VXLAN Design and Deployment

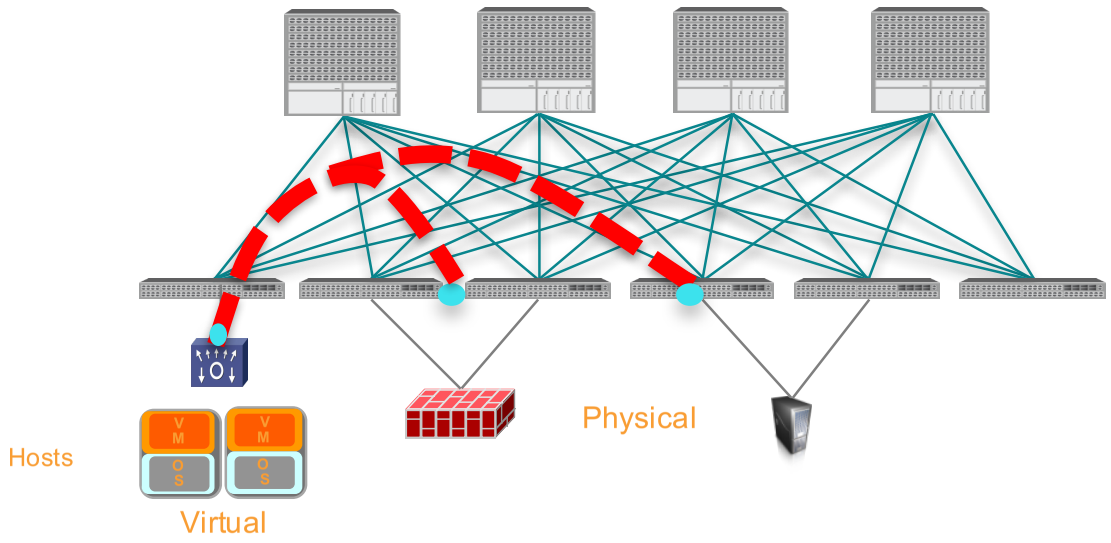
Marian Klas, Systems Engineer

mklas@cisco.com

Agenda

- Why VXLAN?
- VXLAN Fundamentals
- Underlay Deployment Considerations
- Overlay Deployment Considerations
- Summary and Conclusion

Trend: Flexible Data Center Fabrics



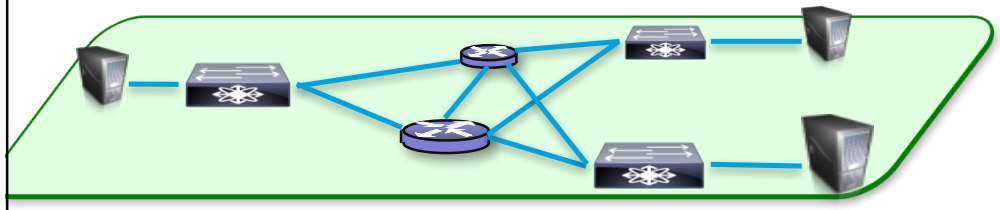
Use VXLAN to Create DC Fabrics

Mobility
Segmentation + Policy
Scale
Automated & Programmable
Full Cross Sectional BW
L2 + L3 Connectivity
Physical + Virtual

VXLAN Fundamentals

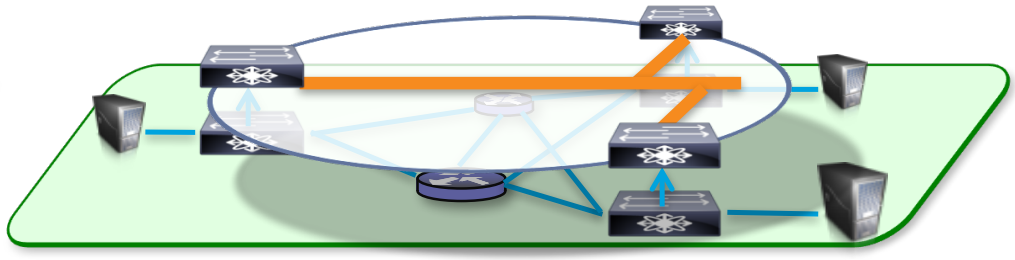
Why Overlays?

Seek well integrated best in class Overlays and Underlays



Robust Underlay/Fabric

- High Capacity Resilient Fabric
- Intelligent Packet Handling
- Programmable & Manageable



Flexible Overlay Virtual Network

- Mobility – Track end-point attach at edges
- Segmentation
- Scale – Reduce core state
  - Distribute and partition state to network edge
- Flexibility/Programmability
  - Reduced number of touch points

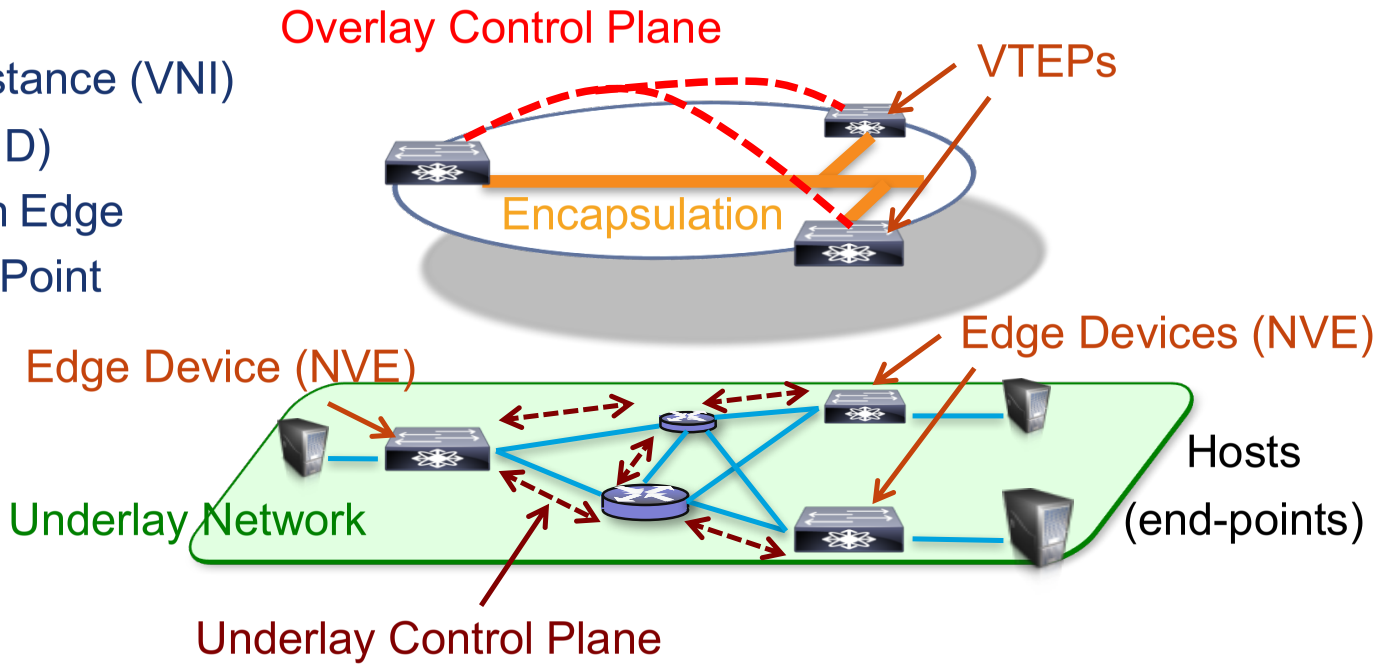
Overlay Taxonomy

Service = Virtual Network Instance (VNI)

