VxLAN Routing and Control Plane on Nexus 9000 Series Switches

- Lilian Quan – Technical Marketing Engineering, INSBU
- Chad Hintz – TSA, US Commercial

Contributors and Acknowledgements

- Lukas Krattiger
- Victor Moreno
- Yves Louis
- Brenden Buresh
- Jason Gmitter
- Chad Hintz
- Errol Roberts
- Cesar Obediente

- Leo Boulton
- Vaughn Suazo
- Dave Malik
- Lilian Quan
- Mike Herbert
- Juan Lage
- Jason Pfiefer
- Lilian Quan

- David Jansen
- Kevin Corbin
- Babi Seal
- James Christopher
- Jim Pisano
- Matt Smorto
- Priyam Reddy

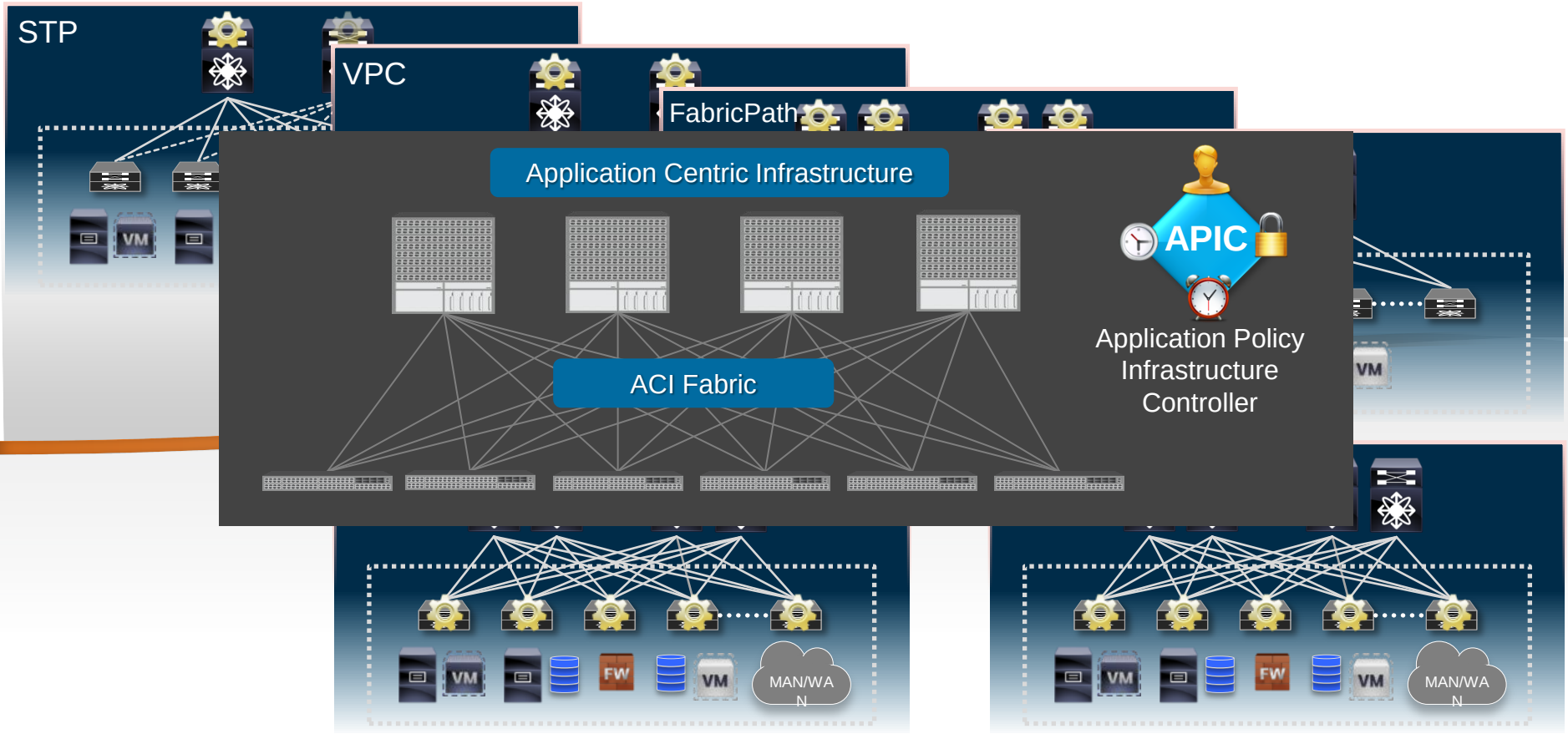
Agenda

- VxLAN Overview
- MP-BGP EVPN Basics
- MP-BGP EVPN Control Plane
- VXLAN Design Options
- MP-BGP EVPN VXLAN Configuration
- VxLAN Capability on Nexus 9000 Series Switches

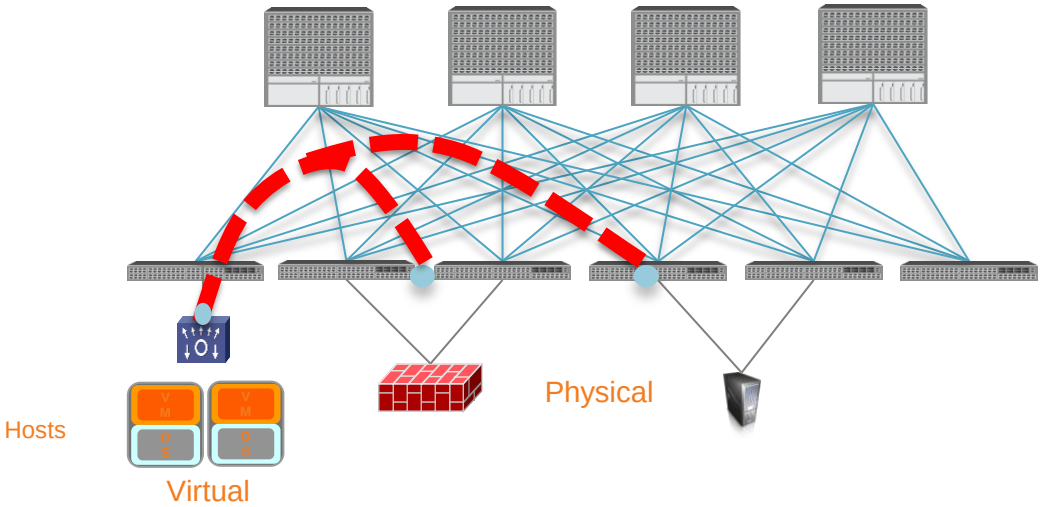
Agenda

- VxLAN Overview
- MP-BGP EVPN Basics
- MP-BGP EVPN Control Plane
- VXLAN Design Options
- MP-BGP EVPN VXLAN Configuration
- VxLAN Capability on Nexus 9000 Series Switches

Data Center “Fabric” Journey



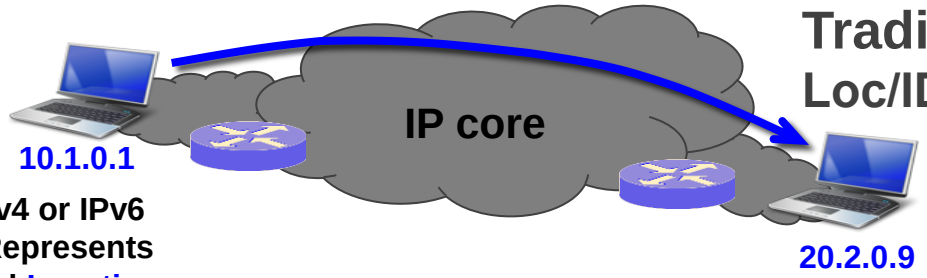
Trend: Flexible Data Center Fabrics



Workload Mobility
Workload Placement
Segmentation
Scale
Automation & Programmability
L2 + L3 Connectivity
Physical + Virtual
Open

Why Do We Need Overlays?

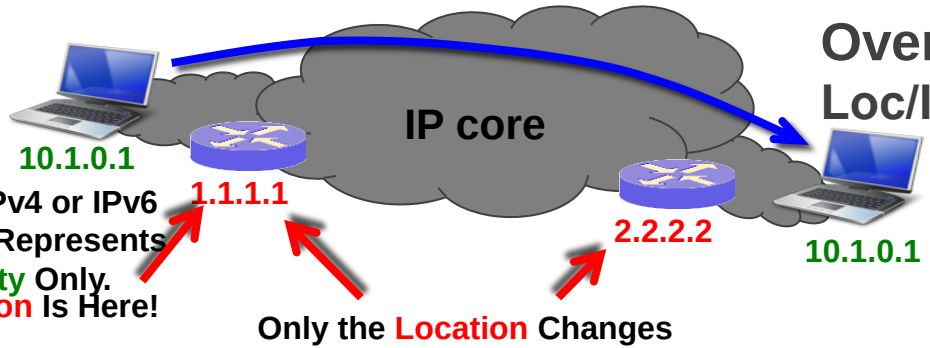
Location and Identity Separation



Traditional Behaviour
Loc/ID "Overloaded" Semantic

Device IPv4 or IPv6 Address Represents Identity and Location

When the Device Moves, It Gets a New IPv4 or IPv6 Address for Its New Identity and Location



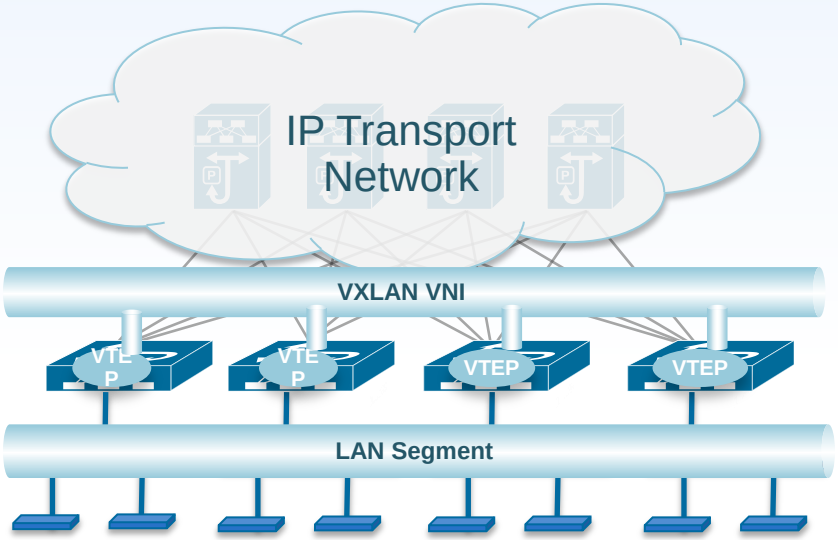
Overlay Behaviour
Loc/ID "Split"

Device IPv4 or IPv6 Address Represents Identity Only. Its Location Is Here!

When the Device Moves, Keeps Its IPv4 or IPv6 Address. It Has the Same Identity

Only the Location Changes

Network Virtualization with VXLAN



Underlay Network:

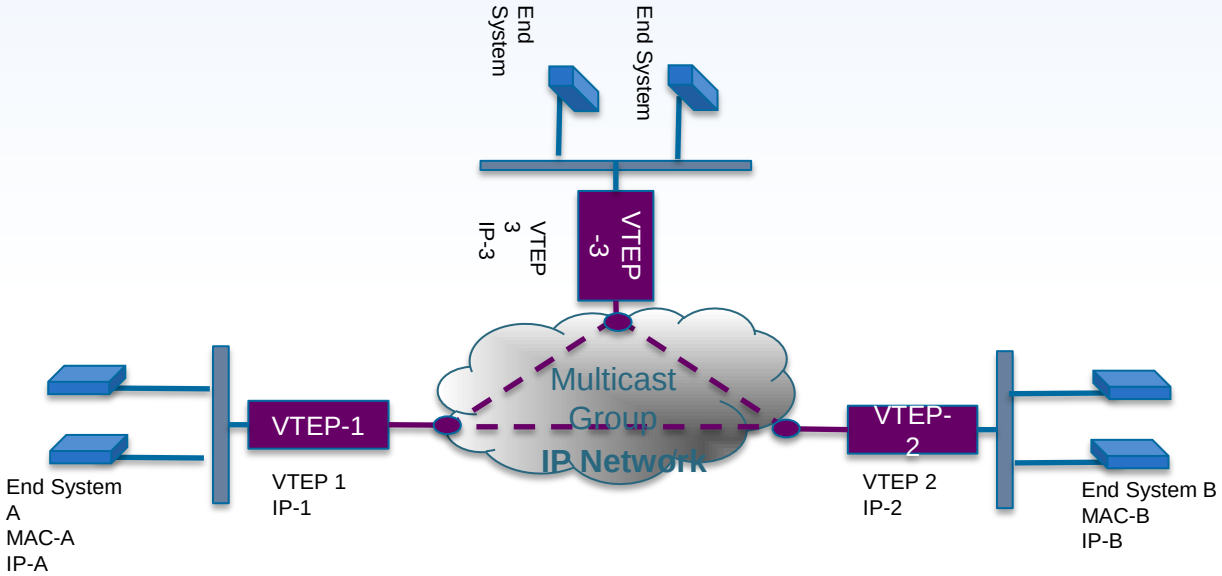
- IP routing – proven, stable, scalable
- ECMP – utilize all available network paths

Overlay Network:

- Standards-based overlay
- Layer-2 extensibility and mobility
- Expanded Layer-2 name space
- Scalable network domain
- Multi-Tenancy

Multicast-Based VxLAN

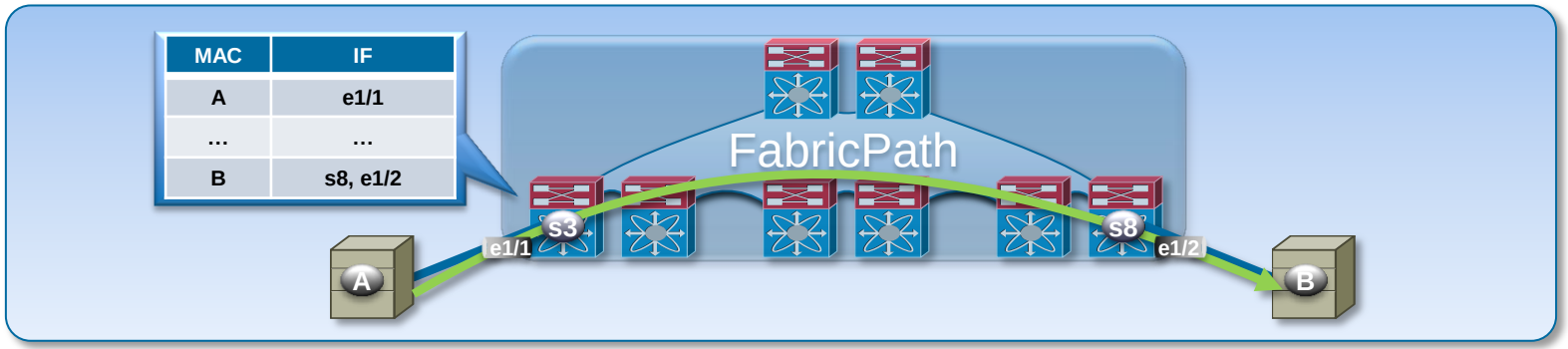
- No VXLAN control plane
- Data driven flood-&-learn
- Multicast transport for VXLAN BUM (Broadcast, Unknown Unicast and Multicast) traffic.



Sound Familiar?

FabricPath

Shortest path any to any



- Single address lookup at the ingress edge identifies the exit port across the fabric
- Traffic is then switched using the shortest path available
- Reliable L2 and L3 connectivity any to any
  (L2 as if it was within the same switch, **no STP inside**)

The Secret Sauce is the Control Plane, not the Encapsulation

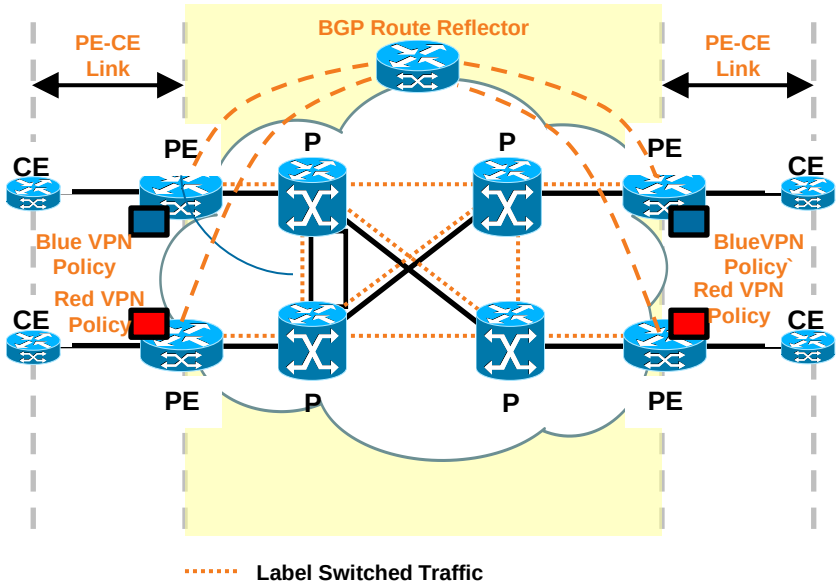
Agenda

- VxLAN Overview
- MP-BGP EVPN Basics
- MP-BGP EVPN Control Plane
- VxLAN Design Options
- MP-BGP EVPN VXLAN Configuration
- VxLAN Capability on Nexus 9000 Series Switches

MP-BGP with MPLS VPN Route Distribution

Exchange of VPN Policies Among PE Routers

- Full mesh of BGP sessions among all PE routers
  - BGP Route Reflector
- Multi-Protocol BGP extensions (MP-iBGP) to carry VPN policies
- PE-CE routing options
  - Static routes
  - eBGP
  - OSPF
  - IS-IS



VPN Control Plane Processing

VRF Parameters

Make customer routes unique:

- **Route Distinguisher (RD):**
  8-byte field, VRF parameters; unique value to make VPN IP routes unique
- **VPNv4 address**: RD + VPN IP prefix

