bios:version

interface loopback0
  ip address 1.1.1.51/32
  ip address 10.1.1.100/32 secondary
  ip router ospf UNDERLAY area 0.0.0.0
  ip pim sparse-mode

LEAF-1#
```

Secondary IP on Leaf Switch 1

```
LEAF-2# sh run int lo

!Command: show running-config interface loopback0
!Running configuration last done at: Sat Dec 28 07:52:05 2024
!Time: Sat Dec 28 07:55:37 2024

version 7.0(3)I7(9) Bios:version

interface loopback0
  ip address 1.1.1.52/32
  ip address 10.1.1.100/32 secondary
  ip router ospf UNDERLAY area 0.0.0.0
  ip pim sparse-mode

LEAF-2#
```

Secondary IP on Leaf Switch 2

Step 4. Enable VXLAN and related features.

- Network Virtualization (nV) overlay - enables VXLAN
- Feature nV overlay EVPN- enables EVPN Control Plane
- Feature fabric forwarding - enables Host Mobility Manager
- Feature Virtual Network (VN)-segment-VLAN-based - enables VLAN-based VXLAN

```
LEAF-1# sh run | sec "feature|nv over"
nv overlay evpn
feature ospf
feature bgp
feature pim
feature fabric forwarding
feature interface-vlan
feature vn-segment-vlan-based
feature lacp
feature vpc
feature nv overlay
LEAF-1#
```

Features on Leaf Switch

```
SPINE-1# sh run | sec "feature|nv over"
nv overlay evpn
feature ospf
feature bgp
feature pim
feature nv overlay
SPINE-1#
```

Features on Spine Switch

- Since the spine does not require knowledge of the VLAN information of the client, the VN-segment and fabric features does not need to be enabled.

Step 5. Bring up the BGP neighborship.

- BGP between the Leaf and Spine switches must be enabled. The spine will serve as a route reflector in the lab.
- Although, it is optional to configure Route Reflector (RR), for the sake of scalability, Cisco recommends RR.

```
LEAF-1# sh run bgp

!Command: show running-config bgp
!Running configuration last done at: Sat Dec 28 07:51:58 2024
!Time: Sat Dec 28 08:07:35 2024

version 7.0(3)I7(9) Bios:version
feature bgp

router bgp 65000
  router-id 1.1.1.51
  neighbor 1.1.1.71
    remote-as 65000
    update-source loopback0
    address-family l2vpn evpn
      send-community extended
  neighbor 1.1.1.72
    remote-as 65000
    update-source loopback0
    address-family l2vpn evpn
      send-community extended
```

Enabling BGP on Leaf Switch

```
SPINE-1# sh run bgp

!Command: show running-config bgp
!Running configuration last done at: Sat Dec 28 07:16:33 2024
!Time: Sat Dec 28 08:08:21 2024

version 7.0(3)I7(9) Bios:version
feature bgp

router bgp 65000
  router-id 1.1.1.71
  neighbor 1.1.1.51
    remote-as 65000
    update-source loopback0
    address-family l2vpn evpn
      send-community extended
      route-reflector-client
  neighbor 1.1.1.52
    remote-as 65000
    update-source loopback0
    address-family l2vpn evpn
      send-community extended
      route-reflector-client
  neighbor 1.1.1.53
    remote-as 65000
    update-source loopback0
    address-family l2vpn evpn
      send-community extended
      route-reflector-client
  neighbor 1.1.1.54
    remote-as 65000
    update-source loopback0
    address-family l2vpn evpn
      send-community extended
      route-reflector-client


SPINE-1#
```

Enabling BGP on Spine Switch

```
LEAF-1# show bgp l2vpn evpn summary
BGP summary information for VRF default, address family L2VPN EVPN
BGP router identifier 1.1.1.51, local AS number 65000
BGP table version is 62, L2VPN EVPN config peers 2, capable peers 2
10 network entries and 13 paths using 2228 bytes of memory
BGP attribute entries [10/1600], BGP AS path entries [0/0]
BGP community entries [0/0], BGP clusterlist entries [4/16]