Identifier = VN Identifier (VNID)

NVE = Network Virtualization Edge

VTEP = VXLAN Tunnel End-Point



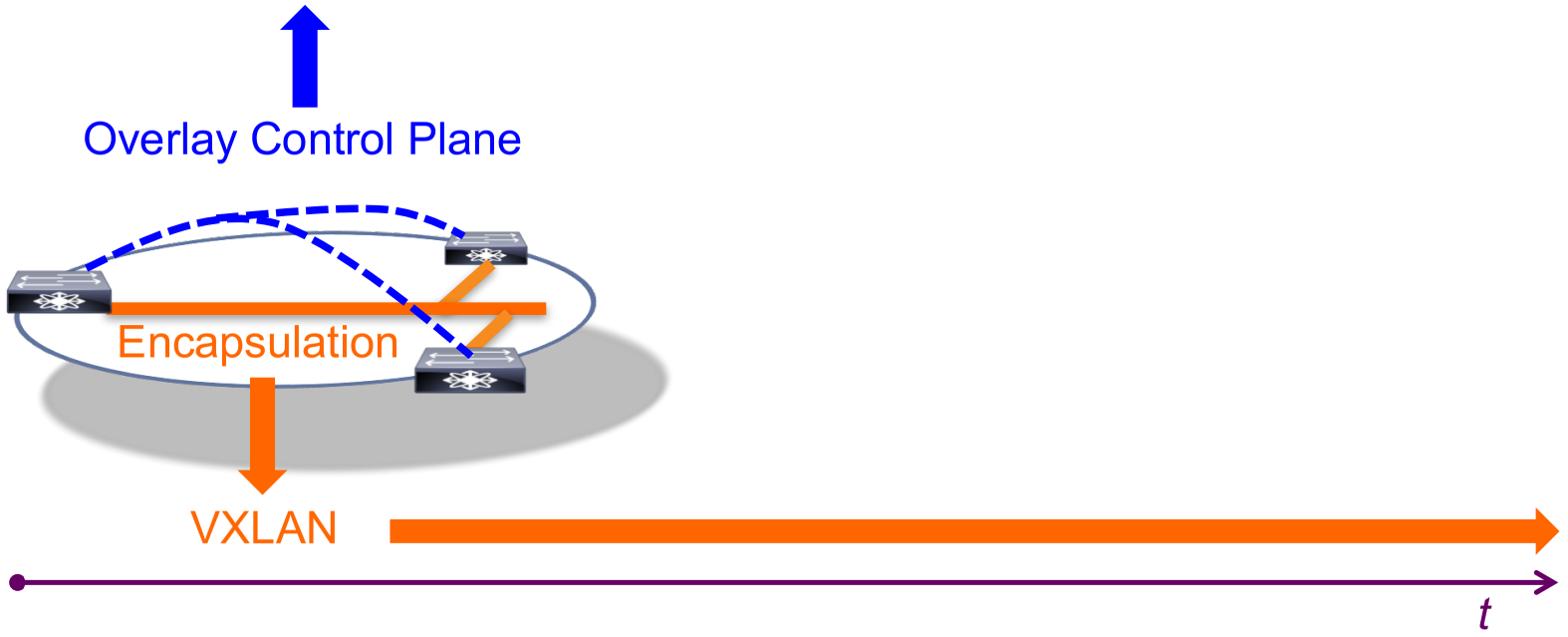
VXLAN is an Overlay Encapsulation

Data Plane Learning

Flood and Learn over a multidestination distribution tree joined by all edge devices