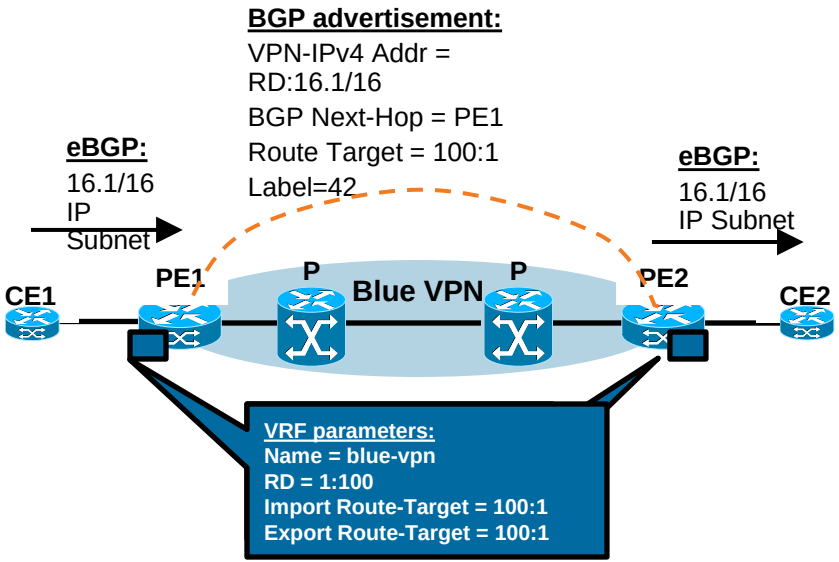
Selective distribute VPN routes:

- **Route Target (RT):** 8-byte field, VRF parameter, unique value to define the import/export rules for VPNv4 routes
- MP-iBGP: advertises VPNv4 prefixes + labels

VPN Control Plane Processing

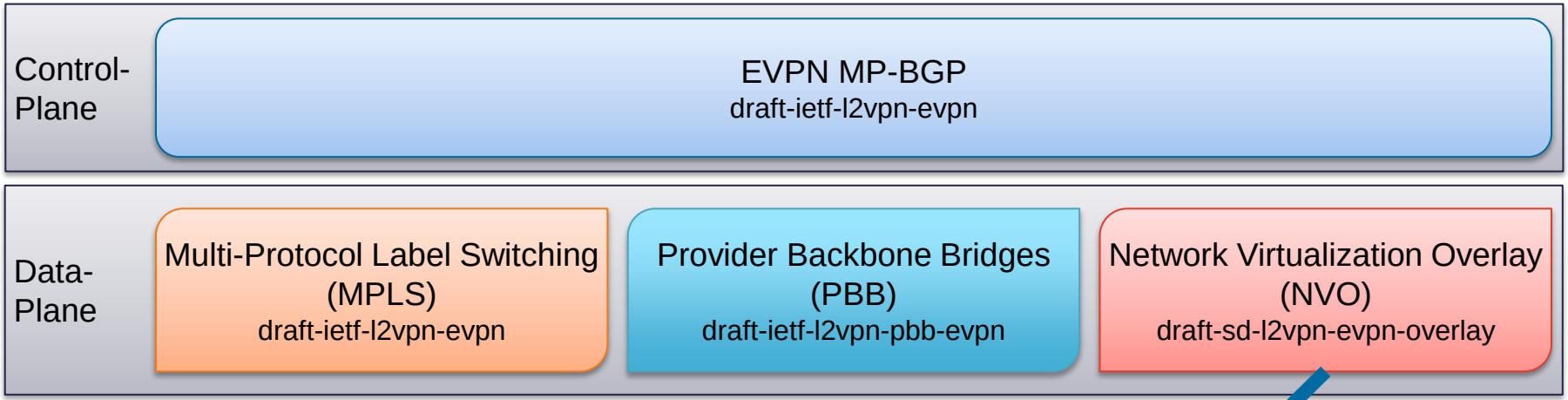
Interactions Between VRF and BGP VPN Signaling

1. CE1 redistribute IPv4 route to PE1 via eBGP
2. PE1 allocates VPN label for prefix learnt from CE1 to create unique VPNv4 route
3. PE1 redistributes VPNv4 route into MP-iBGP, it sets itself as a next hop and relays VPN site routes to PE2
4. PE2 receives VPNv4 route and, via processing in local VRF (green), it redistributes original IPv4 route to CE2



EVPN – Ethernet VPN

VXLAN Evolution

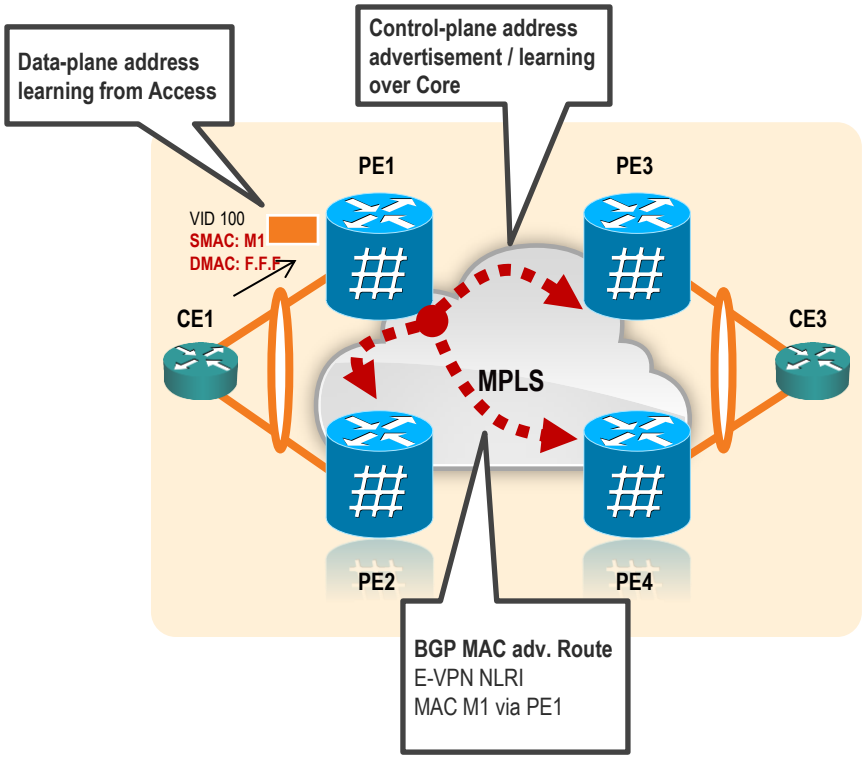


- EVPN over NVO Tunnels (VXLAN, NVGRE, MPLSoE) for Data Center Fabric encapsulations
- Provides Layer-2 and Layer-3 Overlays over simple IP Networks

Ethernet VPN

Highlights

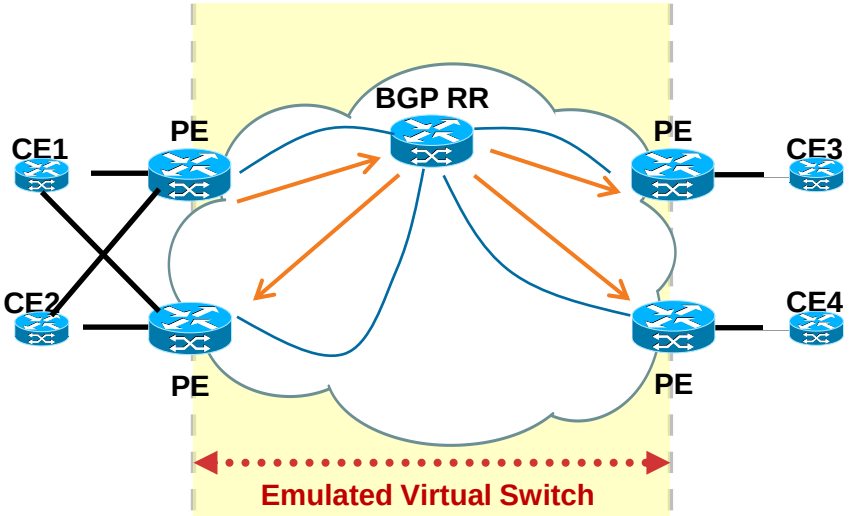
- Next generation solution for Ethernet multipoint connectivity services
  - Leverage similarities with L3VPN
- PEs run Multi-Protocol BGP to advertise & learn MAC addresses over Core
- Learning on PE Access Circuits via data-plane transparent learning
- No pseudowire full-mesh required
  - Unicast: use MP2P tunnels
  - Multicast: use ingress replication over MP2P tunnels or use LSM
- Under standardization at IETF – draft-ietf-l2vpn-evpn



EVPN

- Multi-Protocol BGP (MP-BGP) based Control-Plane using EVPN NLRI (Network Layer Reachability Information)
- Make Forwarding decisions at VTEPs for Layer-2 (MAC) and Layer-3 (IP)
- Discovery: BGP, using MPLS VPN mechanisms (RT)
- Signaling: BGP
- Learning: **Control** plane (BGP)

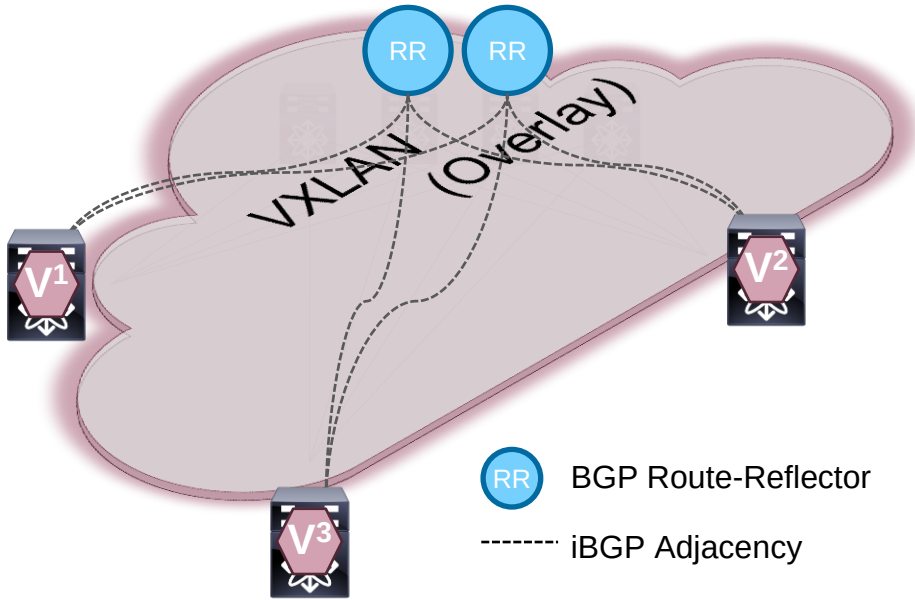
BGP advertisement:
L2VPN/EVPN Addr = **CE1.MAC**
BGP Next-Hop = PE1
Route Target = 100:1
Label=42



Host and Subnet Route Distribution

VXLAN/EVPN

- Host Route Distribution decoupled from the Underlay protocol
- Use MultiProtocol-BGP (MP-BGP) on the Leaf nodes to distribute internal Host/Subnet Routes and external reachability information
- Route-Reflectors deployed for scaling purposes



Agenda

- VxLAN Overview
- MP-BGP EVPN Basics
- MP-BGP EVPN Control Plane
- VXLAN Design Options
- MP-BGP EVPN VXLAN Configuration
- VxLAN Capability on Nexus 9000 Series Switches

Decoding an Overlay Technology

Overlay services
- Layer-2
- Layer-3
- Layer-2 + Layer-3

Tunnel Encapsulation

Underlay transport network

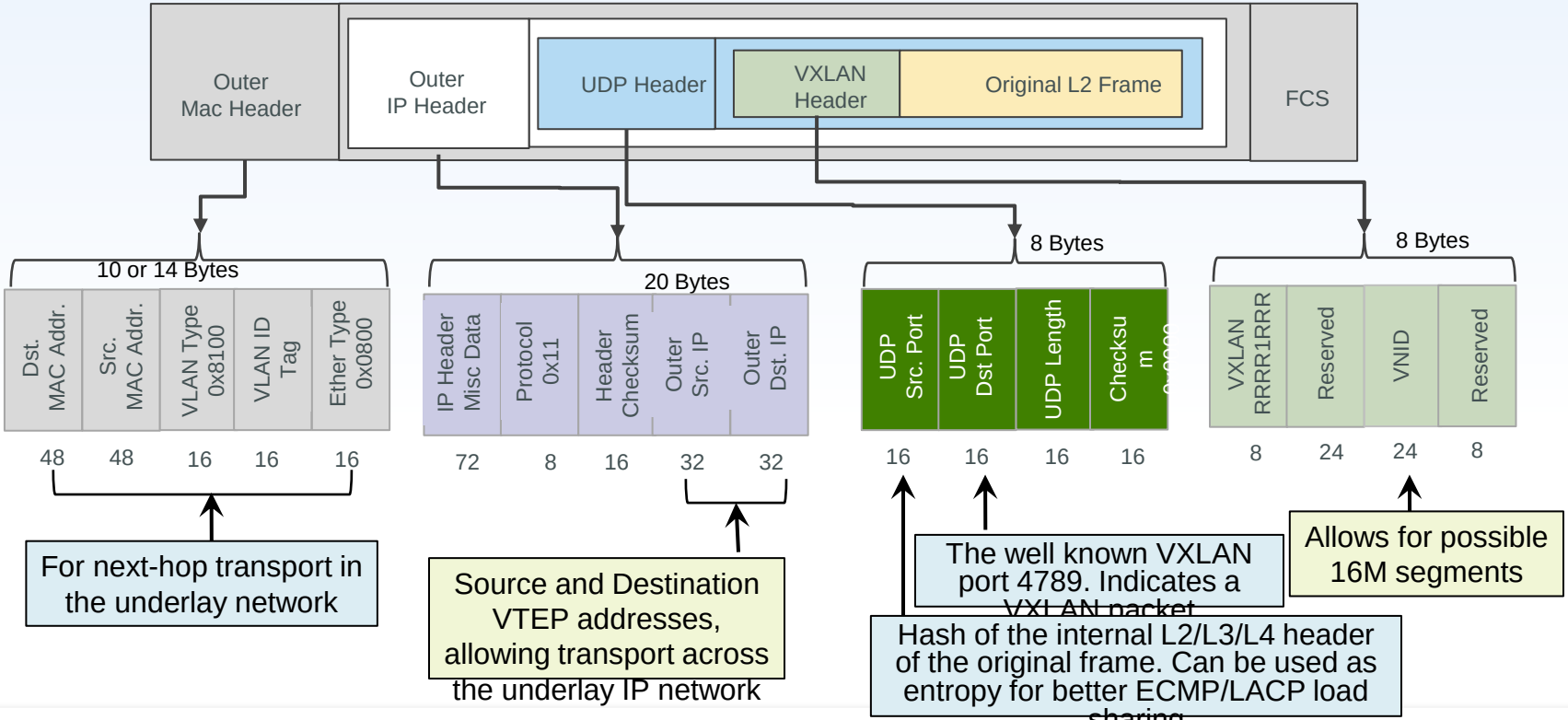
Control Plane

- Peer discovery mechanism
- Route learning and distribution mechanism
  - Local learning
  - Remote learning

Data Plane

- Overlay L2/L3 Unicast traffic
- Overlay Broadcast, Unknown (Layer-2) traffic, Multicast traffic (BUM traffic) forwarding

VxLAN Tunnel Encapsulation(MAC-in-UDP)



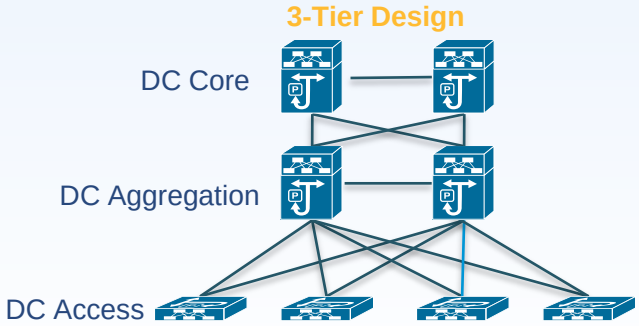
VXLAN Underlay Network

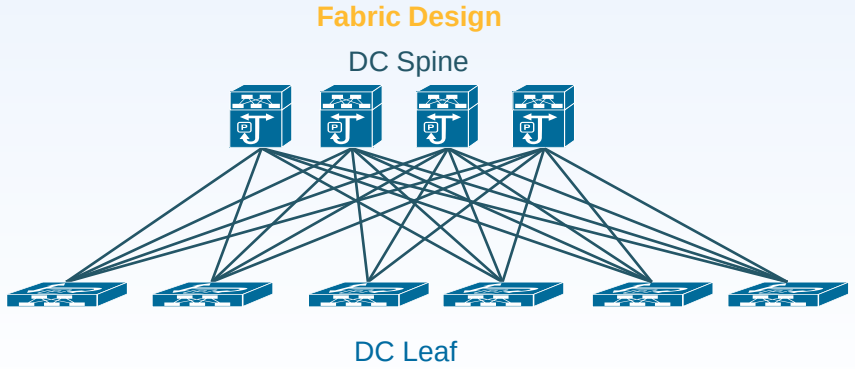
- IP routed Network

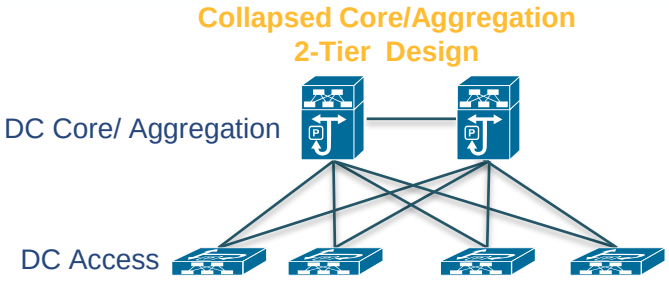
  - Flexible topologies
  - Recommend a network with redundant paths using ECMP for load sharing