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
1.1.1.71        4 65000     146     121       62    0    0 01:45:52 3
1.1.1.72        4 65000     141     114       62    0    0 01:39:12 3
LEAF-1#
```

BGP Status on Leaf Switch

```
SPINE-1# show bgp l2vpn evpn summary
BGP summary information for VRF default, address family L2VPN EVPN
BGP router identifier 1.1.1.71, local AS number 65000
BGP table version is 98, L2VPN EVPN config peers 4, capable peers 4
9 network entries and 9 paths using 2124 bytes of memory
BGP attribute entries [7/1120], BGP AS path entries [0/0]
BGP community entries [0/0], BGP clusterlist entries [0/0]

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
1.1.1.51        4 65000     147     124       98    0    0 01:46:29 2
1.1.1.52        4 65000     147     124       98    0    0 01:46:30 2
1.1.1.53        4 65000     128     155       98    0    0 02:01:15 1
1.1.1.54        4 65000     191     225       98    0    0 03:03:08 2
SPINE-1#
```

BGP Status on Spine Switch

Step 6. Enable VRF context on leaf switches. VRF separates customer traffic and facilitates communication between two distinct L2VNIs via L3VNI.

- Allocate L3VNI 50000 under VRF TENANT1.

```
vrf context TENANT1
  vni 50000
  ip pim ssm range 232.0.0.0/8
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
```

L3VNI Allocation

Step 7. Network Virtual Interface (NVE), VXLAN identifier (VNI), and VLAN configuration.

- Set up the NVE Interface, using Loopback 0 as the source. Define the Multicast group for each VNI, where Layer 2 Broadcast, Unknown unicast, and Multicast (BUM) traffic will be delivered, then attach the VNI 100010 and 100020 IDs to the NVE interface. The VXLAN header contains the information that the VNI uses in order to identify which VXLAN segments it belongs to.
- The L3VNI 50000 is linked to the VRF instance (when sending it to the spine switch, VNI 50000 was attached in the VRF table).
- The **host-reachability protocol BGP** command activates the EVPN address family in the VXLAN tunnel, which means that MAC addresses and IP addresses are learned via the BGP protocol in the control plane and not in the data plane.
- Configure **suppress-arp** under the NVE interface.
- Attach Layer 2 and Layer 3 VLAN to relevant VNI.

Suppress-Address Resolution Protocol (ARP):

The Multi-Protocol (MP)-BGP EVPN control plane offers an improvement called ARP suppression in order to lessen network flooding brought on by broadcast traffic from ARP requests. Each of a VNIs VTEPs keeps an ARP suppression cache table for known IP hosts and the MAC addresses that correspond to them in the VNI segment when ARP suppression is enabled for that VNI. Its local VTEP intercepts the ARP request and

looks for the ARP-resolved IP address in its ARP suppression cache table whenever an end host in the VNI submits an ARP request for another end-host IP address. On behalf of the remote end host, the local VTEP sends an ARP response if it discovers a match. The ARP response then provides the local host with the remote hosts MAC address. The ARP request is flooded to the other VTEPs in the VNI if the local VTEP does not have the ARP-resolved IP address in its ARP suppression table. For the first ARP request to a silent network host, this ARP flooding can take place.