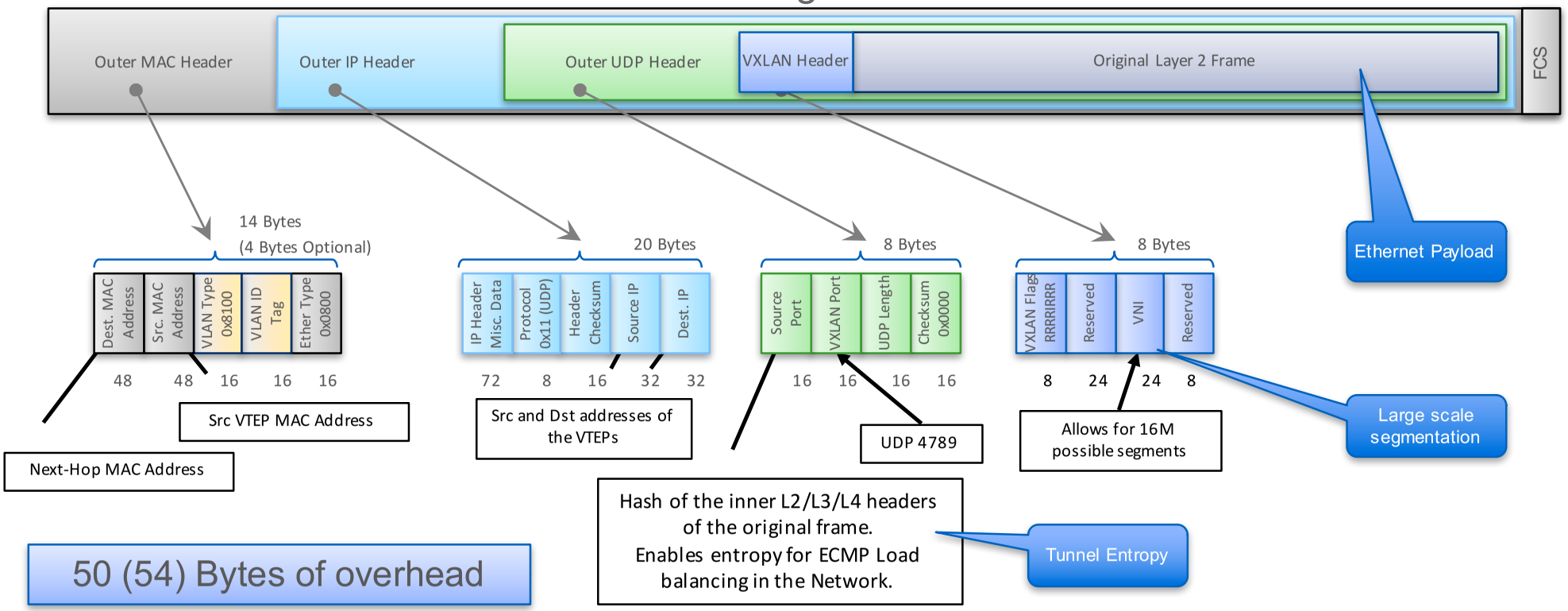
Protocol Learning

Advertise hosts in a protocol amongst edge devices



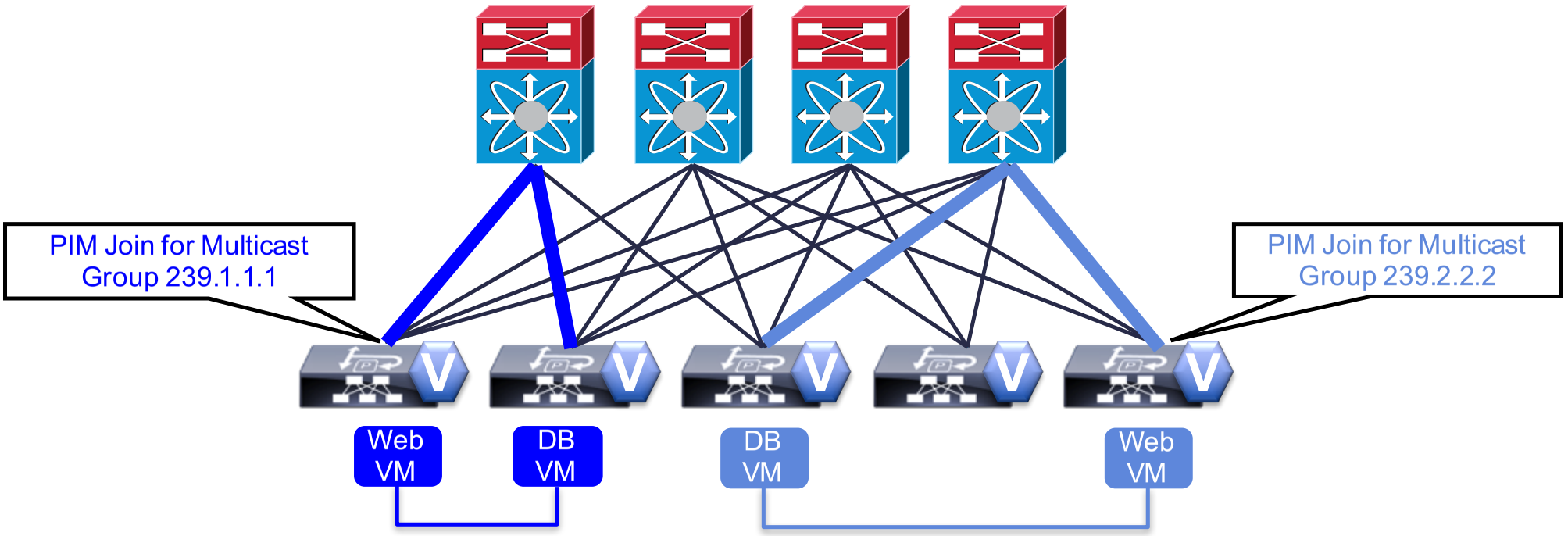
VXLAN Packet Structure

Ethernet in IP with a shim for scalable segmentation



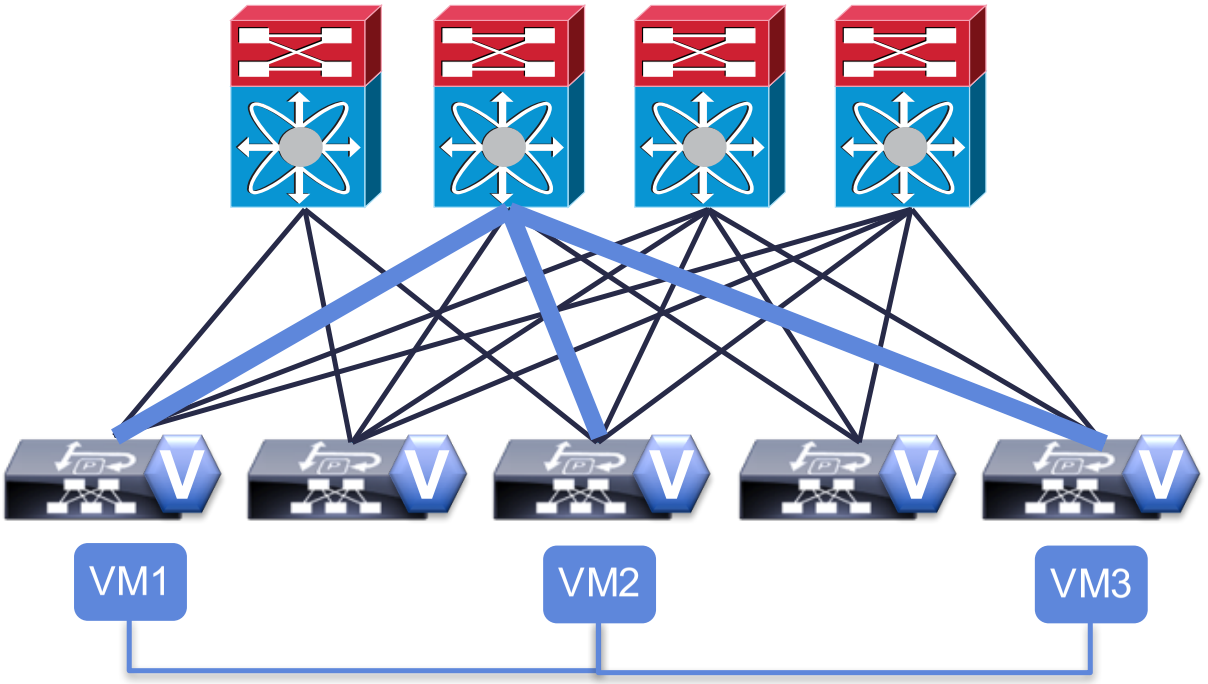
Data Plane Learning

Dedicated Multicast Distribution Tree per VNI



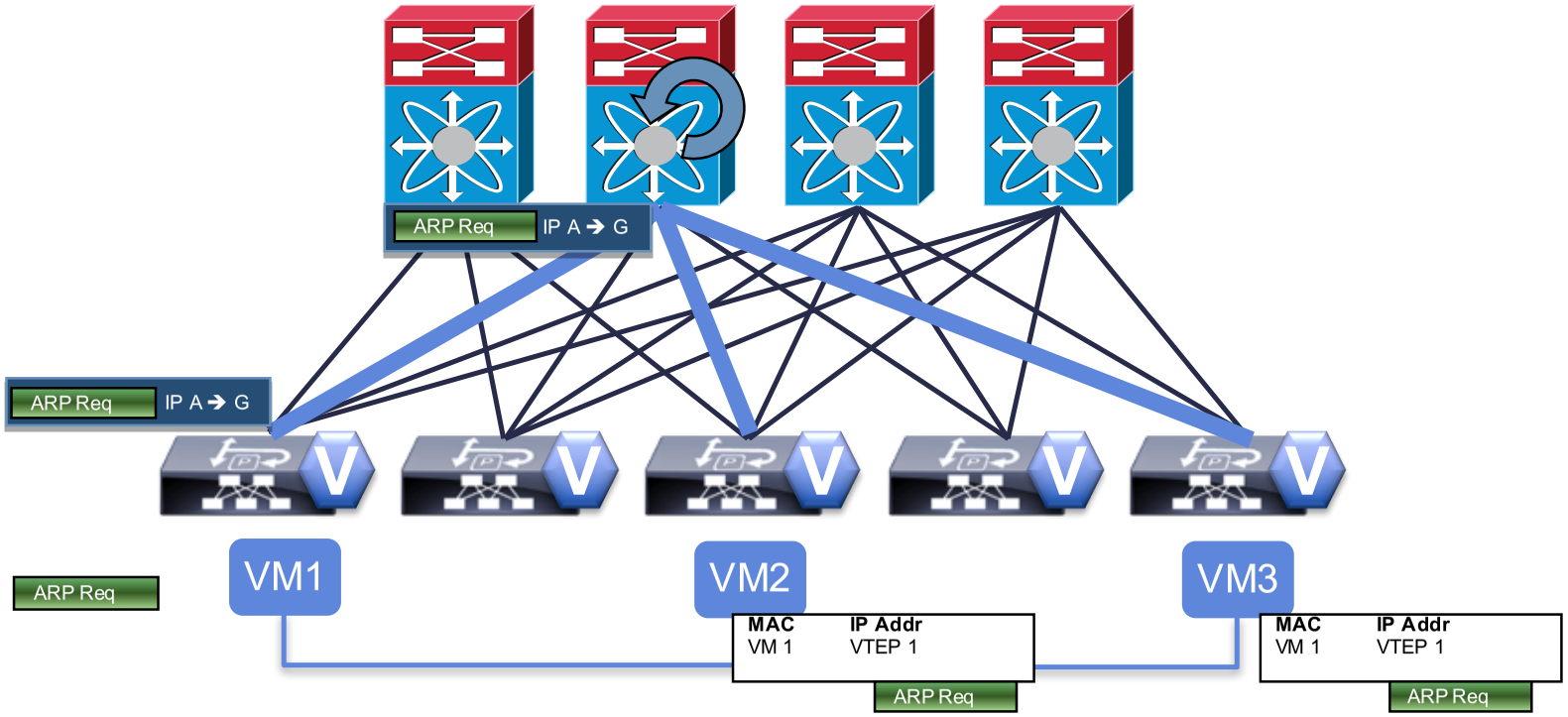