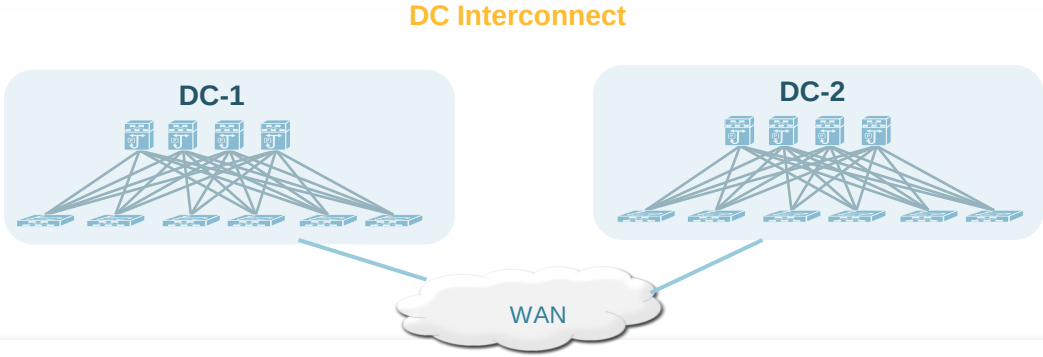
- Support any routing protocols --- OSFP, EIGRP, IS-IS, BGP, etc.
- Multicast is needed if using multicast for overlay BUM replication and transport

VXLAN Underlay Network – Typical DC Topologies



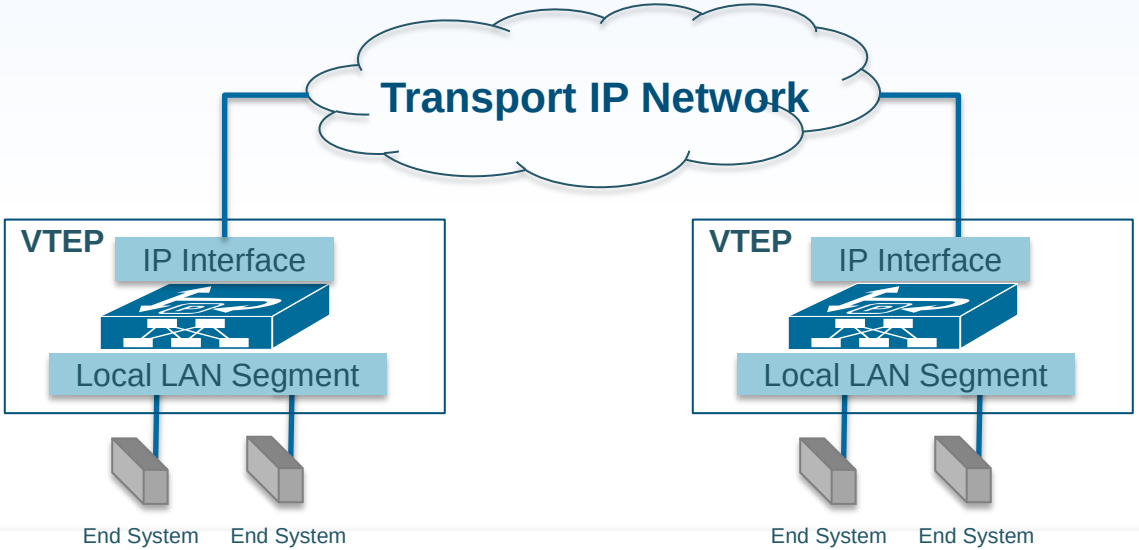






VXLAN VTEP

VXLAN terminates its tunnels on VTEPs (Virtual Tunnel End Point).
Each VTEP has two interfaces, one is to provide bridging function for local hosts, the other has an IP identification in the core network for VXLAN encapsulation/decapsulation.



Agenda

- VxLAN Overview
- MP-BGP EVPN Basics
- MP-BGP EVPN Control Plane
- VXLAN Design Options
- MP-BGP EVPN VXLAN Configuration
- VxLAN Capability on Nexus 9000 Series Switches

VXLAN: Flood-&-Learn vs EVPN Control Plane

	Flood-&-Learn	EVPN Control Plane
Overlay Services	L2+L3	L2+L3
Underlay Network	IP network with ECMP	IP network with ECMP
Encapsulation	MAC in UDP	MAC in UDP
Peer Discovery	Data-driven flood-&-learn	MP-BGP
Peer Authentication	Not available	MP-BGP
Host Route Learning	Local hosts: Data-driven flood-&-learn Remote hosts: Data-driven flood-&-learn	Local Host: Data-driven Remote host: MP-BGP
Host Route Distribution	No route distribution.	MP-BGP
L2/L3 Unicast Forwarding	Unicast encap	Unicast encap
BUM Traffic forwarding	Multicast replication Unicast/Ingress replication	Multicast replication Unicast/Ingress replication

EVPN

MP-BGP for EVPN

- MP-BGP is the routing protocol for EVPN
- Multi-tenancy construct using VRF (Rout Distinguisher, Route Targets)
- New address-family “l2vpn evpn” for distributing EVPN routes
- EVPN routes = [MAC] + [IP]
- iBGP or eBGP support

```
vrf context evpn-tenant-1
  vni 39000
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
```

```
evpn
  vni 20000 l2
    rd auto
    route-target import auto
    route-target export auto
```

```
router bgp 100
  router-id 10.1.1.11
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  vrf evpn-tenant-1
    address-family ipv4 unicast
      advertise l2vpn evpn
```

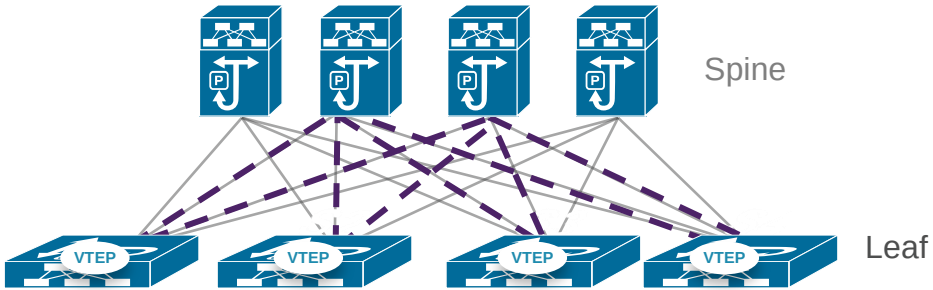
VXLAN EVPN Control Plane Functions in Bronte Release

- **Host MAC/IP advertisements through MP-BGP**
- **VTEP Peer Auto-discovery and Authentication via MP-BGP**
- **Anycast IP gateway**
- **ARP Suppression**
- **Ingress Replication with Head-end Auto-discovery (planned for Bronte+)**

EVPN Control Plane – Reachability Distribution

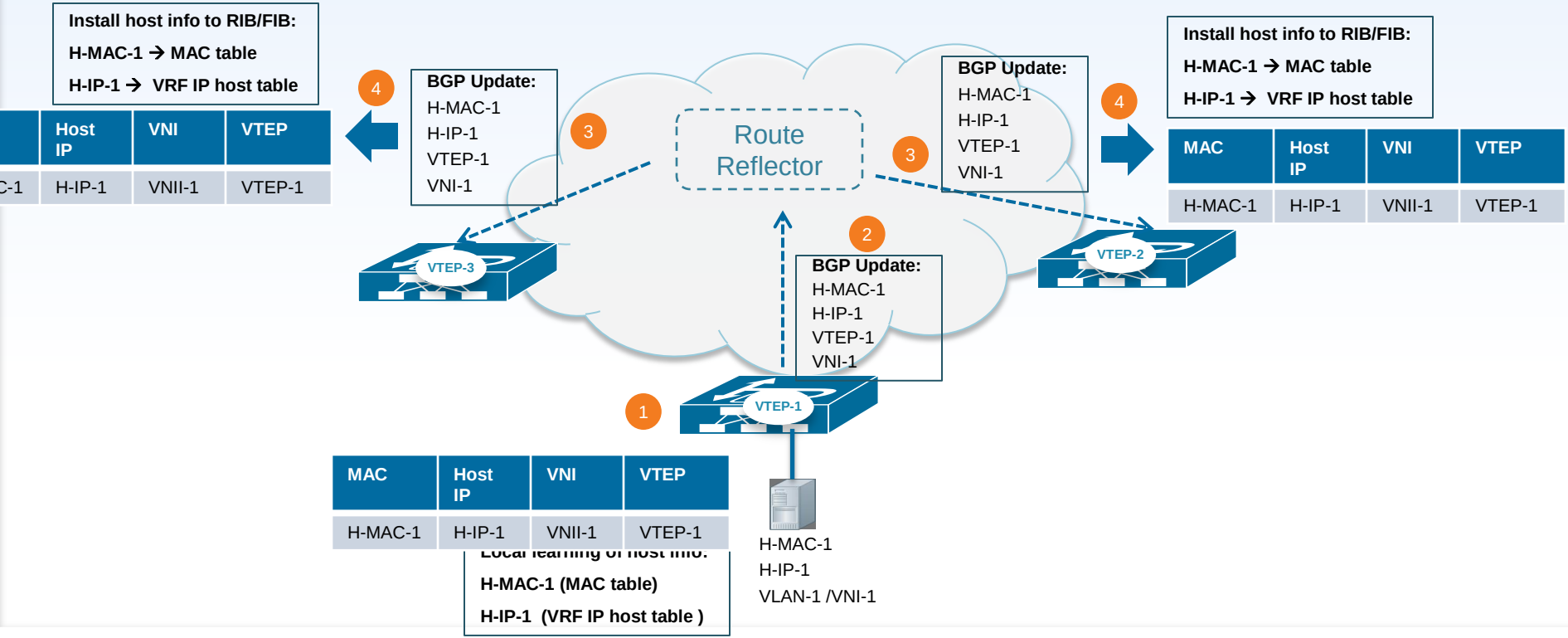
EVPN Control Plane -- Host and Subnet Route Distribution

BGP Update
- Host-MAC
- Host-IP
- Internal IP Subnet
- External Prefixes



- Use MP-BGP with EVPN Address Family on VTEPs to distribute internal host MAC/IP addresses, subnet routes and external reachability information
- MP-BGP enhancements to carry up to 100s of thousands of routes with reduced convergence time

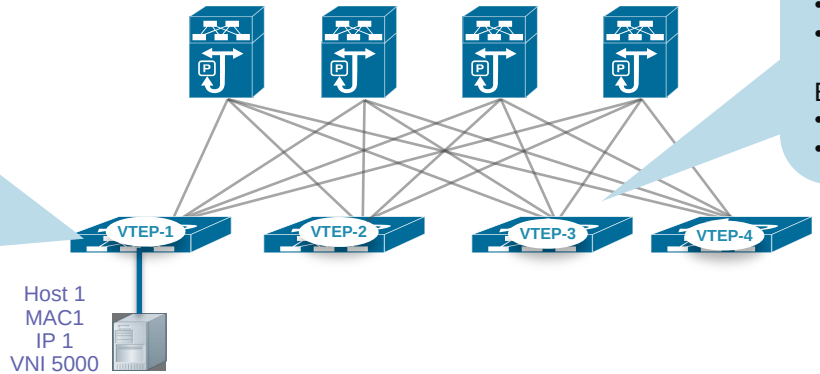
Host Advertisement



EVPN Control Plane --- Host Movement

NLRI:
- Host MAC1, IP1
- NVE IP 1
- VNI 5000
- Next-Hop: VTEP-1

Ext. Community:
- Encapsulation: VXLAN
- Cost/Sequence: 0



NLRI:
- Host MAC1, IP1
- NVE IP 1
- VNI 5000
- Next-Hop: VTEP-3

Ext. Community:
- Encapsulation: VXLAN
- Cost/Sequence: 1

MAC	IP	VNI	Next-Hop	Encap	Seq
MAC-1	IP-1	5000	VTEP-1	VXLAN	0

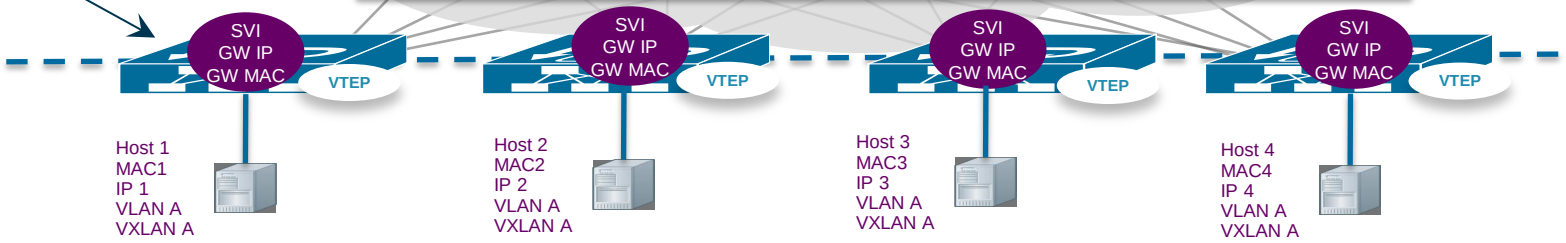
MAC	IP	VNI	Next-Hop	Encap	Seq
MAC-1	IP-1	5000	VTEP-3	VXLAN	1

1. VTEP-1 detects Host1 and advertise an EVPN route for Host1 with seq# 0
2. Host1 Moves behind VTEP-3
3. VTEP-3 detects Host1 and advertises an EVPN route for Host1 with seq #1
4. VTEP-1 sees more recent route and withdraws its advertisement

Anycast-Gateway

The same anycast gateway virtual IP address and MAC address need to be configured on all VTEPs in the VNI

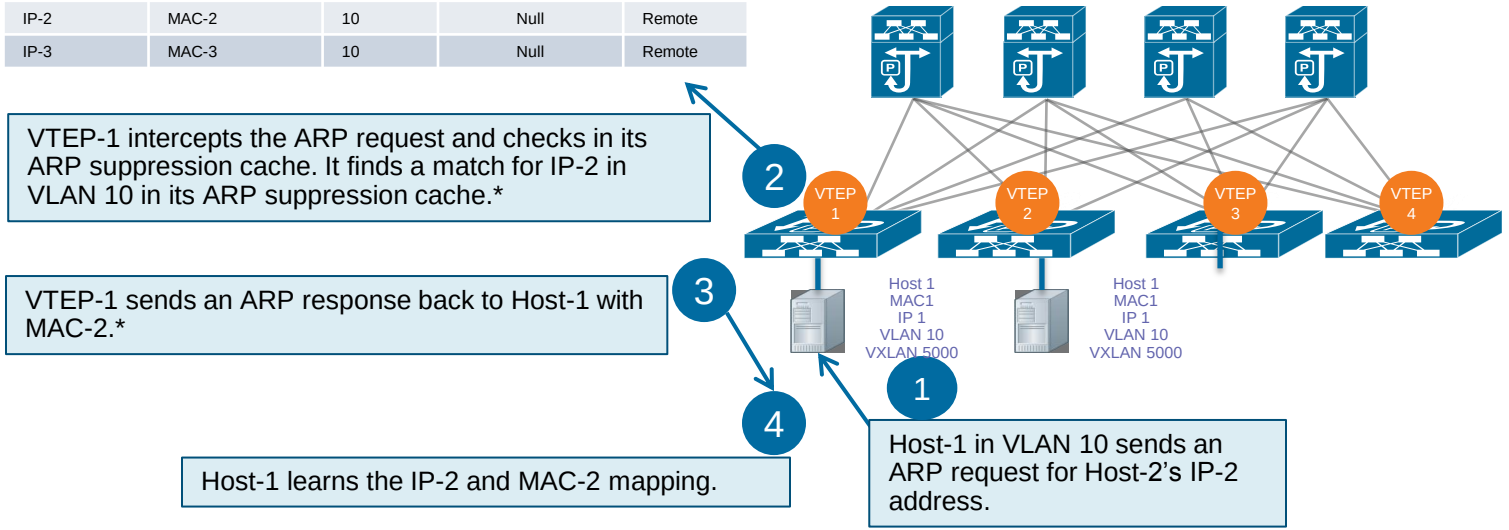
```
# VLAN to VNI mapping
vlan 200
 vn-segment 5200

# Anycast Gateway MAC, identically configured on all VTEPs
fabric forwarding anycast-gateway-mac 0002.0002.0002

# Distributed IP Anycast Gateway (SVI)
# Gateway IP address needs to be identically configured on all VTEPs
interface vlan 200
 no shutdown
 vrf member Tenant-A
 ip address 20.0.0.1/24
 fabric forwarding mode anycast-gateway
```



ARP Suppression in MP-BGP EVPN

ARP suppression reduces network flooding due to host learning

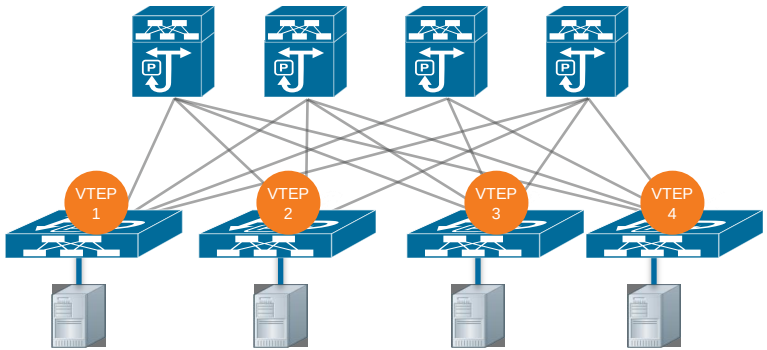
IP Address	MAC Address	VLAN	Physical Interface Index (ifindex)	Flags
IP-1	MAC-1	10	E1/1	Local
IP-2	MAC-2	10	Null	Remote
IP-3	MAC-3	10	Null	Remote



* If VTEP-1 doesn't have a match for IP-2 in its ARP suppression cache table, it will flood the ARP request to all other VTEPs in this VNI

ARP Suppression in MP-BGP EVPN (Cont'ed)

- ARP Suppression can be enabled on a per-VNI basis under the interface nve1 configuration.