```
LEAF-1# sh run interface nve 1

!Command: show running-config interface nve1
!Running configuration last done at: Sat Dec 28 07:51:58 2024
!Time: Sat Dec 28 08:44:44 2024

version 7.0(3)I7(9) Bios:version

interface nve1
  no shutdown
  host-reachability protocol bgp
  source-interface loopback0
  member vni 50000 associate-vrf
  member vni 100010
    suppress-arp
    mcast-group 239.0.0.10
  member vni 100020
    suppress-arp
    mcast-group 239.0.0.20

LEAF-1#
```

NVE Interface

```
LEAF-1# sh run vlan

!Command: show running-config vlan
!Running configuration last done at: Sat Dec 28 07:51:58 2024
!Time: Sat Dec 28 08:46:44 2024

version 7.0(3)I7(9) Bios:version
vlan 1,10,20,500
vlan 10
  vn-segment 100010
vlan 20
  vn-segment 100020
vlan 500
  vn-segment 50000


LEAF-1#
```

VLAN to VN-Segment Mapping

- By sending Spine a PIM join message, the NVE interface will join the multicast groups 239.0.0.10 and 239.0.0.20, respectively, as soon as it boots up.
- You can see other (S, G) tables as well (1.1.1.54,239.0.0.20) and (10.1.1.100, 239.0.0.10/239.0.0.20)

in the image and those are already registered with Spine from different Leaf Switches.

```
LEAF-1# sh ip mroute summary
IP Multicast Routing Table for VRF "default"
Route Statistics unavailable - only liveness detected

Total number of routes: 7
Total number of (*,G) routes: 2
Total number of (S,G) routes: 4
Total number of (*,G-prefix) routes: 1
Group count: 2, rough average sources per group: 2.0

Group: 232.0.0.0/8, Source count: 0
Source          packets      bytes          aps   pps       bit-rate     oifs
(*,G)           0            0              0     0         0.000   bps  0

Group: 239.0.0.10/32, Source count: 2
Source          packets      bytes          aps   pps       bit-rate     oifs
(*,G)           1            100            100   0         0.000   bps  1
1.1.1.53        48           4644           96    0         78.267  bps  1
10.1.1.100      1124         113514         100   0         131.467 bps  1

Group: 239.0.0.20/32, Source count: 2
Source          packets      bytes          aps   pps       bit-rate     oifs
(*,G)           1            100            100   0         0.000   bps  1
1.1.1.54        51           4944           96    0         63.200  bps  1
10.1.1.100      1116         112729         101   0         70.667  bps  1
LEAF-1#
```

Mroute Table

Step 8. Enable EVPN instance.

- Enable EVPN instance along with address-family for EVPN and VRF under BGP.

```
LEAF-1# sh run bgp

!Command: show running-config bgp
!Running configuration last done at: Sat Dec 28 09:22:19 2024
!Time: Sat Dec 28 09:43:07 2024

version 7.0(3)I7(9) Bios:version
feature bgp

router bgp 65000
  router-id 1.1.1.51
  neighbor 1.1.1.71
    remote-as 65000
    update-source loopback0
    address-family l2vpn evpn
      send-community extended
  neighbor 1.1.1.72
    remote-as 65000
    update-source loopback0
    address-family l2vpn evpn
      send-community extended
  vrf TENANT1
    address-family ipv4 unicast
      redistribute direct route-map REDIST
evpn
  vni 100010 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 100020 l2
    rd auto
    route-target import auto
    route-target export auto
vrf context TENANT1
```

EVPN Instance

- The only purpose of route-map REDIST is to allow everything.
- Using the redistribute direct command, the connected VRF-aware routes are promoted into MP-BGP (type 5 routes).
- The EVPN configuration displayed above is identical to the network statement used by BGP in order to advertise MAC routes (type 2 routes).

Step 9. Configure Switch Virtual Interface (SVI) for each VLAN for the end host under VRF.

- On each leaf switch, the SVI is configured for locally configured VLAN and one SVI for L3VNI VLAN in order to achieve the Symmetric Routing Information Base (RIB).

Symmetric RIB:

- When the End host sends the data packet to a different network and it receives to the Leaf Switch, it will be processed into L2VNI first and then it will be placed to L3VNI using VRF and sent to the remote Leaf.
- Remote Leaf first receives the packets in the VRF table using Routing and then bridging to L2VNI and sends it to the end host.
- By that way, the Symmetric Routing (B-R-R-B) is achieved.