Data Plane Learning

Dedicated Multicast Distribution Tree per VNI



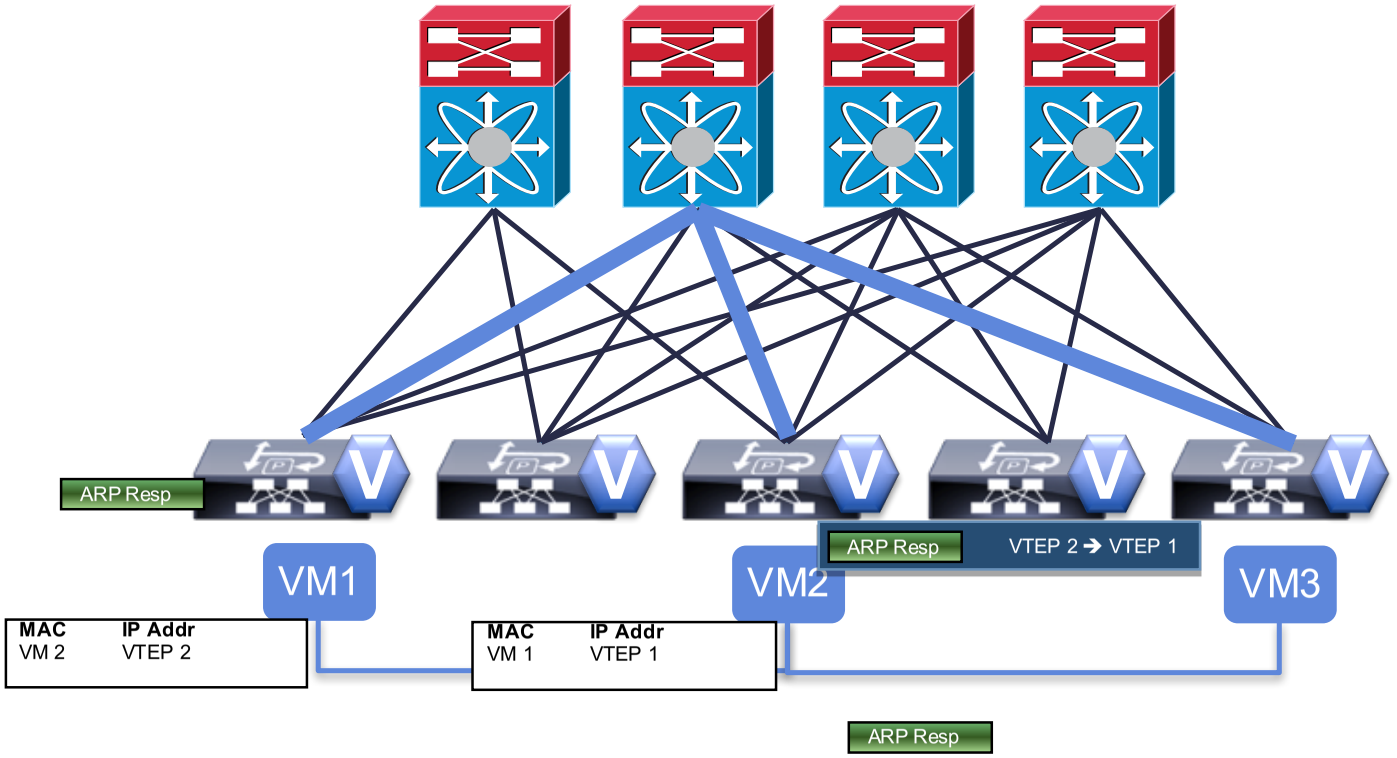
Data Plane Learning

Learning on Broadcast Source - ARP Request Example



Data Plane Learning

Learning on Unicast Source - ARP Response Example

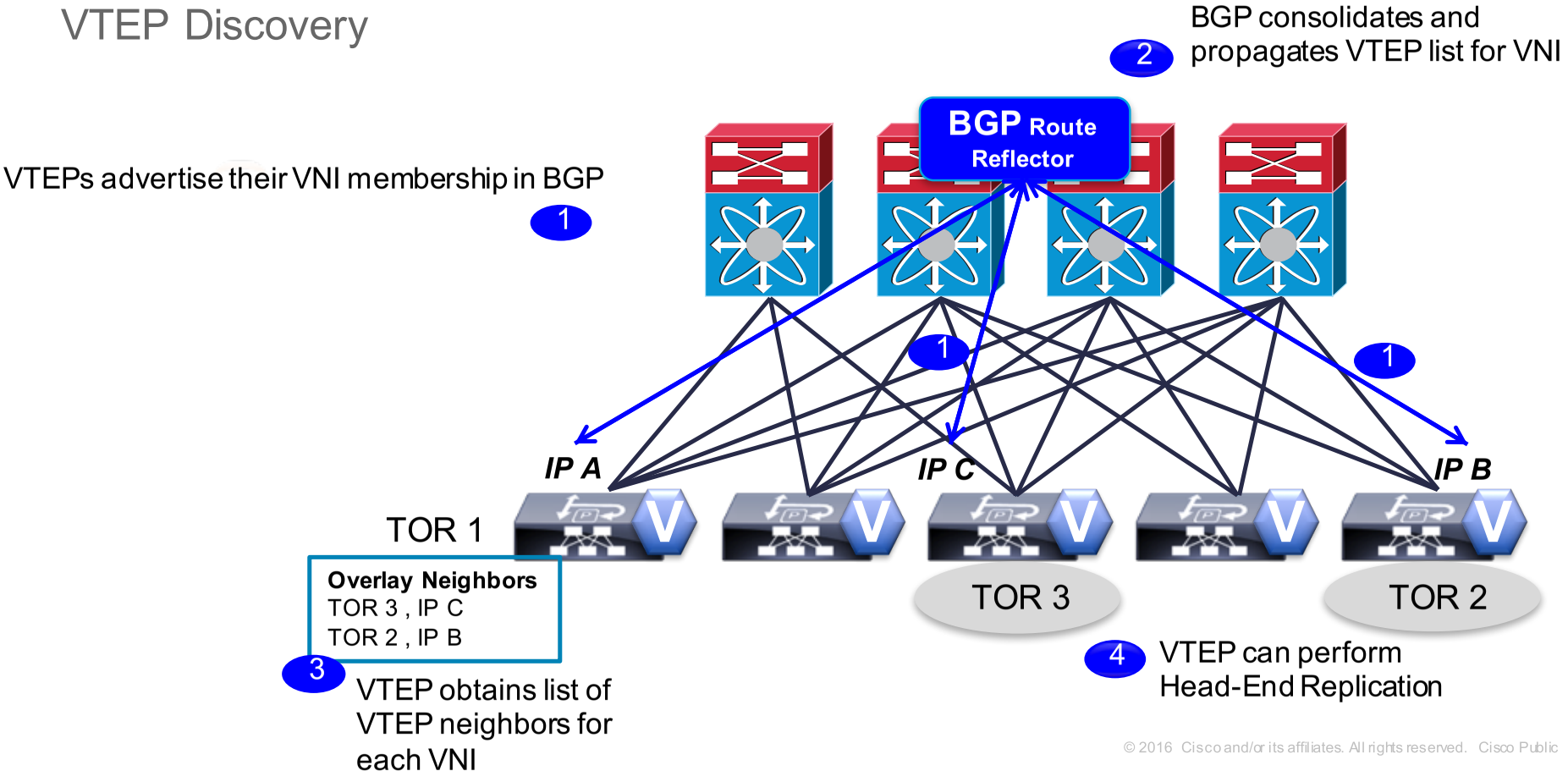


VXLAN Evolution

Multicast Independent	• Head-end replication enables unicast-only mode • Control Plane provides dynamic VTEP discovery
Protocol Learning prevents floods	• Workload MAC addresses learnt by VXLAN NVEs • Advertise L2/L3 address-to-VTEP association information in a protocol
External Connectivity	• VXLAN HW Gateways to other encaps/networks • VXLAN HW Gateway redundancy • Enable hybrid overlays
IP Services	• VXLAN Routing • Distributed IP Gateways