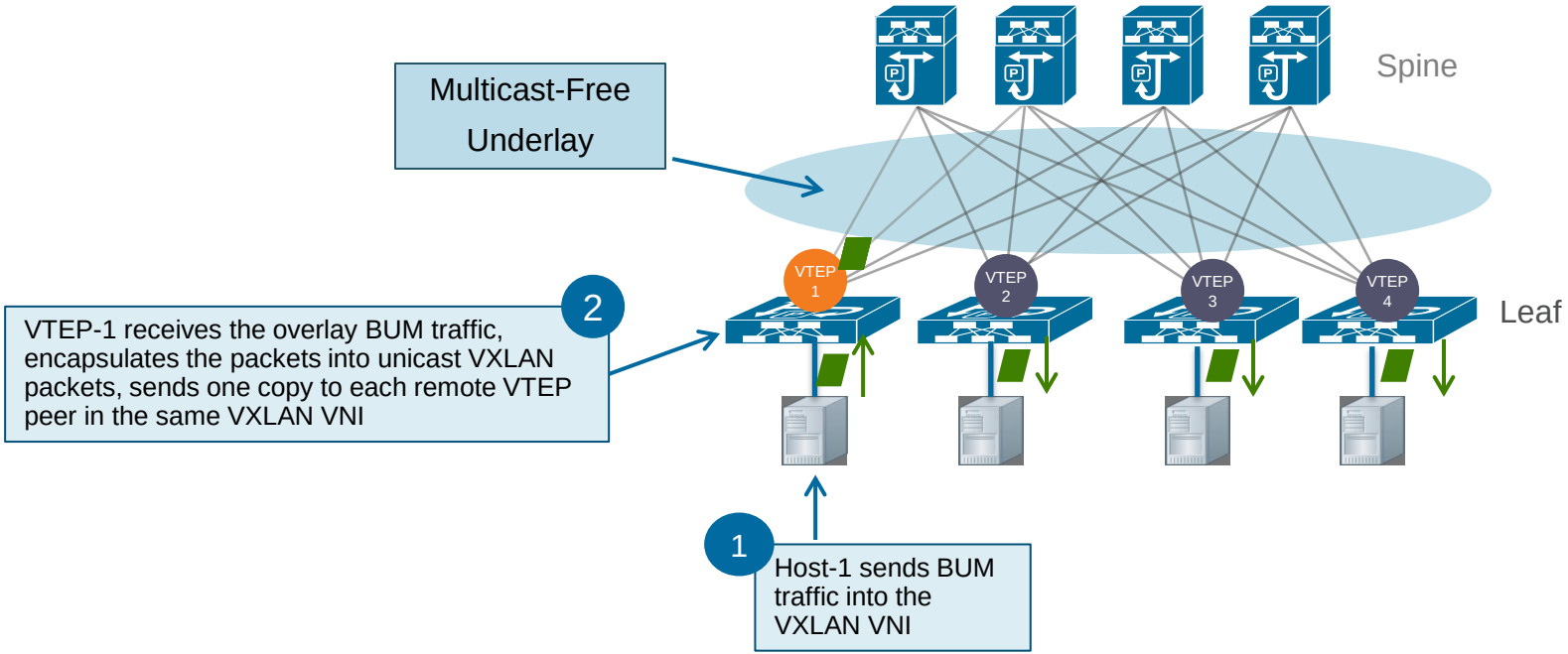

```
interface nve1
  no shutdown
  source-interface loopback0
  host-reachability protocol bgp
  member vni 20000
    suppress-arp
    mcast-group 239.1.1.1
  member vni 21000
    suppress-arp
    mcast-group 239.1.1.2
  member vni 39000 associate-vrf
  member vni 39010 associate-vrf
```

```
n9396-vtep-1.sakommu-lab.com# sh ip arp suppression topo-info

ARP L2RIB Topology information
Topo-id  ARP-suppression mode
100      L2 ARP Suppression
200      L2/L3 ARP Suppression
201      L2/L3 ARP Suppression
```

Head-end Replication

Head-end Replication (aka. Ingress replication):

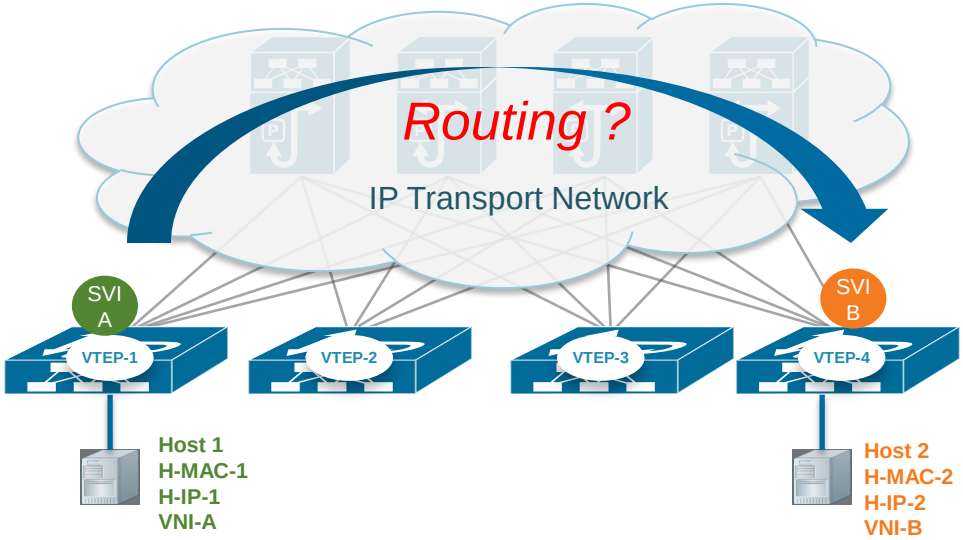
Eliminate the need for underlay multicast to transport overlay BUM traffic



Different integrated Route/Bridge (IRB) Modes

VXLAN Routing

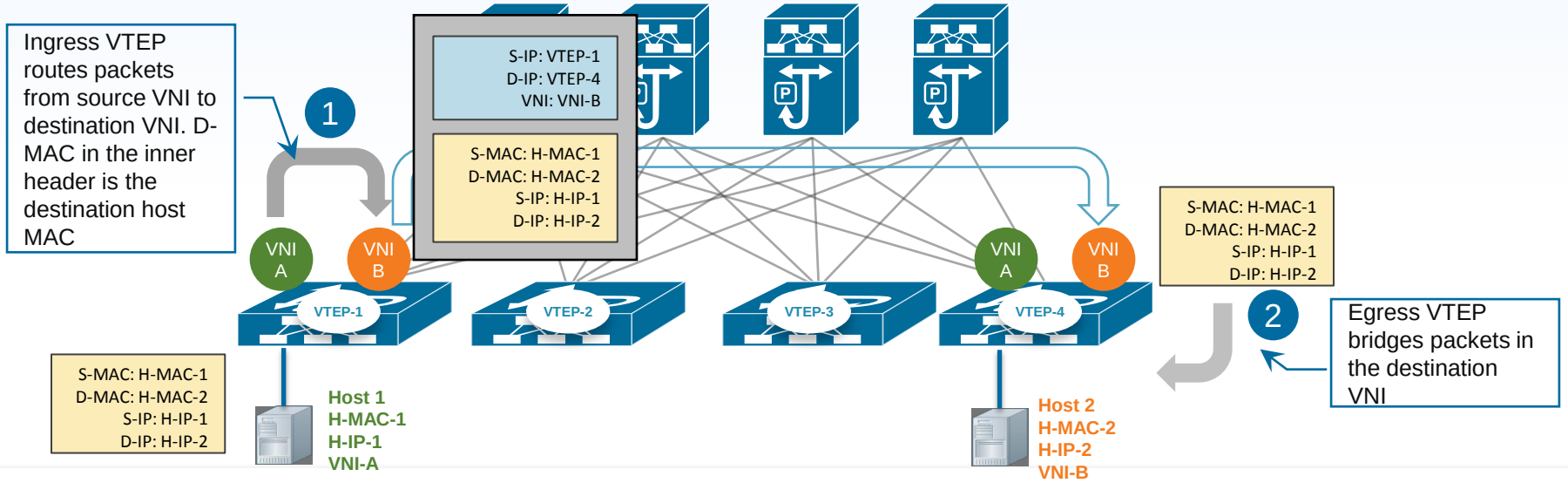
- Overlay Networks do follow two slightly different integrated Route/Bridge (IRB) semantics
- Asymmetric
  - Uses different “path” from Source to Destination and back
- Symmetric
  - Uses same “path” from Source to Destination and back
- Cisco follows Symmetric IRB



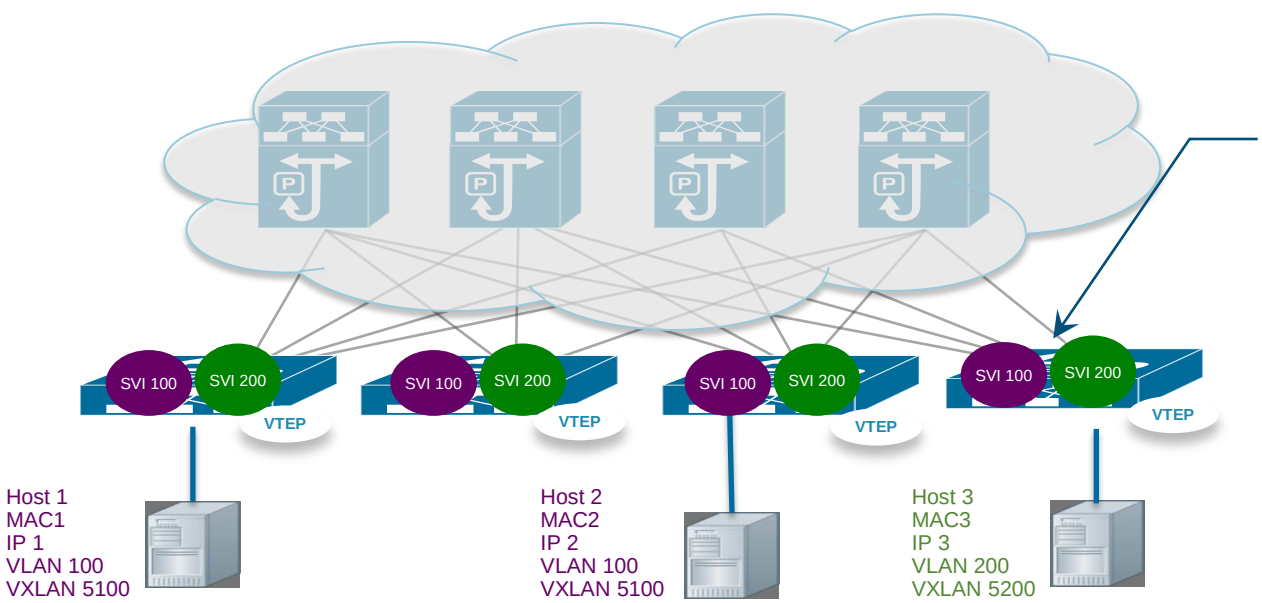
Asymmetric IRB (Cont'ed)

Asymmetric

- Routing and Bridging on the ingress VTEP
- Bridging on the egress VTEP
- Both source and destination VNIs need to reside on the ingress VTEP



VTEP VNI Membership Asymmetric IRB



Every VTEP needs to be in all VNIs

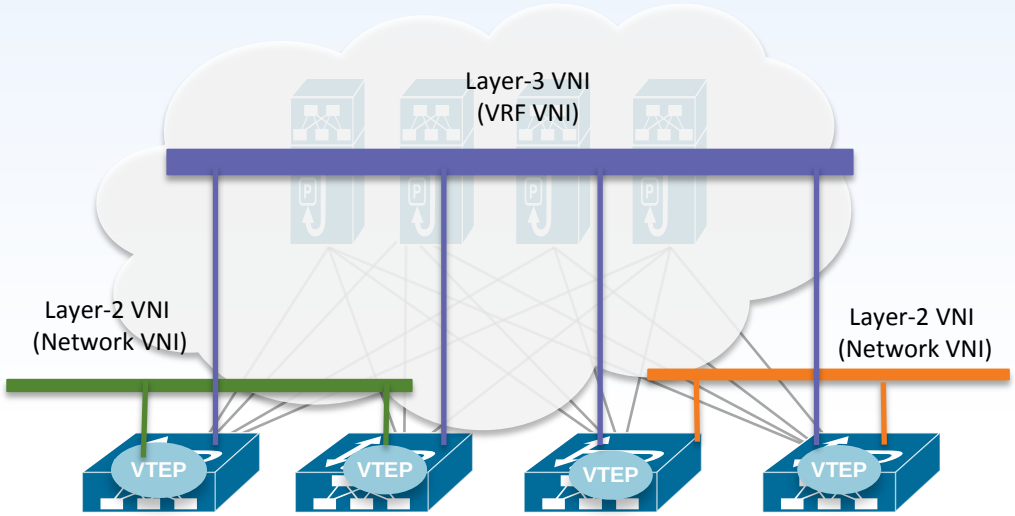
Every VTEP needs to maintain MAC tables for all VNIs, including those they don't have local hosts for.

1. All VTEPs in a VNI can be the virtual IP gateway for the local hosts
2. Optimized south-north bound forwarding for routed traffic without hair-pinning

Symmetric IRB

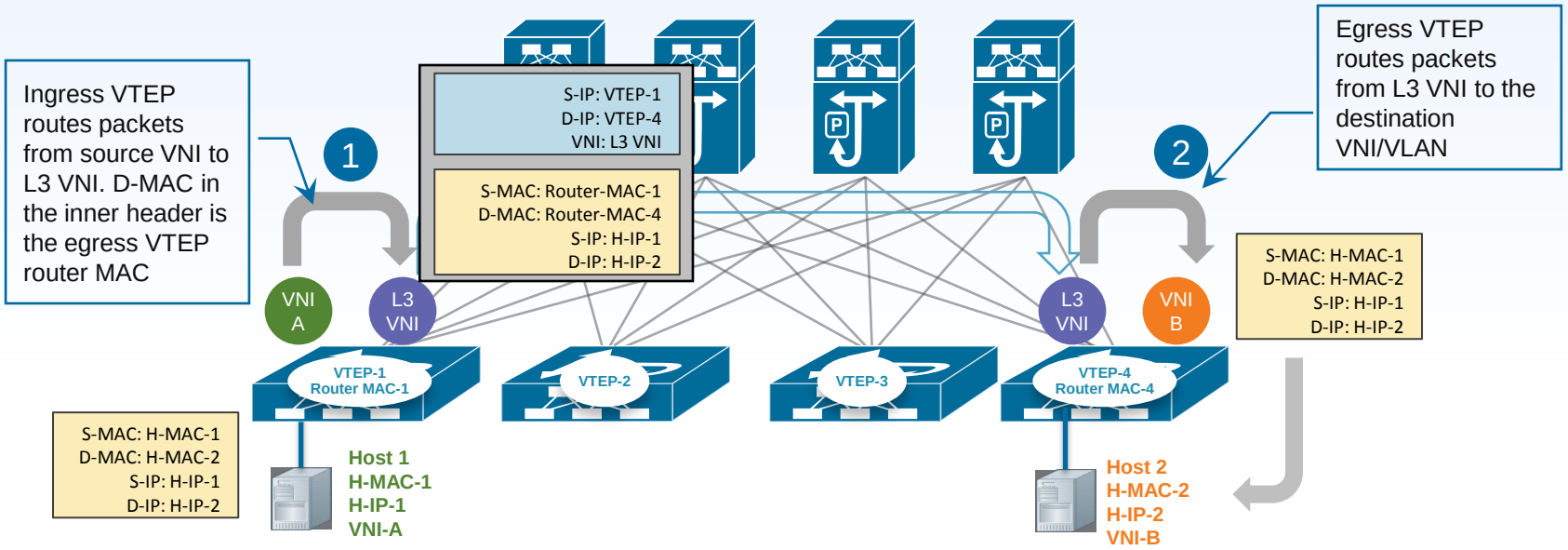
- Routing on both ingress and egress VTEPs
- Layer-3 VNI
  - Tenant VPN indicator
  - One per tenant VRF
- VTEP Router MAC
- Ingress VTEP routes packets onto the Layer-3 VNI
- Egress VTEP routes packets to the destination Layer-2 VNI

EVPN Multi-Tenancy and VNI Types (Cont'ed)

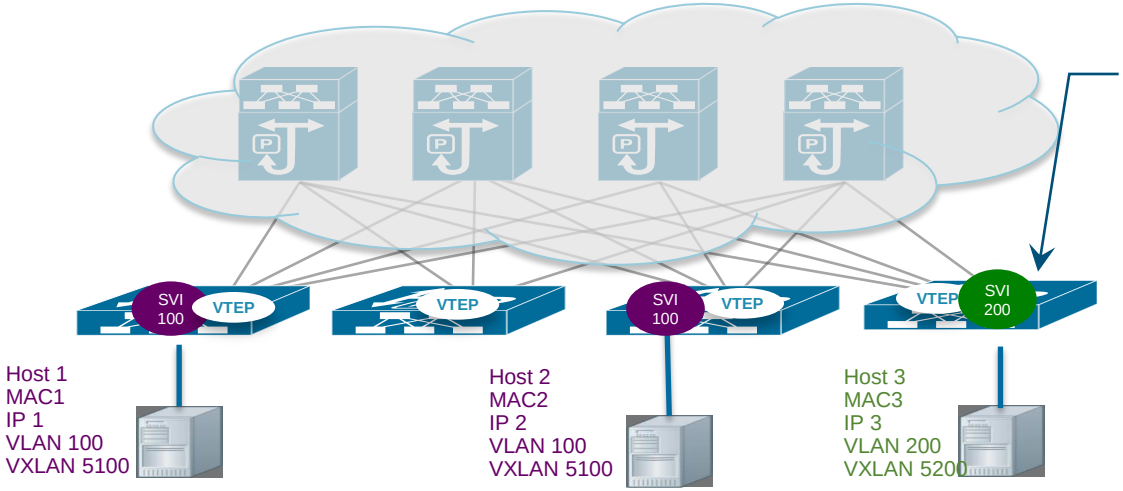



```
vlan 200
  vn-segment 20000
vlan 201
  vn-segment 20100

vlan 3900
  name l3-vni-vlan-for-tenant-1
  vn-segment 39000

interface Vlan3900
  description l3-vni-for-tenant-1-routing
  no shutdown
  vrf member evpn-tenant-1
  ip address 39.0.0.1/16
  fabric forwarding mode anycast-gateway

vrf context evpn-tenant-1
  vni 39000
  rd auto
  address-family ipv4 unicast
    route-target import 39000:39000
    route-target export 39000:39000
    route-target both auto evpn

interface Vlan200
  no shutdown
  vrf member evpn-tenant-1
  ip address 20.0.0.1/24
  fabric forwarding mode anycast-gateway

interface Vlan201
  no shutdown
  vrf member evpn-tenant-1
  ip address 20.1.0.1/24
  fabric forwarding mode anycast-gateway
```