```
LEAF-1# sh run interface vlan 10,vlan 20,vlan 500

!Command: show running-config interface Vlan10, Vlan20, Vlan500
!Running configuration last done at: Sat Dec 28 09:22:19 2024
!Time: Sat Dec 28 10:00:26 2024

version 7.0(3)I7(9) Bios:version

interface Vlan10
  no shutdown
  mtu 9216
  vrf member TENANT1
  no ip redirects
  ip address 192.168.10.254/24
  no ipv6 redirects
  ip pim sparse-mode
  fabric forwarding mode anycast-gateway

interface Vlan20
  no shutdown
  mtu 9216
  vrf member TENANT1
  no ip redirects
  ip address 192.168.20.254/24
  no ipv6 redirects
  ip pim sparse-mode
  fabric forwarding mode anycast-gateway

interface Vlan500
  no shutdown
  vrf member TENANT1
  no ip redirects
  ip forward
  no ipv6 redirects

LEAF-1#
```

VLAN Interfaces

- The **IP forward** command under the VLAN 500 is used to enable Layer 3 forwarding for all VXLANs. There is no need to configure IP address, as it just processes the packet from the L2VNI table to the L3VNI table.

```
LEAF-1# show bgp vpnv4 unicast vrf TENANT1
BGP routing table information for VRF default, address family VPNv4 Unicast
BGP table version is 15, Local Router ID is 1.1.1.51
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: 1.1.1.51:3    (VRF TENANT1)
*>r192.168.10.0/24    0.0.0.0                  0        100      32768 ?
*>i192.168.10.20/32   1.1.1.53                          100          0 i
*>r192.168.20.0/24    0.0.0.0                  0        100      32768 ?
*>i192.168.20.20/32   1.1.1.54                          100          0 i

LEAF-1#
```

Learning BGP VPNv4 Routes for VRF TENANT1

- The IP address for each VLAN will be common for all the SVIs on all leaf switches. This is called anycast IP and it is used in Mobility Management where the end can communicate to another host

seamlessly without any disruption.

Step 10. Enable fabric forwarding anycast gateway MAC for the end host.

- It ensures seamless Layer 3 gateway redundancy and optimized forwarding for devices connected to the fabric.
- The Anycast Gateway MAC address is a globally consistent MAC address used for all Layer 3 gateways in a fabric.
- The concept is identical to that which is employed in First Hop Redundancy Protocol (FHRP), where each group is issued a virtual MAC.

```
LEAF-1# show running-config fabric forwarding

!Command: show running-config fabric forwarding
!Running configuration last done at: Sat Dec 28 09:22:19 2024
!Time: Sat Dec 28 10:08:08 2024

version 7.0(3)I7(9) Bios:version
nv overlay evpn
feature fabric forwarding

fabric forwarding anycast-gateway-mac 0000.1234.5678


interface Vlan10
  fabric forwarding mode anycast-gateway

interface Vlan20
  fabric forwarding mode anycast-gateway

LEAF-1#
```

Enabling Fabric Forwarding

Step 11. Enable the Access/Trunk VLAN to the member ports.

vPC Switch:

```
LEAF-1# sh run int po10 membership

!Command: show running-config interface port-channel10 membership
!Running configuration last done at: Sat Dec 28 09:22:19 2024
!Time: Sat Dec 28 10:13:19 2024

version 7.0(3)I7(9) Bios:version

interface port-channel10
  switchport mode trunk
  vpc 10

interface Ethernet1/5

  switchport mode trunk
  channel-group 10 mode active

LEAF-1#
```

Enabling Trunk Ports to the vPC Member Interface

Non-vPC Switch:

```
LEAF-3# show running-config interface e1/3

!Command: show running-config interface Ethernet1/3
!Running configuration last done at: Sat Dec 28 09:28:18 2024
!Time: Sat Dec 28 10:14:42 2024

version 7.0(3)I7(9) Bios:version

interface Ethernet1/3
  switchport access vlan 10
  spanning-tree port type edge

LEAF-3#
```

Enabling Trunk Ports to the Non vPC Member Interface

Verification

- Check the ARP and MAC address table.