VXLAN Evolution: Using a Control Protocol

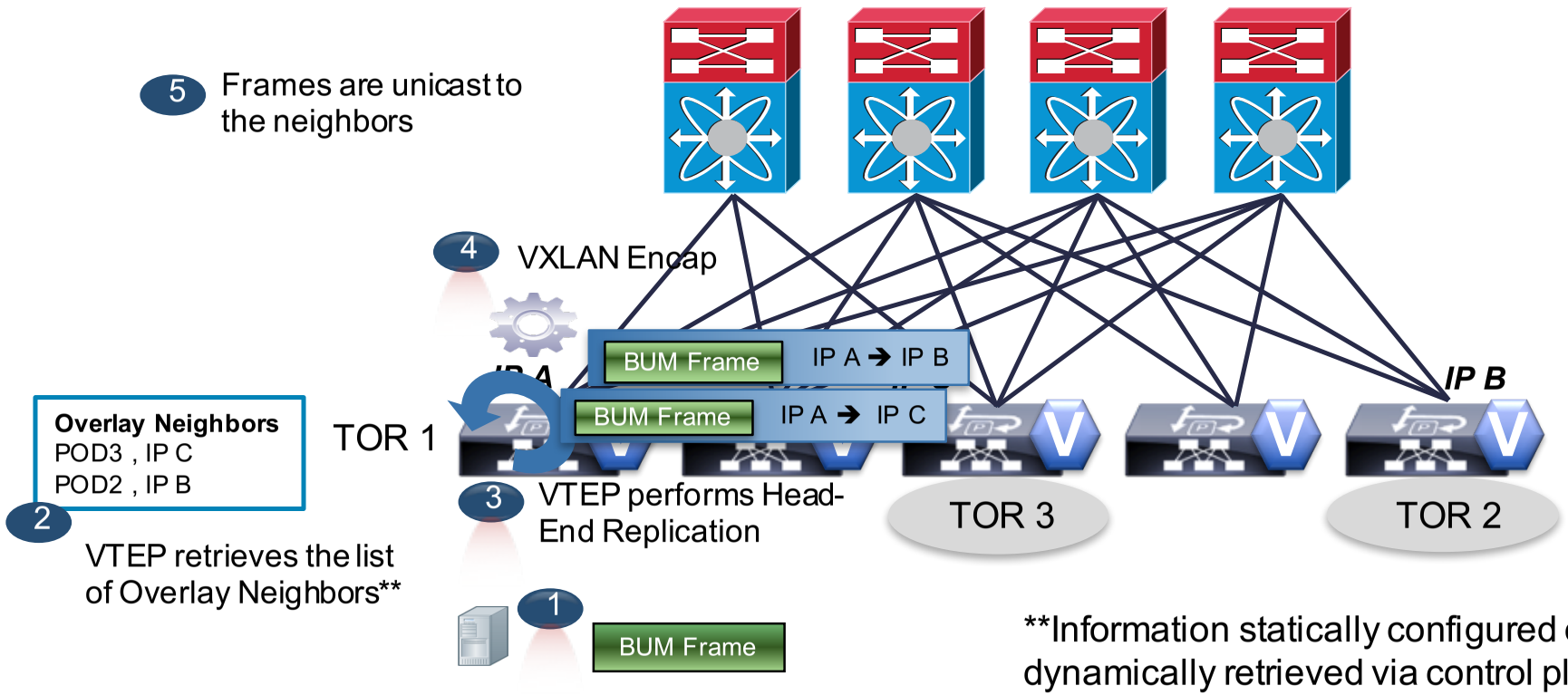
VTEP Discovery



VXLAN Unicast Mode

Head-end replication

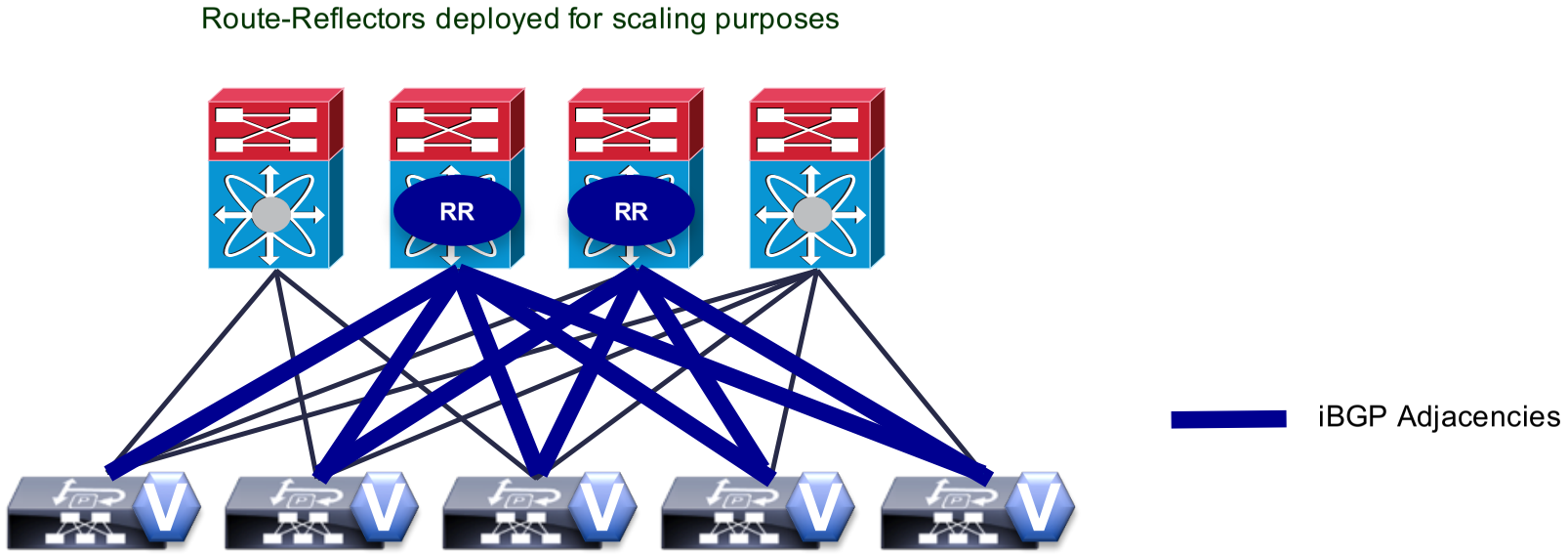
*Broadcast, Unknown Unicast or Multicast



**Information statically configured or dynamically retrieved via control plane (VTEP discovery)

BGP EVPN Control Plane for VXLAN

Host and Subnet Route Distribution



- Host Route Distribution decoupled from the Underlay protocol
- Use MP-BGP on the leaf nodes to distribute internal host/subnet routes and external reachability information

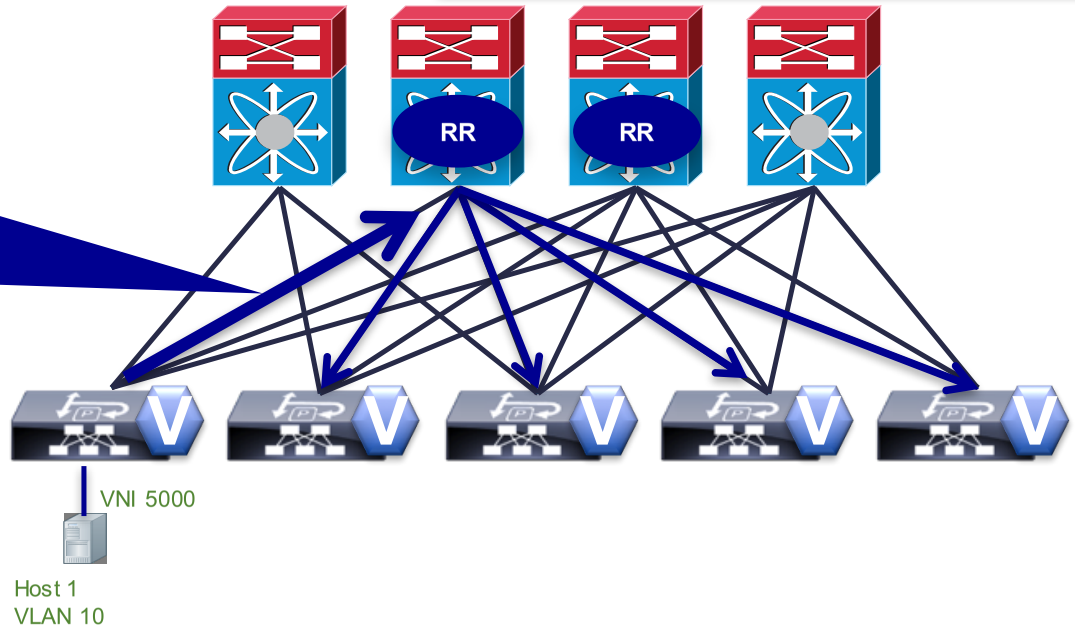
BGP EVPN Control Plane

Host Advertisement

MAC	IP	VNI	Next-Hop	Encap	Seq
1	1	5000	IP L1 MAC L1	VXLAN	0

NLRI:
- Host MAC1, IP1
- NVE IP L1/MAC L1
- VNI 5000

Ext.Community:
- Encapsulation: VXLAN, NVGRE
- Sequence 0



1. Host Attaches
2. Attachment NVE advertises host's MAC (+IP) through BGP RR
3. Choice of encapsulation is also advertised