Symmetric IRB (Cont'ed)



VTEP VNI Membership Symmetric IRB

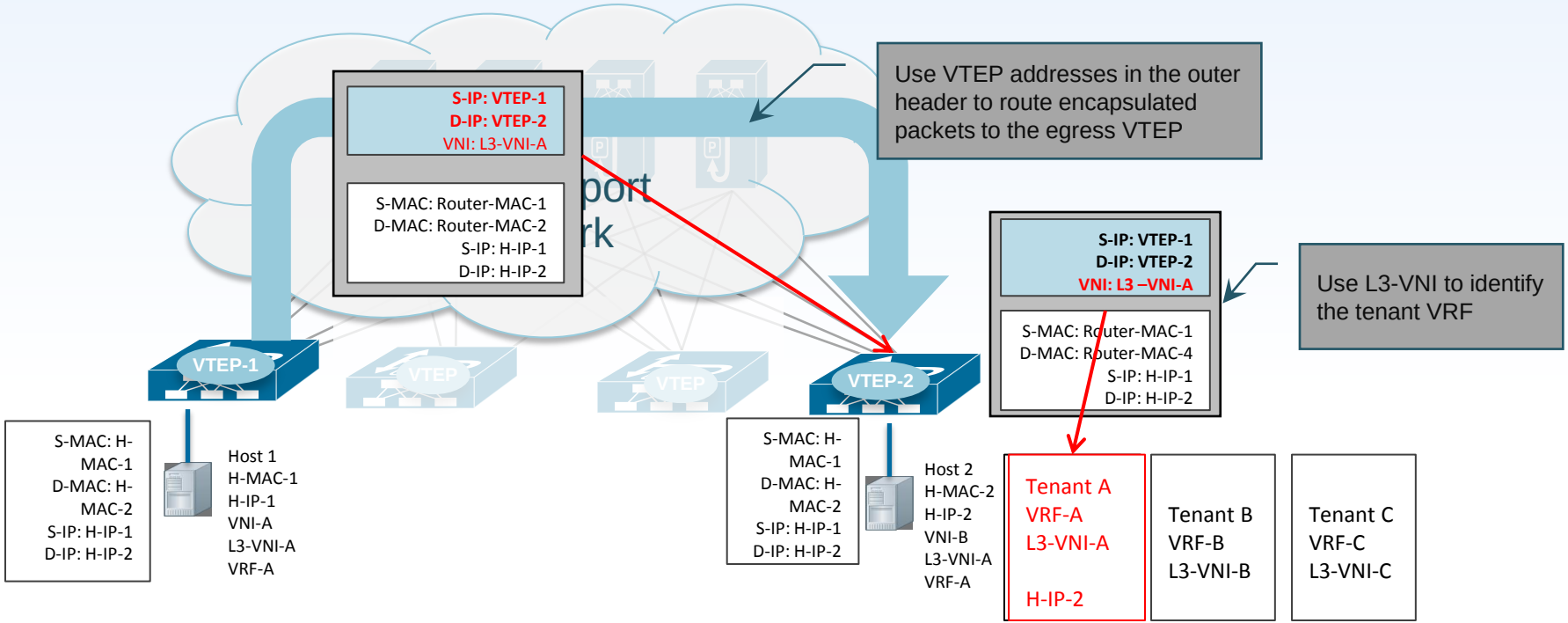


Every VTEP only needs to be in VNIs that it has local hosts for.

VTEPs don't need to maintain MAC tables for VNIs that they don't have local hosts for.

1. Optimal utilization of ARP and MAC tables
2. A VTEP only needs to be in the VNIs which it has local hosts for.

Multi-tenant Packet Forwarding in Symmetric IRB



Symmetric IRB vs Asymmetric IRB

- Symmetric IRB has optimal utilization of ARP and MAC tables on a VTEP
- Symmetric IRB scales better for end hosts
- Symmetric IRB scales better in terms of the total number of VNIs a VXLAN overlay network can support

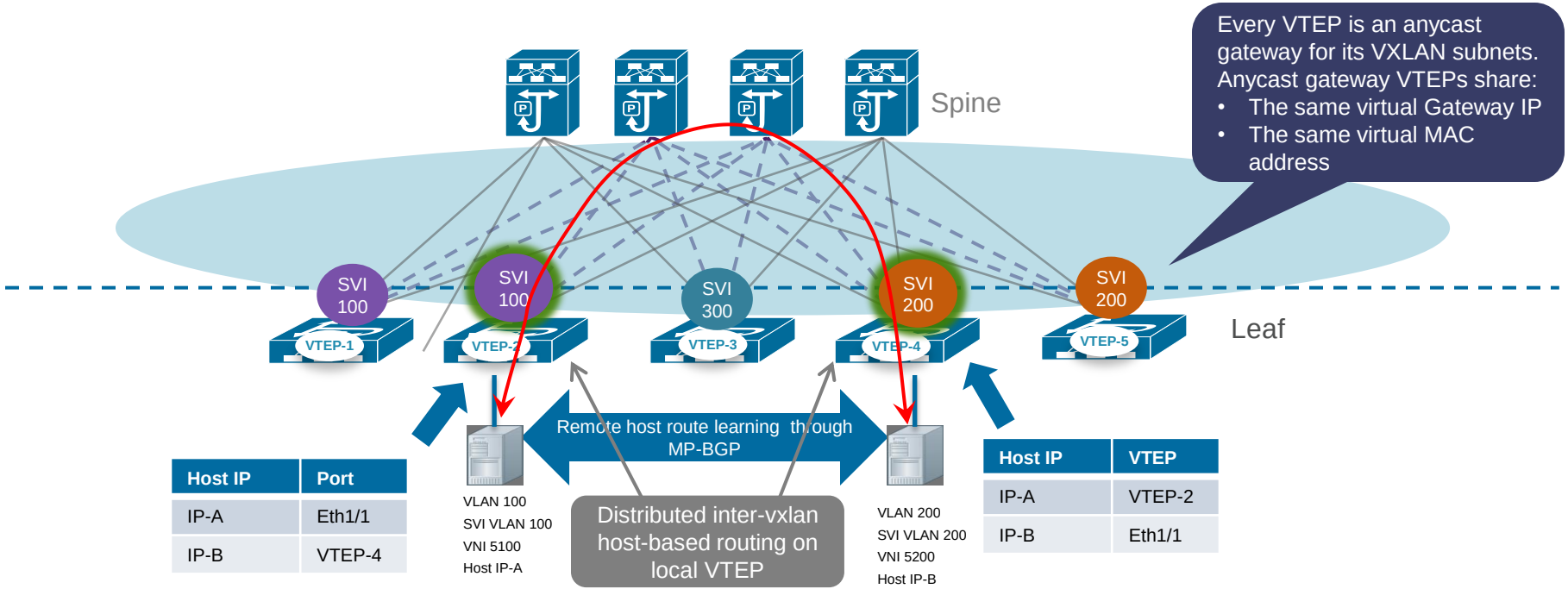
Multi-vendor interoperability:

- Some vendors implemented Asymmetric IRB
- It’s been agreed upon among multiple vendors that Symmetric IRB is the ultimate solution
- Cisco implemented Symmetric IRB
- Cisco will introduce backward compatability with asymmetric IRB by adding the support for it.

Optimal VXLAN Routing with Symmetric IRB and Anycast Gateway

Host-based fabric routing and bridging

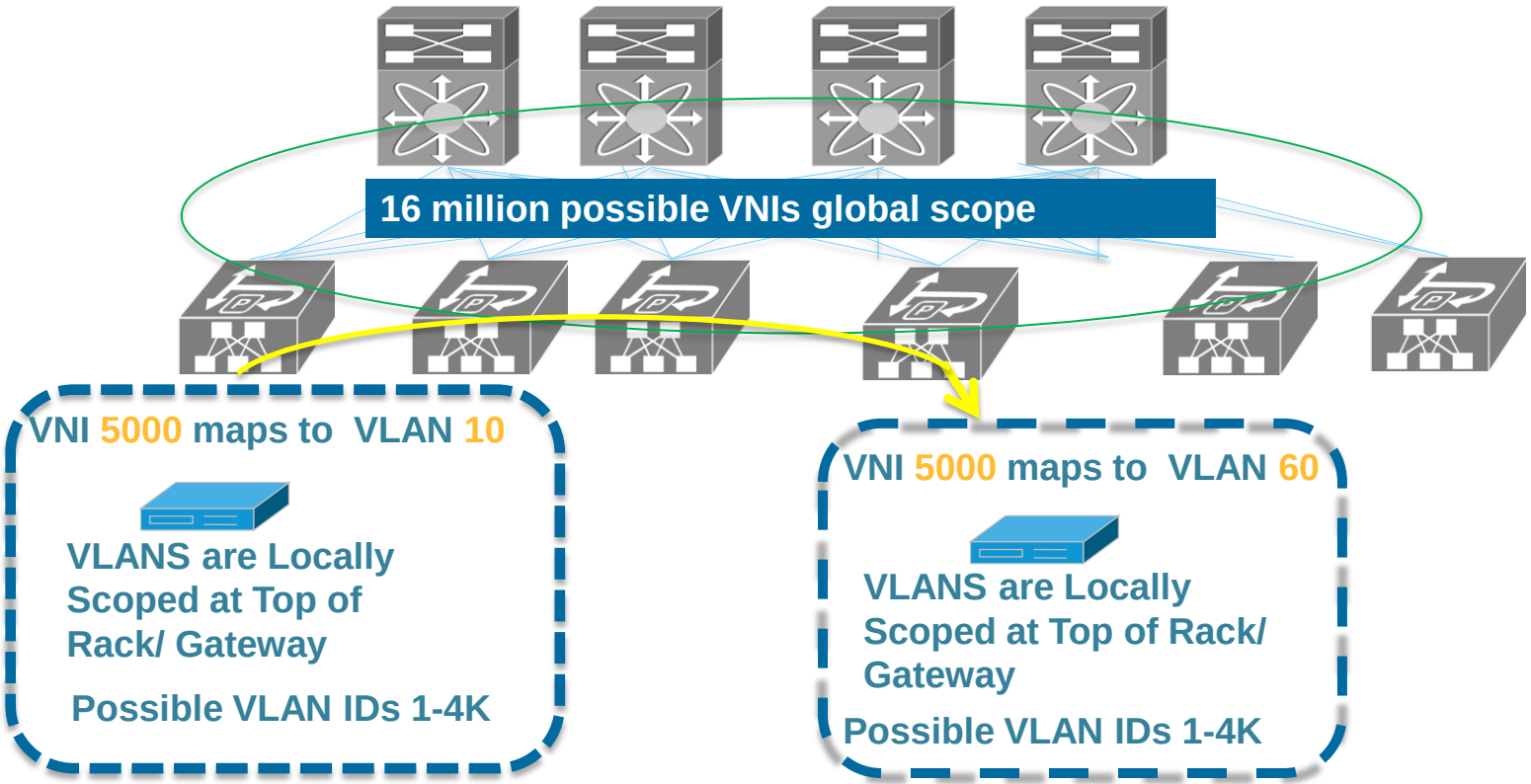
with optimal and flexible VXLAN VNI placement



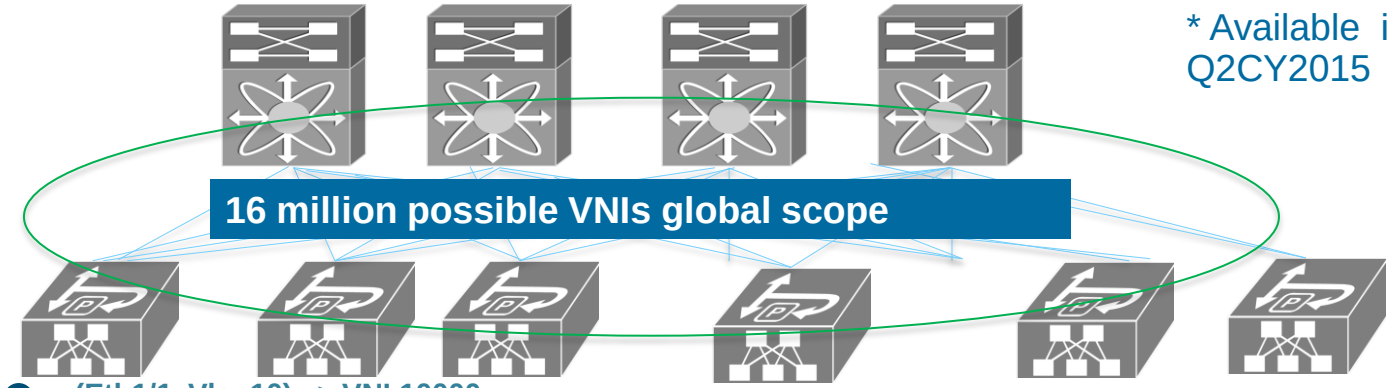
Host IP	Port
IP-A	Eth1/1
IP-B	VTEP-4

Host IP	VTEP
IP-A	VTEP-2
IP-B	Eth1/1

Local Scoping of VLANs –ToR Local



Local Scoping of VLANs – Port Local*



* Available in Q2CY2015

(Eth1/1, Vlan10) => VNI 10000
(Eth1/2, Vlan10) => VNI 10001
(Eth1/2, Vlan11) => VNI 10000

VNI 5000 maps to (E1/1, VLAN 10)

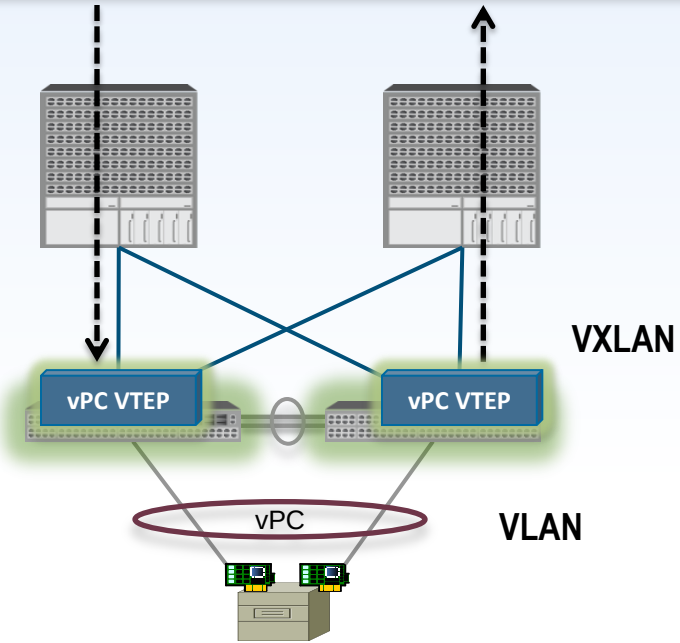
VLANS are Locally Scoped
VLAN to VNI mapping is per-port significant
Possible VLAN IDs 1-4K

VNI 5000 maps to (E1/2, VLAN 60

VLANS are Locally Scoped
VLAN to VNI mapping is per-port significant
Possible VLAN IDs 1-4K

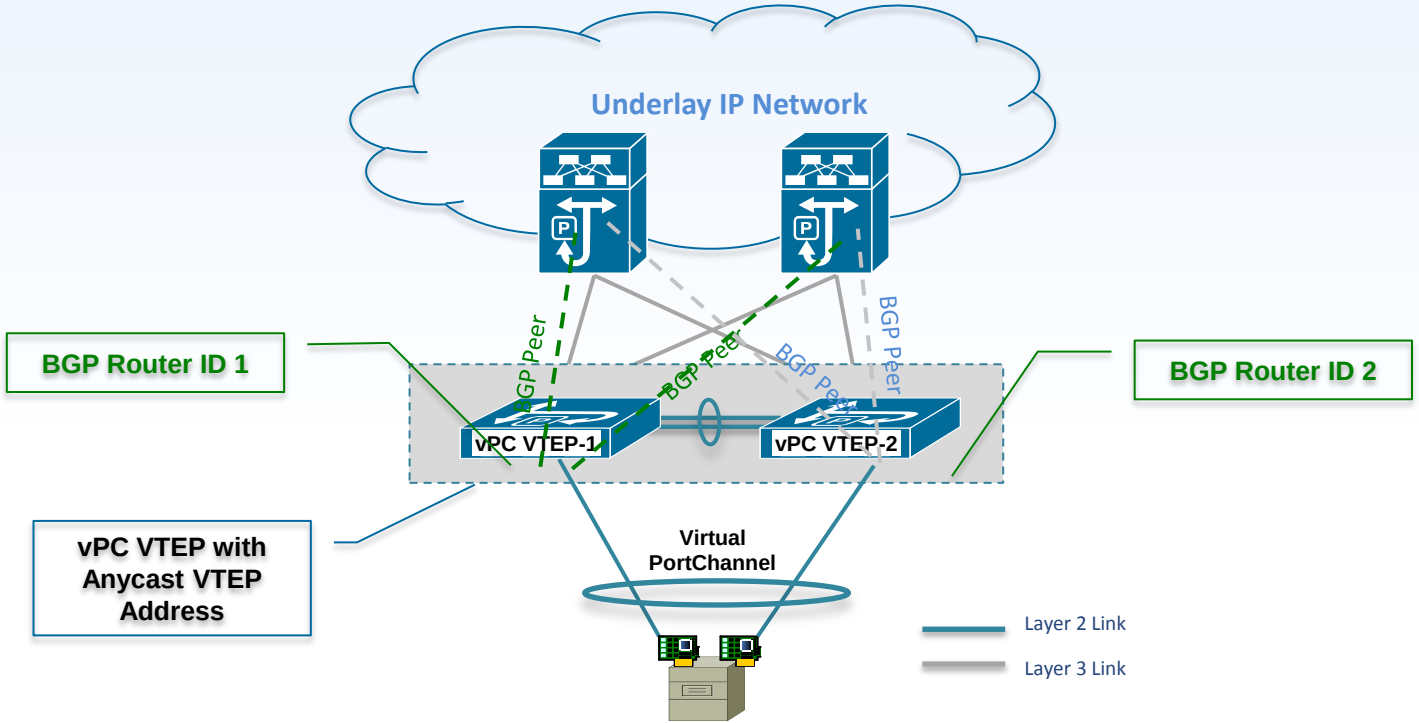
vPC VTEP for VXLAN Bridging and Routing

- When vPC is enabled an ‘anycast’ VTEP address is programmed on both vPC peers
- Symmetrical forwarding behavior on both peers provides
- Multicast topology prevents BUM traffic being sent to the same IP address across the L3 network (prevents duplication of flooded packets)
- vPC peer-gateway feature must be enabled on both peers
- VXLAN header is ‘not’ carried on the vPC Peer link (MCT link)




```
interface loopback0
  ip address 10.1.1.13/32
  ip address 10.1.1.134/32 secondary
```

vPC VTEPs in MP-BGP EVPN



EVPN Control Plane Advantages

A multi-tenant fabric solution with host-based forwarding

- Industry standard protocol for multi-vendor interoperability
- Build-in multi-tenancy support
  - Leverage MP-BGP to deliver VXLAN with L3VPN characteristics
- Truly scalable with protocol-driven learning
  - Host MAC/IP address advertisement through EVPN MP-BGP
- Fast convergence upon host movements or network failures
  - MP-BGP protocol driven re-learning and convergence
  - Upon host movement, the new VTEP will send out a BGP update to advertise the new location of the host