```
LEAF-1# sh ip arp vrf TENANT1

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 for context TENANT1
Total number of entries: 2
Address         Age       MAC Address     Interface       Flags
192.168.20.10   00:00:36  0000.0000.bbbb  Vlan20
192.168.10.10   00:04:19  0000.0000.aaaa  Vlan10
LEAF-1#  sh ip arp suppression-cache deta

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

Ip Address      Age      Mac Address    Vlan Physical-ifindex    Flags    Remote Vtep Addrs

192.168.10.10   00:04:33 0000.0000.aaaa   10 port-channel10      L
192.168.10.20   00:55:53 0000.0000.cccc   10 (null)              R        1.1.1.53
192.168.20.10   00:00:50 0000.0000.bbbb   20 port-channel20      L
192.168.20.20   03:26:04 0000.0000.dddd   20 (null)              R        1.1.1.54
LEAF-1#
```

ARP and MAC Table on LEAF Switch 1

```
LEAF-2# show ip arp vrf TENANT1

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 for context TENANT1
Total number of entries: 2
Address         Age       MAC Address     Interface       Flags
192.168.20.10   00:01:28  0000.0000.bbbb  vlan20          +
192.168.10.10   00:00:11  0000.0000.aaaa  vlan10          +
LEAF-2#
```

ARP and MAC Table on LEAF Switch 2

- Both peers maintain the ARP entries.
- Check the Network Virtual Interface (NVI) status.

vPC Switch:

```
LEAF-1# show nve peers
Interface Peer-IP          State LearnType Uptime   Router-Mac
--------- ---------------  ----- --------- -------- -----------------
nve1      1.1.1.53         Up    CP        01:09:04 5000.0003.0007
nve1      1.1.1.54         Up    CP        03:39:16 5000.0004.0007

LEAF-1# 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

Interface VNI      Multicast-group   State Mode Type [BD/VRF]      Flags
--------- -------- ----------------- ----- ---- ------------------ -----
nve1      50000    n/a               Up    CP   L3 [TENANT1]
nve1      100010   239.0.0.10        Up    CP   L2 [10]            SA
nve1      100020   239.0.0.20        Up    CP   L2 [20]            SA

LEAF-1#
```

NVE Peers on vPC Switch

Non-vPC Switch:

```
LEAF-3# show nve peers
Interface Peer-IP          State LearnType Uptime   Router-Mac
--------- ---------------  ----- --------- -------- -----------------
nve1      1.1.1.54         Up    CP        01:14:00 5000.0004.0007
nve1      10.1.1.100       Up    CP        01:14:16 5000.0001.0007

LEAF-3#
```

NVE Peers on Non vPC Switch

- Here, you will notice that the peer IP is 10.1.1.100 instead of the primary loopback IP address, so the return packet will be routed for that IP to any of the vPC switches.
- Check BGP EVPN routes.

```
LEAF-1# 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
Topology    Mac Address    Prod   Flags         Seq No     Host IP         Next-Hops
----------- -------------- ------ ----------- ---------- --------------- --------------
10          0000.0000.aaaa HMM    --          0          192.168.10.10   Local
10          0000.0000.cccc BGP    --          0          192.168.10.20   1.1.1.53
20          0000.0000.bbbb HMM    --          0          192.168.20.10   Local
20          0000.0000.dddd BGP    --          0          192.168.20.20   1.1.1.54
LEAF-1#
```

BGP l2route EVPN MAC-IP