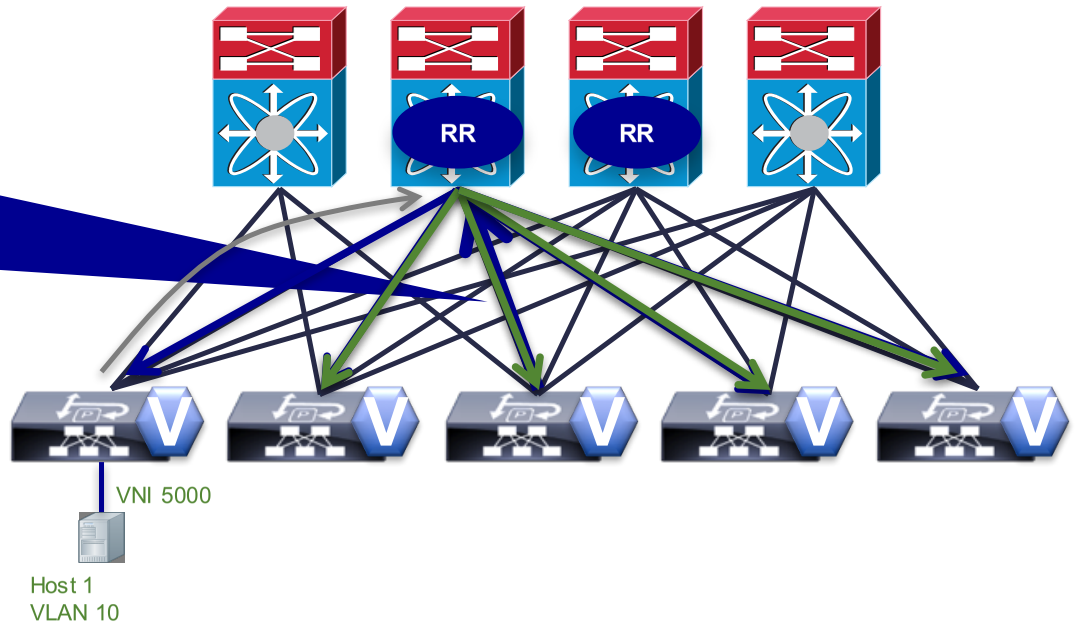
BGP EVPN Control Plane

Host Moves

MAC	IP	VNI	Next-Hop	Encap	Seq
1	1	5000	IP L3 MAC L3	VXLAN	1

NLRI:
- Host MAC1, IP1
- NVE IP L3/MAC L3
- VNI 5000

Ext.Community:
- Encapsulation: VXLAN, NVGRE
- Sequence 1

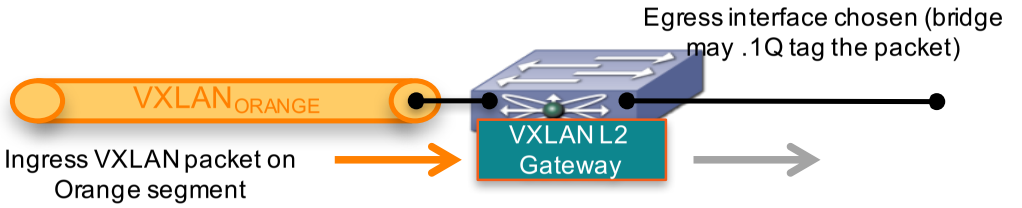


1. Host Moves to NVE3
2. NVE3 detects Host1 and advertises H1 with seq#1
3. NVE1 sees more recent route and withdraws its advertisement

VXLAN L2 and L3 Gateways

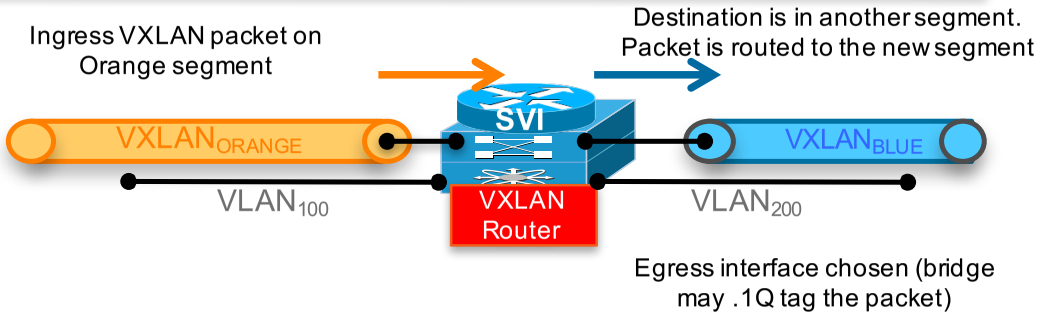
Connecting VXLAN to the broader network

L2 Gateway: VXLAN to VLAN Bridging



L3 Gateway: VXLAN to X Routing

- VXLAN
- VLAN



	N7K w/F3 LC	Nexus 3K	N5600	N9K	ASR1K/ASR 9K
L2 Gateway	Yes	Yes	Yes	Yes	Yes
L3 Gateway	Yes	No	Yes	Yes	Yes

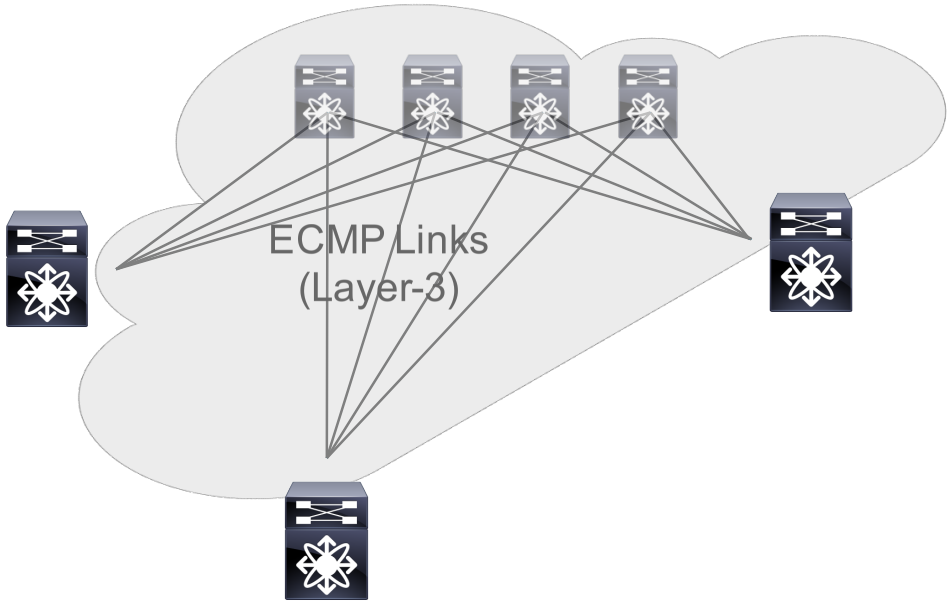
VXLAN BGP EVPN Nexus Programmable Fabric Support

Platform	Spine Function	Leaf Function	Border Leaf Function	DCNM Fabric Management
Nexus 5600	Yes 7.3(0)N1(1)	Yes 7.3(0)N1(1)	Yes 7.3(0)N1(1)	Yes 7.2(3)
Nexus 7000/7700 (F3 Only)	Yes 7.3(0)D1(1)	Yes 7.3(0)D1(1)	Yes 7.3(0)D1(1)	Yes 7.2(3)
Nexus 9300	Yes 7.0(3)I1(3)	Yes 7.0(3)I1(3)	Yes 7.0(3)I1(3)	Yes 7.2(3)
Nexus 9500	Yes 7.0(3)I1(3)	Yes 7.0(3)I1(3)	Yes 7.0(3)I1(3)	Yes 7.2(3)