EVPN Control Plane Advantages (Cont'ed)

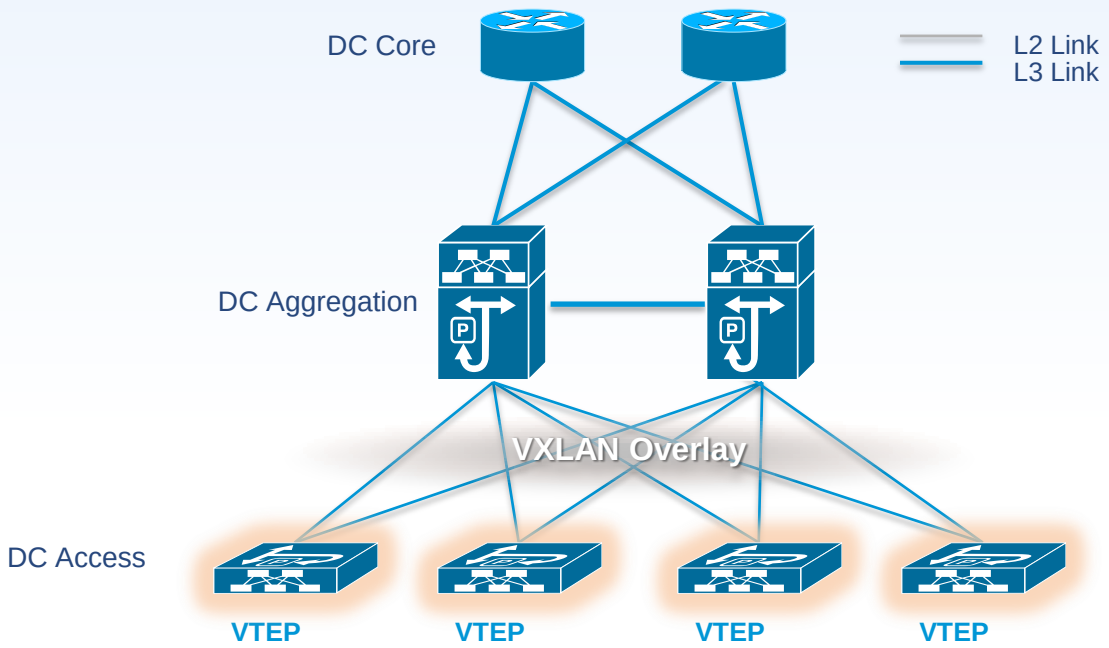
A multi-tenant fabric solution with host-based forwarding

- Optimal traffic forwarding supporting host mobility
  - Anycast IP gateway for optimal forwarding for host generated traffic
  - No need for hair-pinning to to reach the IP gateway
- ARP suppression
  - Minimize ARP flooding in overlay
- Head-end Replication with dynamically learned remote-VTEP list
  - Head-end replication enables multicast-free underlay network
  - Dynamically learned remote-VTEP list minimizes the operational overhead of head-end replication
- VTEP peer authentication via MP-BGP authentication
  - Added security to prevent rogue VTEPs or VTEP spoofing

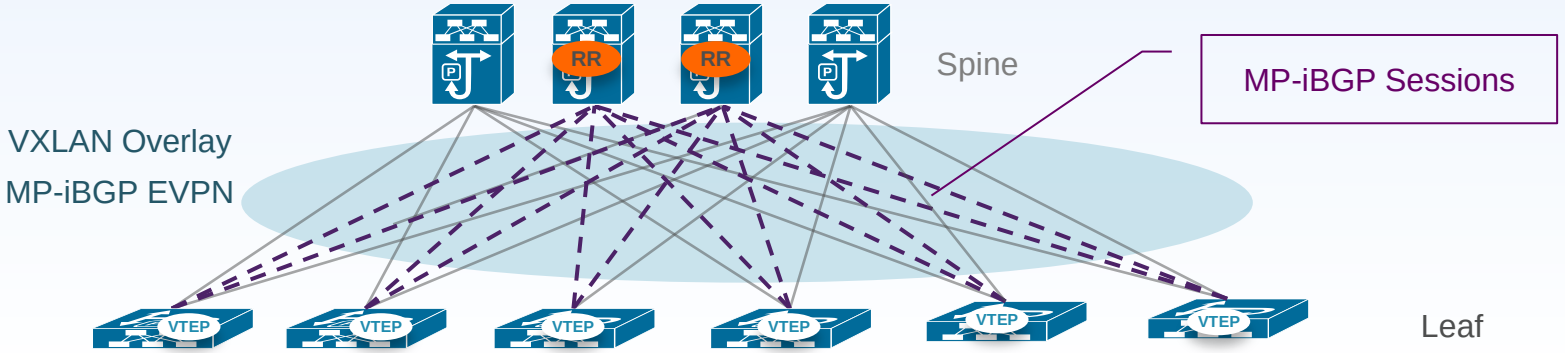
Agenda

- VxLAN Overview
- MP-BGP EVPN Basics
- MP-BGP EVPN Control Plane
- VxLAN Design Options
- MP-BGP EVPN VXLAN Configuration
- VxLAN Capability on Nexus 9000 Series Switches

VXLAN in 3-Tier Network

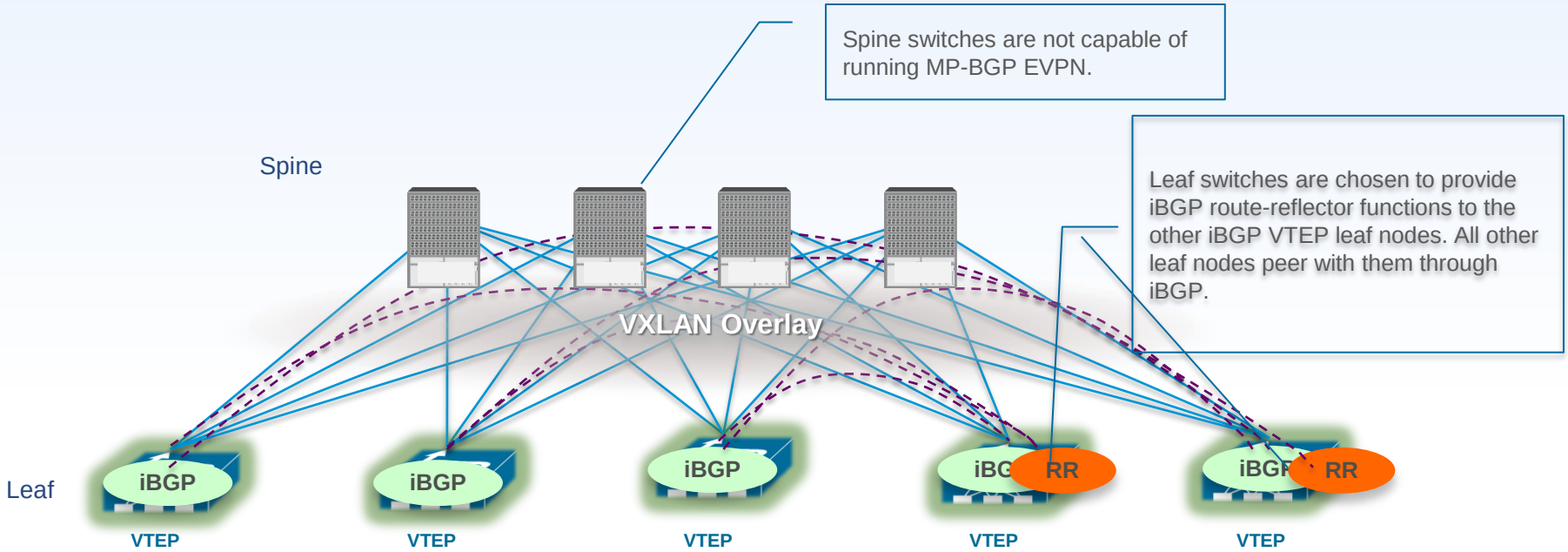


VXLAN Fabric Design with MP-iBGP EVPN

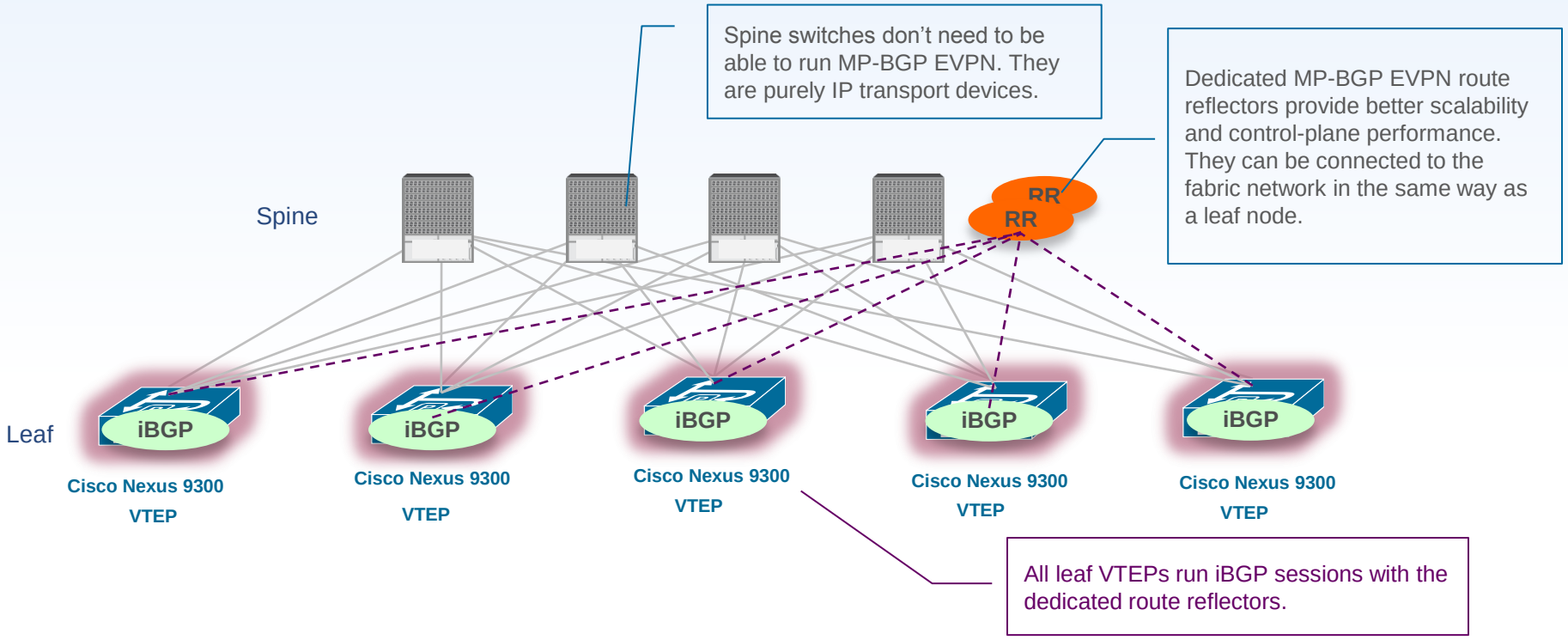


- VTEP Functions are on leaf layer
- Spine nodes are iBGP route reflector
- Spine nodes don’t need to be VTEP

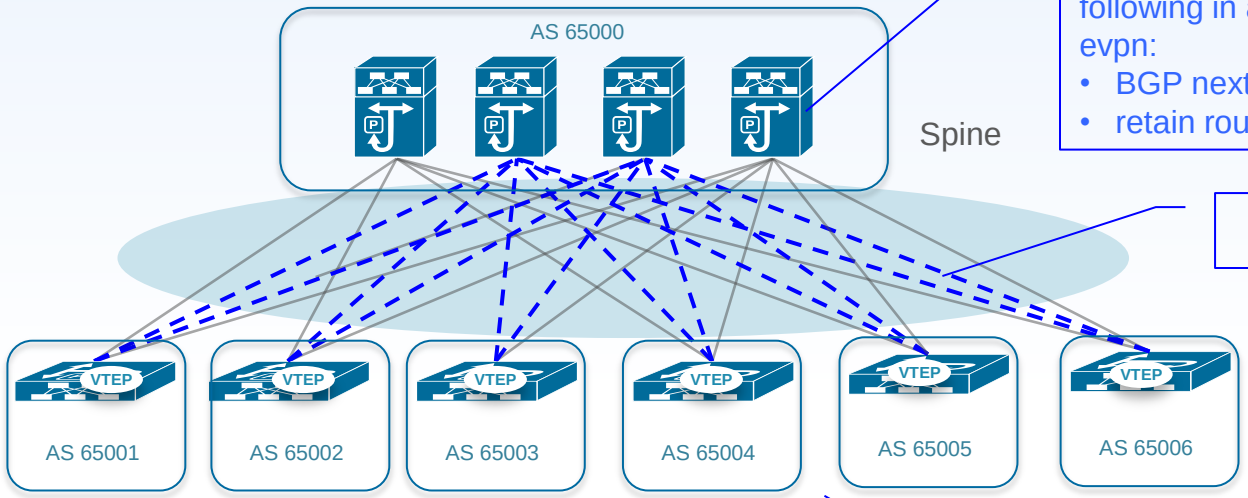
VXLAN EVPN Fabric with MP-iBGP Design (Cont'ed)



VXLAN EVPN Fabric with MP-iBGP Design (Cont'ed)



VXLAN Fabric Design with MP-eBGP EVPN



BGP on Spine needs to have the following in address-family l2vpn evpn:

- BGP next-hop unchanged
- retain route-target all

MP-eBGP Sessions

- VTEP Functions are on leaf layer
- Spine nodes are MP-eBGP Peers
- Spine nodes don't need to be VTEP

Need to manually configure Route-targets on each VTEP

VXLAN Fabric Design with MP-eBGP EVPN (Cont'ed)

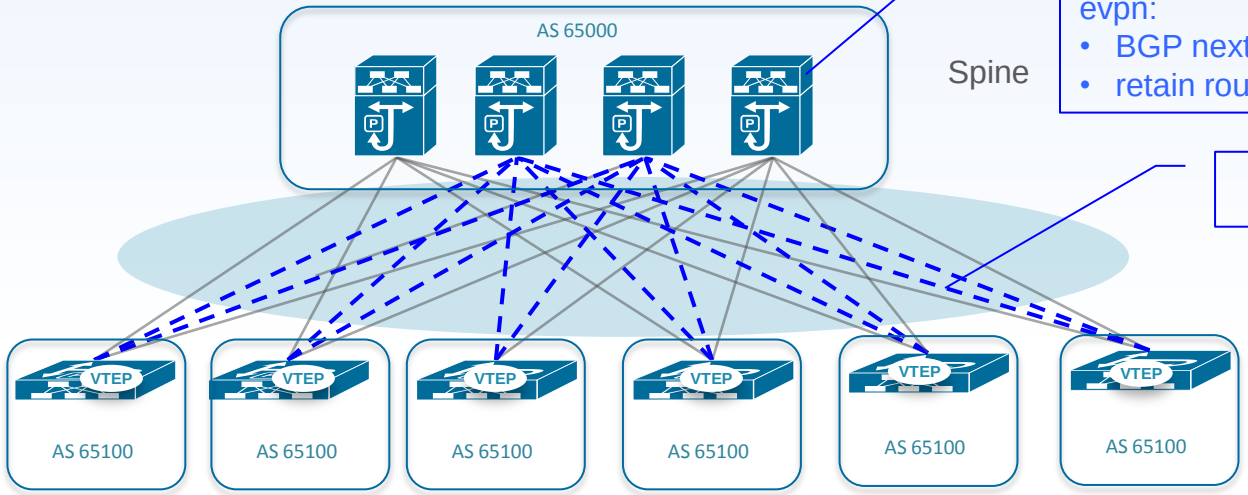
```
[BGP configuration on a spine switch as in Figure 16 design]
route-map permit-all permit 10
  set ip next-hop unchanged

router bgp 65000
  router-id 10.1.1.1
  address-family ipv4 unicast
    redistribute direct route-map permitall
  address-family l2vpn evpn
    nexthop route-map permit-all
    retain route-target all
  neighbor 192.167.11.2 remote-as 65001
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
      route-map permit-all out
  neighbor 192.168.12.2 remote-as 65002
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
      route-map permit-all out
```

Set next-hop policy to not change the next-hop attributes.

Retain routes with all route targets when advertising the EVPN BGP routes to eBGP peers.

Set outbound policy to advertise all routes to this eBGP neighbor.