```
LEAF-1# 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

Topology    Mac Address    Prod   Flags         Seq No     Next-Hops
----------- -------------- ------ ------------ ---------- ----------------
10          0000.0000.aaaa Local  L,           0          Po10
10          0000.0000.cccc BGP    Spl          0          1.1.1.53
20          0000.0000.bbbb Local  L,           0          Po20
20          0000.0000.dddd BGP    SplRcv       0          1.1.1.54
500         5000.0003.0007 VXLAN  Rmac         0          1.1.1.53
500         5000.0004.0007 VXLAN  Rmac         0          1.1.1.54
LEAF-1#
```

BGP l2route EVPN MAC

```
LEAF-1# show bgp l2vpn evpn summary
BGP summary information for VRF default, address family L2VPN EVPN
BGP router identifier 1.1.1.51, local AS number 65000
BGP table version is 134, L2VPN EVPN config peers 2, capable peers 2
12 network entries and 15 paths using 2568 bytes of memory
BGP attribute entries [12/1920], BGP AS path entries [0/0]
BGP community entries [0/0], BGP clusterlist entries [4/16]

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
1.1.1.71        4 65000     312     263      134    0    0 03:46:01 3
1.1.1.72        4 65000     307     256      134    0    0 03:39:21 3
LEAF-1#
```

BGP EVPN Summary

```
LEAF-1# show bgp l2vpn evpn
BGP routing table information for VRF default, address family L2VPN EVPN
BGP table version is 146, Local Router ID is 1.1.1.51
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: 1.1.1.51:32777    (L2VNI 100010)
*>l[2]:[0]:[0]:[48]:[0000.0000.aaaa]:[0]:[0.0.0.0]/216
                      10.1.1.100                             100      32768 i
*>l[2]:[0]:[0]:[48]:[0000.0000.aaaa]:[32]:[192.168.10.10]/272
                      10.1.1.100                             100      32768 i
*>i[2]:[0]:[0]:[48]:[0000.0000.cccc]:[32]:[192.168.10.20]/272
                      1.1.1.53                               100          0 i

Route Distinguisher: 1.1.1.51:32787    (L2VNI 100020)
*>l[2]:[0]:[0]:[48]:[0000.0000.bbbb]:[0]:[0.0.0.0]/216
                      10.1.1.100                             100      32768 i
*>i[2]:[0]:[0]:[48]:[0000.0000.dddd]:[0]:[0.0.0.0]/216
                      1.1.1.54                               100          0 i
*>l[2]:[0]:[0]:[48]:[0000.0000.bbbb]:[32]:[192.168.20.10]/272
                      10.1.1.100                             100      32768 i
*>i[2]:[0]:[0]:[48]:[0000.0000.dddd]:[32]:[192.168.20.20]/272
                      1.1.1.54                               100          0 i

Route Distinguisher: 1.1.1.53:32777
*>i[2]:[0]:[0]:[48]:[0000.0000.cccc]:[32]:[192.168.10.20]/272
                      1.1.1.53                               100          0 i
* i                   1.1.1.53                               100          0 i

Route Distinguisher: 1.1.1.54:32787
* i[2]:[0]:[0]:[48]:[0000.0000.dddd]:[0]:[0.0.0.0]/216
                      1.1.1.54                               100          0 i
*>i                   1.1.1.54                               100          0 i
* i[2]:[0]:[0]:[48]:[0000.0000.dddd]:[32]:[192.168.20.20]/272
                      1.1.1.54                               100          0 i
*>i                   1.1.1.54                               100          0 i

Route Distinguisher: 1.1.1.51:3    (L3VNI 50000)
*>i[2]:[0]:[0]:[48]:[0000.0000.cccc]:[32]:[192.168.10.20]/272
                      1.1.1.53                               100          0 i
*>i[2]:[0]:[0]:[48]:[0000.0000.dddd]:[32]:[192.168.20.20]/272
                      1.1.1.54                               100          0 i

LEAF-1#
```

BGP EVPN Routes

- It is common to question how Leaf Switches acquire MAC entries for remote hosts. This process is facilitated by Gratuitous ARP. When a network port is activated, it immediately sends an ARP request in order to verify the uniqueness of the IP address. Each Leaf Switch then records the MAC address and includes it in a BGP Update Packet. This allows other Leaf Switches to update their respective MAC address tables accordingly. But there can be a case where the end host does not generate Gratuitous ARP (Silent host), and in that case, the ARP request will be broadcast to the leaf and as it is a broadcast request, the Leaf switch will generate the multicast request to the respective group for the particular VNI. In this case, it is 239.0.0.10 and 239.0.0.20.

- Lets ping from Host-1 to Host-3 within the same VNI and look at the capture.

```
HOST-1#ping 192.168.10.20 rep 2
Type escape sequence to abort.
Sending 2, 100-byte ICMP Echos to 192.168.10.20, timeout is 2 seconds:
!!
Success rate is 100 percent (2/2), round-trip min/avg/max = 11/11/12 ms
HOST-1#
```

Pinging from HOST-1 to HOST-3

Internet Control Message Protocol (ICMP) Packet over the VXLAN:

```
> Frame 213: 164 bytes on wire (1312 bits), 164 bytes captured (1312 bits) on interface -, id 0
> Ethernet II, Src: 50:00:00:06:00:07 (50:00:00:06:00:07), Dst: 50:00:00:03:00:07 (50:00:00:03:00:07)
> Internet Protocol Version 4, Src: 10.1.1.100, Dst: 1.1.1.53
v User Datagram Protocol, Src Port: 50413, Dst Port: 4789
    Source Port: 50413
    Destination Port: 4789
    Length: 130
  > Checksum: 0x0000 [zero-value ignored]
    [Stream index: 24]
    [Stream Packet Number: 1]
  > [Timestamps]
    UDP payload (122 bytes)
v Virtual eXtensible Local Area Network
  > Flags: 0x0800, VXLAN Network ID (VNI)
    Group Policy ID: 0
    VXLAN Network Identifier (VNI): 100010
    Reserved: 0
> Ethernet II, Src: 00:00:00_00:aa:aa (00:00:00:00:aa:aa), Dst: 00:00:00_00:cc:cc (00:00:00:00:cc:cc)
v Internet Protocol Version 4, Src: 192.168.10.10, Dst: 192.168.10.20
    0100 .... = Version: 4
    .... 0101 = Header Length: 20 bytes (5)
  > Differentiated Services Field: 0x00 (DSCP: CS0, ECN: Not-ECT)
    Total Length: 100
    Identification: 0x0000 (0)
  > 000. .... = Flags: 0x0
    ...0 0000 0000 0000 = Fragment Offset: 0
    Time to Live: 255
    Protocol: ICMP (1)
    Header Checksum: 0x262a [validation disabled]
    [Header checksum status: Unverified]
    Source Address: 192.168.10.10
    Destination Address: 192.168.10.20
    [Stream index: 11]
> Internet Control Message Protocol
```

Wireshark Capture Showing ICMP Request Packet Travelling Through L2VNI 10010

- As you can see, the source IP is 10.1.1.100 with port 4789 as the UDP destination.
- Since it is an intra-VNI communication, VLAN 10 will use VNI 100010, and VLAN 20 will use VNI 1000.
- Lets ping from Host-1 to Host-4 with different VNI and look at the capture.

```
HOST-1#ping 192.168.20.20
Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 192.168.20.20, timeout is 2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 11/13/21 ms
HOST-1#
```

Pinging from HOST-1 to HOST-4

ICMP Packet over the VXLAN:

```
> Frame 27: 164 bytes on wire (1312 bits), 164 bytes captured (1312 bits) on interface -, id 0
> Ethernet II, Src: 50:00:00:05:00:07 (50:00:00:05:00:07), Dst: 50:00:00:04:00:07 (50:00:00:04:00:07)
> Internet Protocol Version 4, Src: 10.1.1.100, Dst: 1.1.1.54
v User Datagram Protocol, Src Port: 54712, Dst Port: 4789
    Source Port: 54712
    Destination Port: 4789
    Length: 130
  > Checksum: 0x0000 [zero-value ignored]
    [Stream index: 3]
    [Stream Packet Number: 1]
  > [Timestamps]
    UDP payload (122 bytes)
v Virtual eXtensible Local Area Network
  > Flags: 0x0800, VXLAN Network ID (VNI)
    Group Policy ID: 0
    VXLAN Network Identifier (VNI): 50000
    Reserved: 0
> Ethernet II, Src: 50:00:00:01:00:07 (50:00:00:01:00:07), Dst: 50:00:00:04:00:07 (50:00:00:04:00:07)
> Internet Protocol Version 4, Src: 192.168.10.10, Dst: 192.168.20.20
> Internet Control Message Protocol
```

Wireshark Capture Showing ICMP Request Packet Travelling Through L3VNI 50000

- Since it is an inter-VNI communication, the L3VNI 50000 will be used.
- Check the ARP table for end host.

```
HOST-1#sh ip arp
Protocol  Address          Age (min)  Hardware Addr   Type   Interface
Internet  192.168.10.10           -   0000.0000.aaaa  ARPA   GigabitEthernet0/0
Internet  192.168.10.20          18   0000.0000.cccc  ARPA   GigabitEthernet0/0
Internet  192.168.10.254          3   0000.1234.5678  ARPA   GigabitEthernet0/0
HOST-1#
```

HOST-1 ARP Entries

```
HOST-2#sh ip arp
Protocol  Address          Age (min)  Hardware Addr   Type   Interface
Internet  192.168.20.10           -   0000.0000.bbbb  ARPA   GigabitEthernet0/0
Internet  192.168.20.20          44   0000.0000.dddd  ARPA   GigabitEthernet0/0
Internet  192.168.20.254          4   0000.1234.5678  ARPA   GigabitEthernet0/0
HOST-2#
```

HOST-2 ARP Entries

```
HOST-3#sh ip arp
Protocol  Address          Age (min)  Hardware Addr   Type   Interface
Internet  192.168.10.10         103   0000.0000.aaaa  ARPA   GigabitEthernet0/0
Internet  192.168.10.20           -   0000.0000.cccc  ARPA   GigabitEthernet0/0
Internet  192.168.10.254         10   0000.1234.5678  ARPA   GigabitEthernet0/0
HOST-3#
```

HOST-3 ARP Entries

```
HOST-4#sh ip arp
Protocol  Address          Age (min)  Hardware Addr   Type   Interface
Internet  192.168.20.10          43   0000.0000.bbbb  ARPA   GigabitEthernet0/0
Internet  192.168.20.20           -   0000.0000.dddd  ARPA   GigabitEthernet0/0
Internet  192.168.20.254          6   0000.1234.5678  ARPA   GigabitEthernet0/0
HOST-4#
```

HOST-4 ARP Entries

```
HOST-4#tclsh
HOST-4(tcl)#set ip_list {192.168.10.10 192.168.10.20 192.168.20.10 192.168.20.20}
192.168.10.10 192.168.10.20 192.168.20.10 192.168.20.20
HOST-4(tcl)#foreach ip $ip_list {
HOST-4(tcl)#foreach ip $ip_list {
+>            puts "Pinging $ip rep 50 size 1500"
+>            set result [exec ping $ip]
+>            puts $result
+>        }
Pinging 192.168.10.10 rep 50 size 1500

Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 192.168.10.10, timeout is 2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 12/14/16 ms
Pinging 192.168.10.20 rep 50 size 1500

Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 192.168.10.20, timeout is 2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 10/12/15 ms
Pinging 192.168.20.10 rep 50 size 1500

Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 192.168.20.10, timeout is 2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 8/11/17 ms
Pinging 192.168.20.20 rep 50 size 1500

Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 192.168.20.20, timeout is 2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 1/1/3 ms

HOST-4(tcl)#
```

Pinging from HOST-4 to All Other End Hosts