VXLAN Fabric Design with MP-eBGP EVPN (Cont'ed)



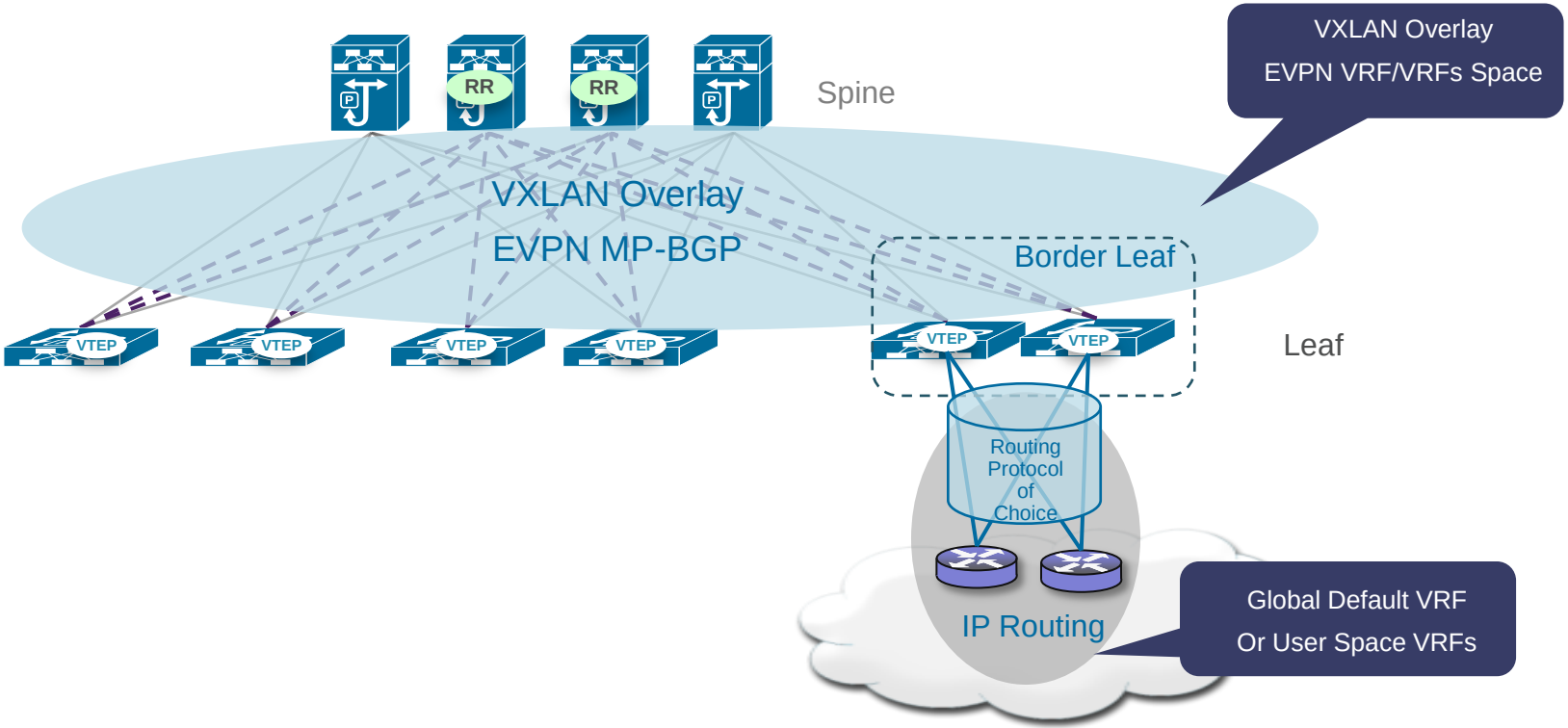
BGP on Spine needs to have the following in address-family l2vpn evpn:

- BGP next-hop unchanged
- retain route-target all

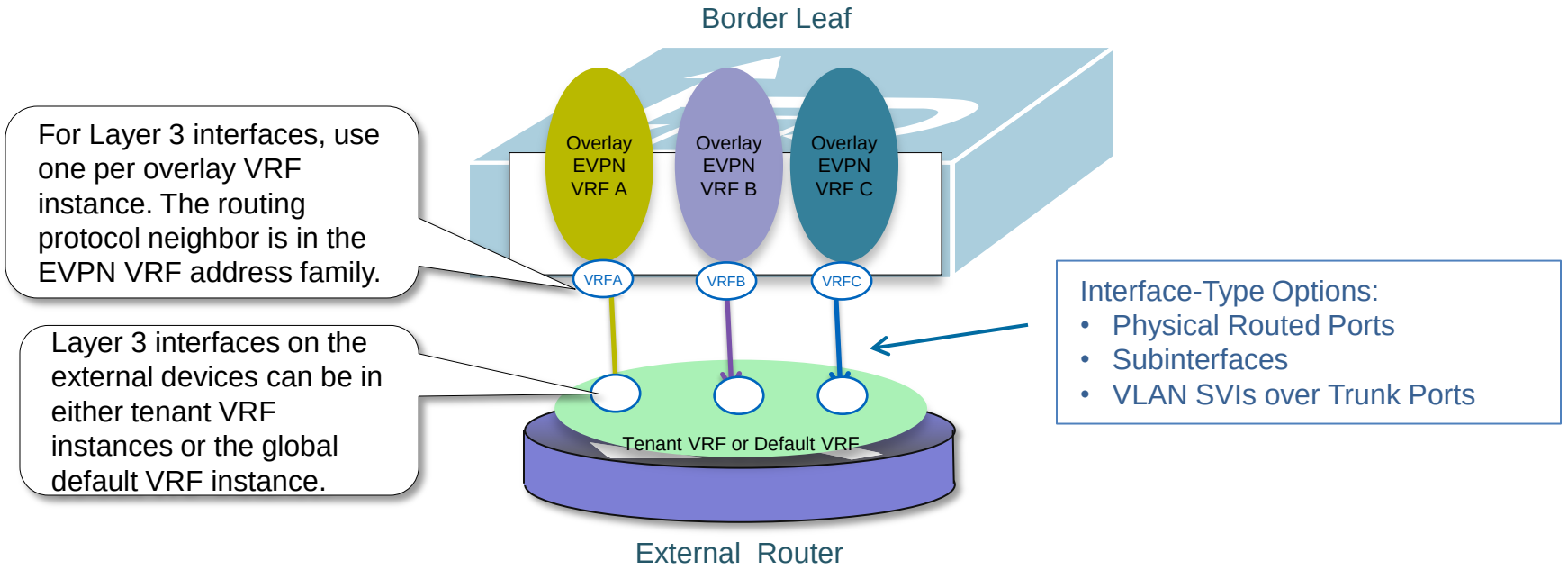
MP-eBGP Sessions

VTEP leafs are in the same BGP AS

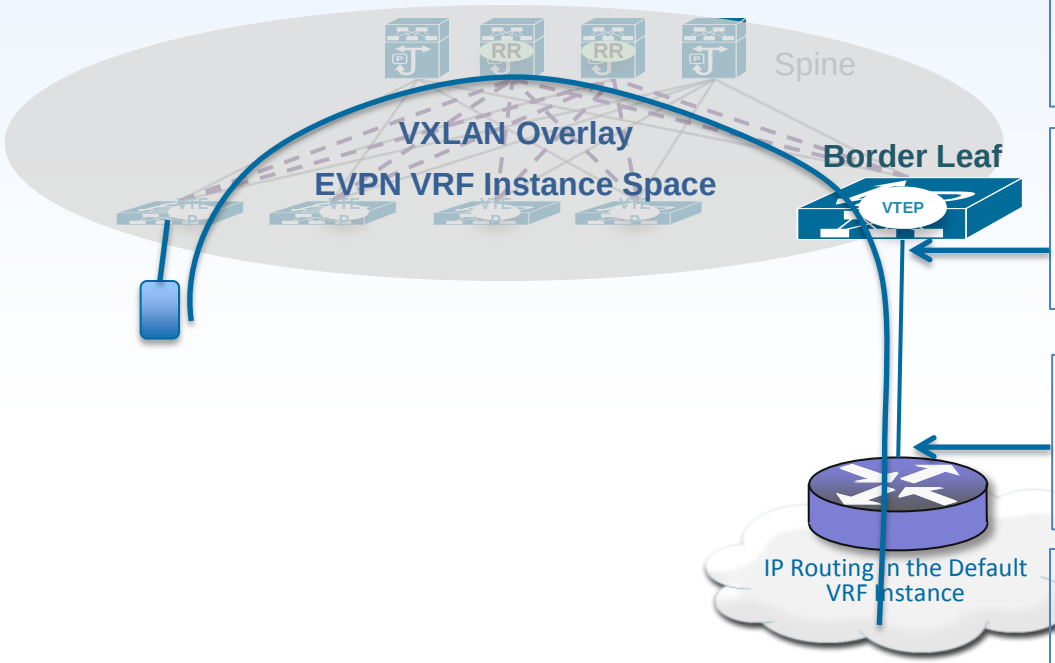
EVPN VXLAN Fabric External Routing



EVPN VXLAN Fabric External Routing (Cont'ed)



EVPN VXLAN External Routing with BGP Sample Configuration




```
Router bgp 100
 vrf evpn-tenant-1
    address-family ipv4 unicast
      network 20.0.0.0/24
    neighbor 30.10.1.2 remote-as 200
      address-family ipv4 unicast
       prefix-list outbound-no-hosts out
```

```
interface Ethernet2/9.10
  mtu 9216
  encapsulation dot1q 10
  vrf member evpn-tenant-1
  ip address 30.10.1.1/30
```

```
interface Ethernet1/50.10
  mtu 9216
  encapsulation dot1q 10
  ip address 30.10.1.2/30
```

```
router bgp 200
 address-family ipv4 unicast
    network 100.0.0.0/24
    network 100.0.1.0/24
  neighbor 30.10.1.1 remote-as 100
    address-family ipv4 unicast
```

EVPN VXLAN External Routing with BGP Sample Configuration – On the Border Leaf

On the VXLAN Border Leaf

```
router bgp 100
  router-id 10.1.1.16
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  neighbor 10.1.1.2 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  vrf evpn-tenant-1
    address-family ipv4 unicast
      network 20.0.0.0/24
    neighbor 30.10.1.2 remote-as 200
      address-family ipv4 unicast
       prefix-list outbound-no-hosts out

ip prefix-list outbound-no-hosts seq 5 deny 0.0.0.0/0 eq 32
ip prefix-list outbound-no-hosts seq 10 permit 0.0.0.0/0 le 32
```

The eBGP neighbor is on the outside.
It's in address-family ipv4 unicast of the tenant VRF routing instance.

For better scalability, apply prefix-list to filter out /32 IP host routes. Advertise prefix routes only to the external eBGP neighbor.

EVPN VXLAN External Routing with BGP

```
n9396-vtep-1# sh ip bgp vrf evpn-tenant-1 100.0.0.0
BGP routing table information for VRF evpn-tenant-1, address family IPv4 Unicast
BGP routing table entry for 100.0.0.0/24, version 70
Paths: (1 available, best #1)
Flags: (0x08041a) on xmit-list, is in urib, is best urib route
  vpn: version 75, (0x100002) on xmit-list

  Advertised path-id 1, VPN AF advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop
             Imported from unknown dest
  AS-Path: NONE, path sourced internal to AS
    10.1.1.16 (metric 3) from 10.1.1.1 (10.1.1.1)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 39000
      Extcommunity:  RT:100:39000 ENCAP:8 Router MAC:6412
      Originator: 10.1.1.16 Cluster list: 10.1.1.1

  VRF advertise information:
  Path-id 1 not advertised to any peer

  VPN AF advertise information:
  Path-id 1 not advertised to any peer
n9396-vtep-1#
n9396-vtep-1# sh ip route vrf evpn-tenant-1 100.0.0.0/24
IP Route Table for VRF "evpn-tenant-1"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>
```

This is the external route.

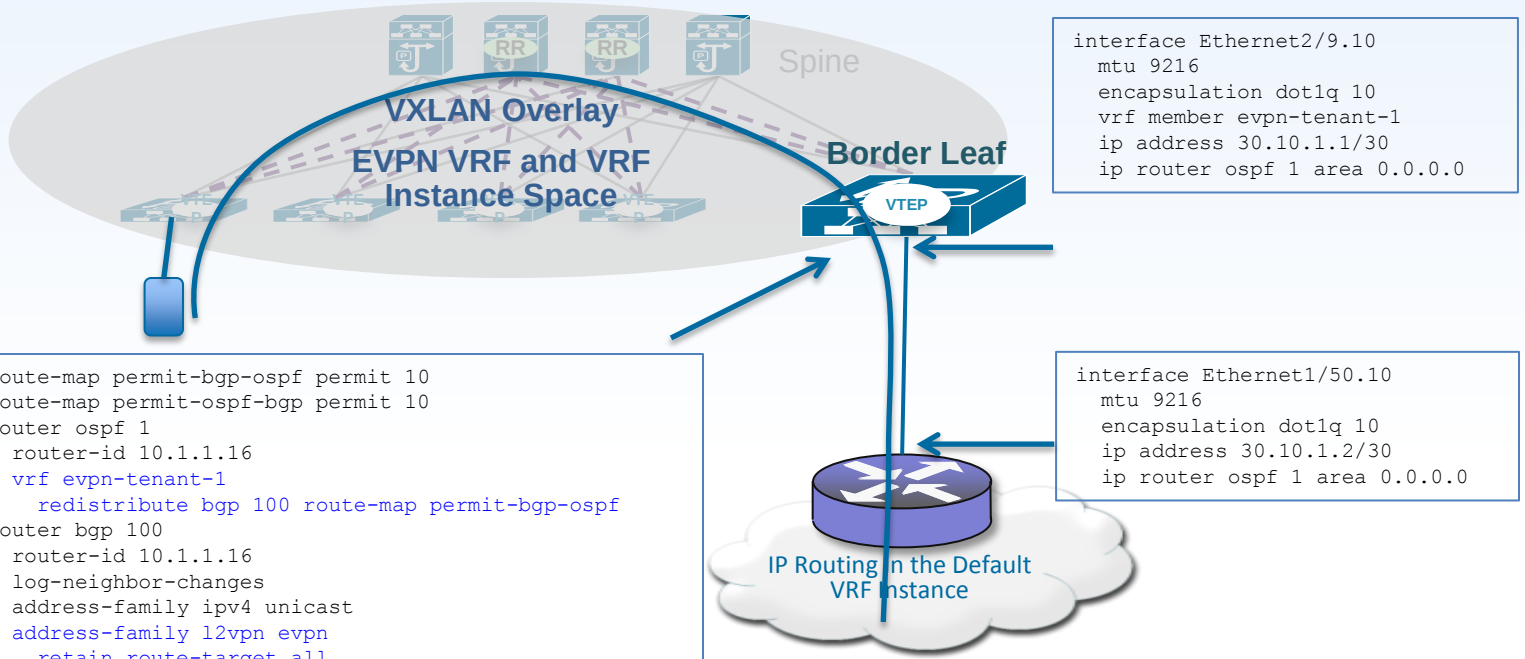
The next hop is the VTEP address of the border lea

The tenant is VRF L3 VNI.

10.1.1.16 is the BGP router ID of the border leaf. 10.1.1.1 is the spine route reflector.

This is the iBGP route. The next hop is the VTEP address of the border leaf.

EVPN VXLAN External Routing with OSPF Sample Configuration




```
interface Ethernet2/9.10
  mtu 9216
  encapsulation dot1q 10
  vrf member evpn-tenant-1
  ip address 30.10.1.1/30
  ip router ospf 1 area 0.0.0.0
```

```
route-map permit-bgp-ospf permit 10
route-map permit-ospf-bgp permit 10
router ospf 1
  router-id 10.1.1.16
  vrf evpn-tenant-1
    redistribute bgp 100 route-map permit-bgp-ospf
router bgp 100
  router-id 10.1.1.16
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
    retain route-target all
 vrf evpn-tenant-1
    address-family ipv4 unicast
      advertise l2vpn evpn
      redistribute ospf 1 route-map permit-ospf-bgp
```

```
interface Ethernet1/50.10
  mtu 9216
  encapsulation dot1q 10
  ip address 30.10.1.2/30
  ip router ospf 1 area 0.0.0.0
```

EVPN VXLAN External Routing with OSPF
Sample Configuration - The relevant configuration on the border leaf

```
ip prefix-list bgp-ospf-no-hosts seq 5 permit 0.0.0.0/0 eq 32
route-map permit-bgp-ospf deny 5
  match ip address prefix-list bgp-ospf-no-hosts
route-map permit-bgp-ospf permit 10
route-map permit-ospf-bgp permit 10

router ospf 1
  router-id 10.1.1.16
  vrf evpn-tenant-1
    redistribute bgp 100 route-map permit-bgp-ospf

router bgp 100
  router-id 10.1.1.16
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
    retain route-target all
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  neighbor 10.1.1.2 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
```

Redistribute BGP routes to OSPF. Filter out /32 IP host routes.

A BGP router will modify route targets in l2vpn evpn routes when it is an autonomous system boundary router. The original route target must be retained.

Redistribute OSPF to BGP. Advertise the redistributed routes to L2VPN EVPN.

EVPN VXLAN External Routing with OSPF
The internal VTEPs learn the external routes through MP-BGP EVPN

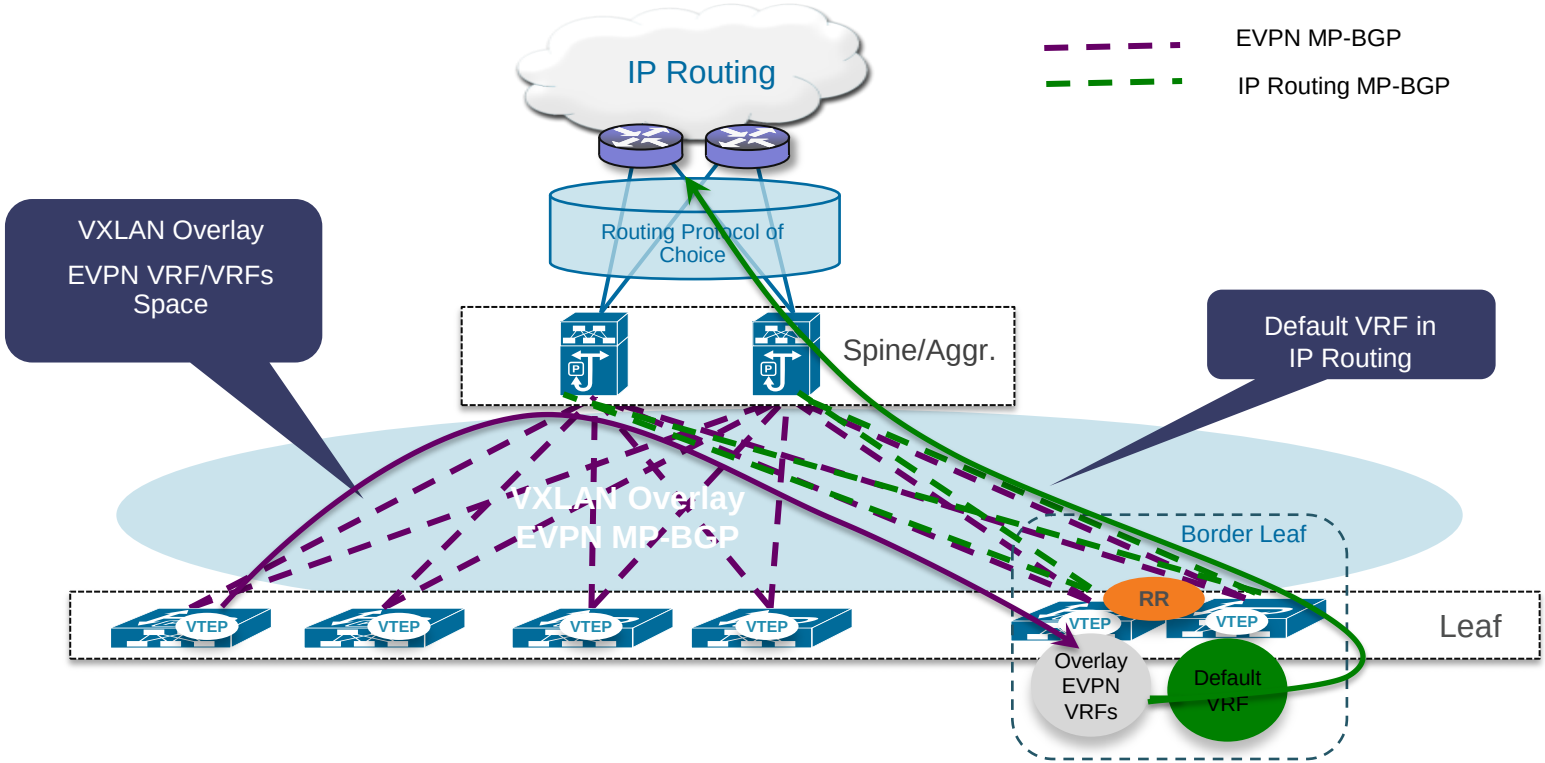
```
n9396-vtep-1# sh vrf evpn-tenant-1 detail
VRF-Name: evpn-tenant-1, VRF-ID: 3, State: Up
    VPNID: unknown
    RD: 10.1.1.11:3
    VNI: 39000
    Max Routes: 0  Mid-Threshold: 0
    Table-ID: 0x80000003, AF: IPv6, Fwd-ID: 0x80000003, State: Up
    Table-ID: 0x00000003, AF: IPv4, Fwd-ID: 0x00000003, State: Up

n9396-vtep-1# sh bgp l2vpn evpn rd 10.1.1.11:3 100.0.0.0
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 10.1.1.11:3    (L3VNI 39000)
BGP routing table entry for [5]:[0]:[0]:[24]:[100.0.0.0]:[0.0.0.0]/224, version 396
Paths: (1 available, best #1)
Flags: (0x00001a) on xmit-list, is in l2rib/evpn

  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop
             Imported from 10.1.1.16:3:[5]:[0]:[0]:[24]:[100.0.0.0]:[0.0.0.0]/120
  AS-Path: NONE, path sourced internal to AS
    10.1.1.16 (metric 3) from 10.1.1.1 (10.1.1.1)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 39000
      Extcommunity:  RT:100:39000 ENCAP:8 Router MAC:6412.2574.6ae7
      Originator: 10.1.1.16 Cluster list: 10.1.1.1

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

The external route learned through MP-BGP EVPN is imported into the tenant VRF.

The next hop is the VTEP address of the border leaf.

This is the Layer 3 VNI of the tenant VRF routing instance.

Alternative Design for EVPN VXLAN External Routing

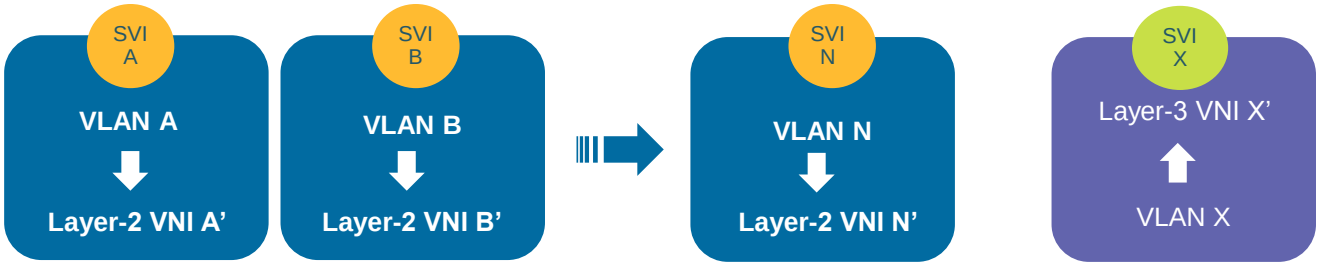


Agenda

- VxLAN Overview
- MP-BGP EVPN Basics
- MP-BGP EVPN Control Plane
- VXLAN Design Options
- MP-BGP EVPN VXLAN Configuration
- VxLAN Capability on Nexus 9000 Series Switches

Logical Construct of Multi-Tenant VXLAN EVPN

Tenant A (VRF A)



- One VLAN maps to one Layer-2 VNI Layer-2 VNI per Layer-2 segment
- A Tenant can have multiple VLANs, therefore multiple Layer-2 VNIs
- Traffic within one Layer-2 VNI is bridged
- Traffic between Layer-2 VINs is routed

- 1 Layer-3 VNI per Tenant (VRF) for routing
- VNI X' is used for routed packets

Initial configuration – Per Switch

Enable VXLAN and MP-BGP EVPN Control Plane

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

- Enable VXLAN
- Enable VLAN-based VXLAN (the currently only mode)
- Enable BGP
- Enable EVPN control plane for VXLAN

Other features may need to be anabled

```
feature ospf
feature pim
feature interface-vlan
```

- Enable OSPF if it's chosen to be the underlay IGP routing protocol
- Enable IP PIM multicast routing in the underlay network
- Enable VLAN SVI interfaces if the VTEP needs to be IP gateway and route for the VXLAN VLAN IP subnet.

EVPN Tenant VRF

Create VXLAN tenant VRF

```
vrf context evpn-tenant-1
  vni 39000
  rd auto
  address-family ipv4 unicast
    route-target import 39000:39000
    route-target export 39000:39000
    route-target both auto evpn
```

```
vrf context evpn-tenant-2
  vni 39010
  rd auto
  address-family ipv4 unicast
    route-target import 39010:39010
    route-target export 39010:39010
    route-target both auto evpn
```

Create a VXLAN Tenant VRF

Specify the Layer-3 VNI for VXLAN routing within the tenant VRF

Define VRF RD (route distinguisher)

Define VRF Route Target and import/export policies in address-family ipv4 unicast

Example to create a 2nd tenant VRF following the above steps

Layer-3 VNI Per Tenant for EVPN Routing

Configure Layer-3 VNI per EVPN Tenant VRF Routing Instant

```
vlan 3900
  name l3-vni-vlan-for-tenant-1
  vn-segment 39000

interface Vlan3900
  description l3-vni-for-tenant-1-routing
  no shutdown
  vrf member evpn-tenant-1

vrf context evpn-tenant-1
  vni 39000
  rd auto
  address-family ipv4 unicast
    route-target import 39000:39000
    route-target export 39000:39000
    route-target both auto evpn
```

Create the VLAN for the Layer-3 VNI.
One Layer-3 VNI per tenant VRF routing instance

Create the SVI interface for the Layer-3 VNI
Put this SVI interface into the tenant VRF context

Associate the Layer-3 VNI with the tenant VRF routing instance.

EVPN Layer-3 VNI Per Tenant for Routing Instance

Configure Layer-3 VNI per EVPN Tenant VRF Routing Instant

```
vlan 3901
  name l3-vni-vlan-for-tenant-2
  vn-segment 39010

interface Vlan3901
  description l3-vni-for-tenant-2-routing
  no shutdown
  vrf member evpn-tenant-2

vrf context evpn-tenant-2
  vni 39010
  rd auto
  address-family ipv4 unicast
    route-target import 39010:39010
    route-target export 39010:39010
    route-target both auto evpn
```

Define Layer-3 VNI for a 2nd tenant following the same steps in the previous slide

EVPN Layer-2 Network VXLAN VNI

Map VLANs to VXLAN VNIs and Configure their MP-BGP EVPN Parameters

```
vlan 200
  vn-segment 20000
vlan 210
  vn-segment 21000
```

Map VLAN to VXLAN VNI

```
evpn
  vni 20000 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 21000 l2
    rd auto
    route-target import auto
    route-target export auto
```

Under EVPN configuration, define RD and RT import/export policies for each Layer-2 VNIs

EVPN Layer-2 Network VXLAN VLAN SVI Interface

Create SVI interface for Layer-2 VNIs for VXLAN routing

```
interface Vlan200
  no shutdown
  vrf member evpn-tenant-1
  ip address 20.1.1.1/8
  fabric forwarding mode anycast-gateway


interface Vlan210
  no shutdown
  vrf member evpn-tenant-1
  ip address 21.1.1.1/8
  fabric forwarding mode anycast-gateway
```

Create SVI interface for a Layer-2 VNI. Associate it with the tenant VRF.

All VTEPs for this VLAN/VNI should have the same SVI interface IP address as the distributed IP gateway.

Enable distributed anycast gateway for this VLAN/VNI

EVPN Distributed Gateway

```
fabric forwarding anycast-gateway-mac 0002.0002.0002

interface Vlan210
  no shutdown
  vrf member evpn-tenant-2
  ip address 21.1.1.1/8
  fabric forwarding mode anycast-gateway
```

Configure distributed gateway virtual MAC address
One virtual MAC per VTEP
All VTEPs should have the same virtual MAC address

Configure virtual IP address
All VTEPs for this VLAN should have the same virtual IP address

Enable distributed gateway for this VLAN

VXLAN Tunnel Interface Configuration

Configure VXLAN tunnel interface nve1

```
interface nve1
  no shutdown
  source-interface loopback0
  host-reachability protocol bgp
  member vni 20000
    suppress-arp
    mcast-group 239.1.1.1
  member vni 21000
    suppress-arp
    mcast-group 239.1.1.2
  member vni 39000 associate-vrf
  member vni 39010 associate-vrf

interface loopback 0
  ip address 10.1.1.11/32
```

Specify loopback0 as the source interface

Define BGP as the mechanism for host reachability advertisement

Associate tenant VNIs to the tunnel interface nve1
Define the mcast group on a per-VNI basis
Enable arp suppression on a per-VNI basis

Add Layer-3 VNIs, one per tenant VRF

MP-BGP Configuration on VTEP

```
router bgp 100
  router-id 10.1.1.11
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  neighbor 10.1.1.2 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended

  vrf evpn-tenant-1
    address-family ipv4 unicast
      advertise l2vpn evpn
  vrf evpn-tenant-2
    address-family ipv4 unicast
      advertise l2vpn evpn
```

Address-family ipv4 unicast for prefix-based routing

Address-family l2vpn evpn for evpn host routes

Define MP-BGP neighbors.
Under each neighbor define address-family ipv4 unicast and l2vpn evpn

Send extended community in l2vpn evpn address-family to distribute EVPN route attributes

Under address-family ipv4 unicast of each tenant VRF instance, enable advertising EVPN routes

MP-BGP Configuration on iBGP Route Reflectore

```
router bgp 100
  router-id 10.1.1.1
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
    retain route-target all
  template peer vtep-peer
    remote-as 100
    update-source loopback0
    address-family ipv4 unicast
      send-community both
      route-reflector-client
    address-family l2vpn evpn
      send-community both
      route-reflector-client
  neighbor 10.1.1.11
    inherit peer vtep-peer
  neighbor 10.1.1.12
    inherit peer vtep-peer
  neighbor 10.1.1.13
    inherit peer vtep-peer
  neighbor 10.1.1.14
    inherit peer vtep-peer
```

Address-family ipv4 unicast for prefix-based routing

Address-family l2vpn evpn for EVPN vxlan host routes
Retain route-targets attributes

iBGP RR client peer template

Send both standard and extended community in address-family ipv4 unicast

Send both standard and extended community in address-family l2vpn evpn

Agenda

- VxLAN Overview
- MP-BGP EVPN Basics
- MP-BGP EVPN Control Plane
- VXLAN Design Options
- MP-BGP EVPN VXLAN Configuration
- VxLAN Capability on Nexus 9000 Series Switches

Nexus 9000 Series
VXLAN Support

VXLAN is supported across the Nexus 9000 series platforms. The VXLAN Gateway functionality is supported across all form factors and line cards. Integrated routing functionality is supported on Nexus 9300 switches and ACI-enabled Modules for Nexus 9500 switches.

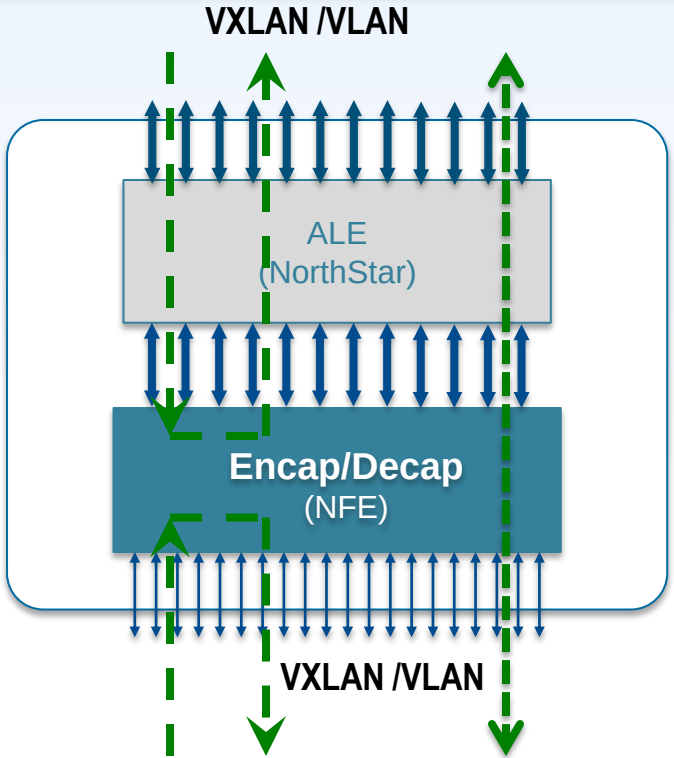
Nexus 9300 Series

Nexus 9500 Series

VXLAN Forwarding on Nexus 9000 NX-OS Mode

VXLAN Bridging and Gateway

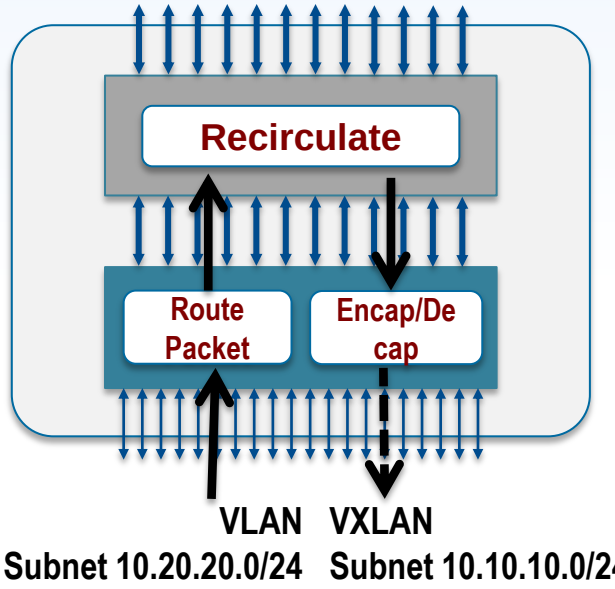
- VXLAN Encapsulation and De-encapsulation occur on T2
- Bridging and Gateway are independent of the port type (1/10/40G ports)
- Encapsulation happens on the egress port
- Decapsulation happens on the ingress port

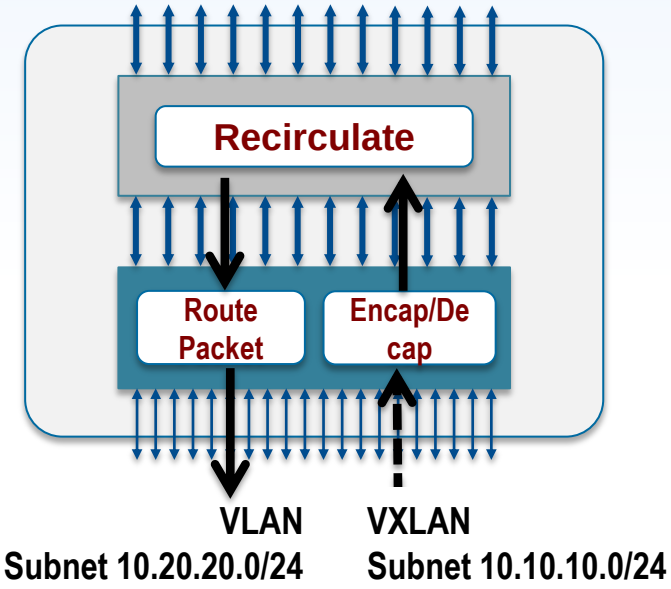


VXLAN Forwarding on Nexus 9000 NX-OS Mode

VXLAN Routing

- VXLAN Routing is not supported currently on Broadcom
- Additional recirculation required for VXLAN routing through NS





VXLAN Scales on Nexus 9000 Series Switches

Scale Parameter	Bronte Target
VxLAN enabled VLANs	1000
VxLAN enabled VRFs	900
VxLAN SVIs	1000
Total VNIs (L2/L3)	1900 (1000/900)
ECMP paths	64
Local VTEPs	1
Remote VTEPs	255


CISCO
TM