Underlay Deployment Considerations

Leaf/Spine Topology – Underlay Considerations

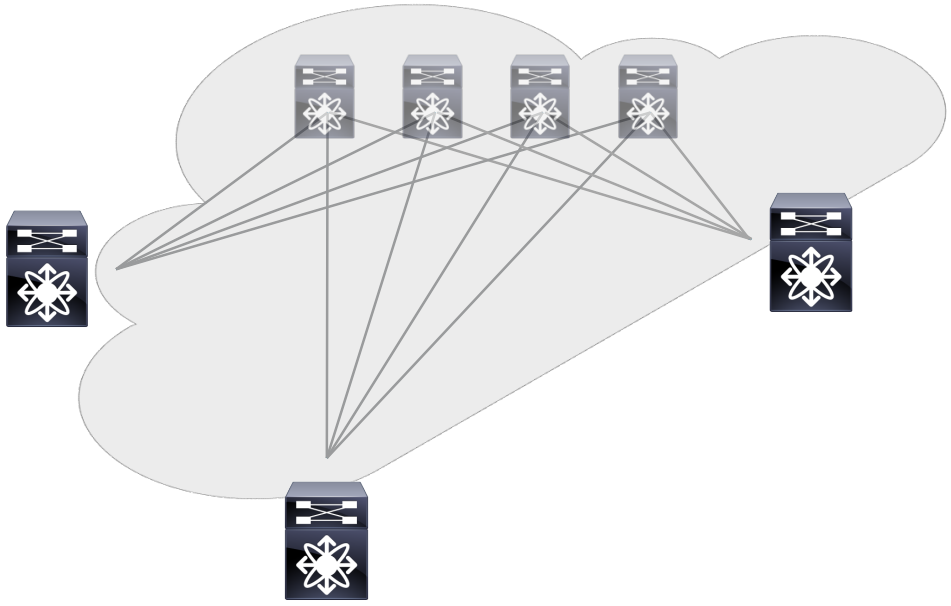
- High Bi-Sectional Bandwidth
- Wide ECMP: Unicast or Multicast
- Uniform Reachability, Deterministic Latency
- High Redundancy: Node/Link Failure
- Line rate, low latency, for all traffic



Deployment Considerations

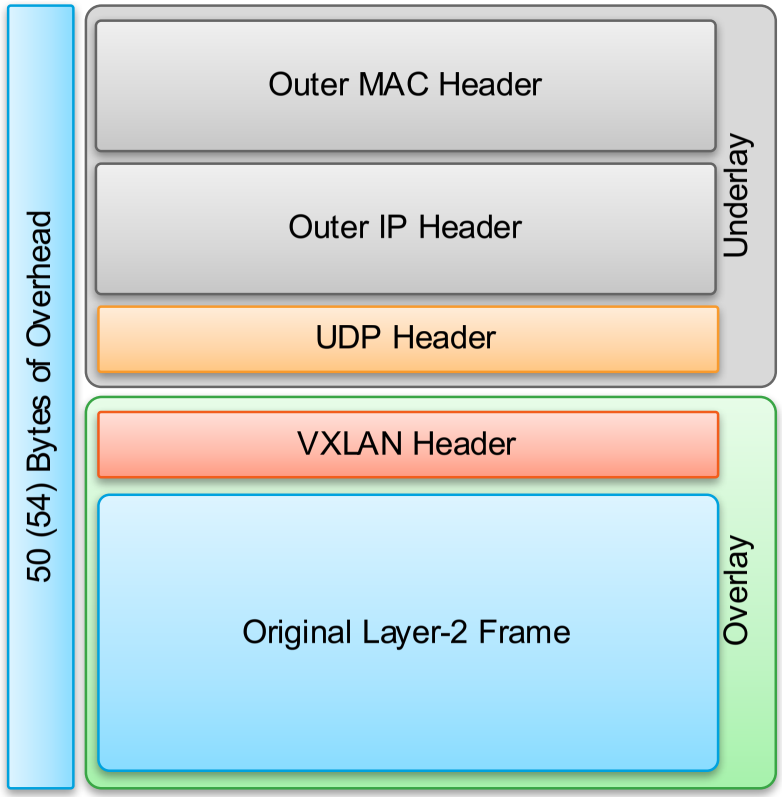
Underlay

- MTU and Overlays
- Unicast Routing Protocol and IP Addressing
- Multicast for BUM* Traffic Replication



MTU and VXLAN

Underlay



No Fragmentation Needed

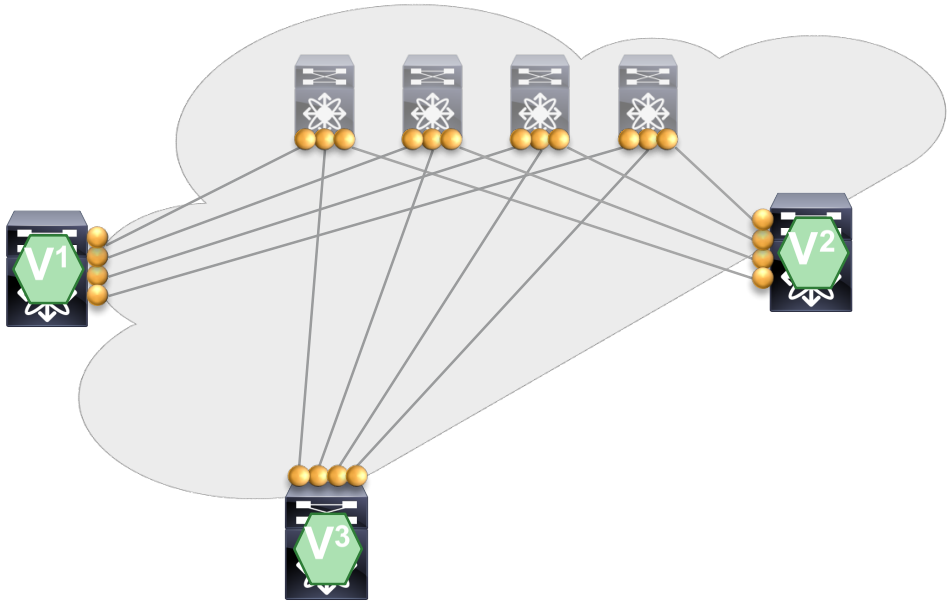
- VXLAN adds 50 Bytes to the Original Ethernet Frame
- Avoid Fragmentation by adjusting the IP Networks MTU
- Data Centers often require Jumbo MTU; most Server NIC do support up to 9000 Bytes
- Using a MTU of 9216* Bytes accommodates VXLAN Overhead plus Server max. MTU

*Cisco Nexus 5600/6000 switches only support 9192 Byte for Layer-3 Traffic

Building your IP Network – Interface Principles

Underlay

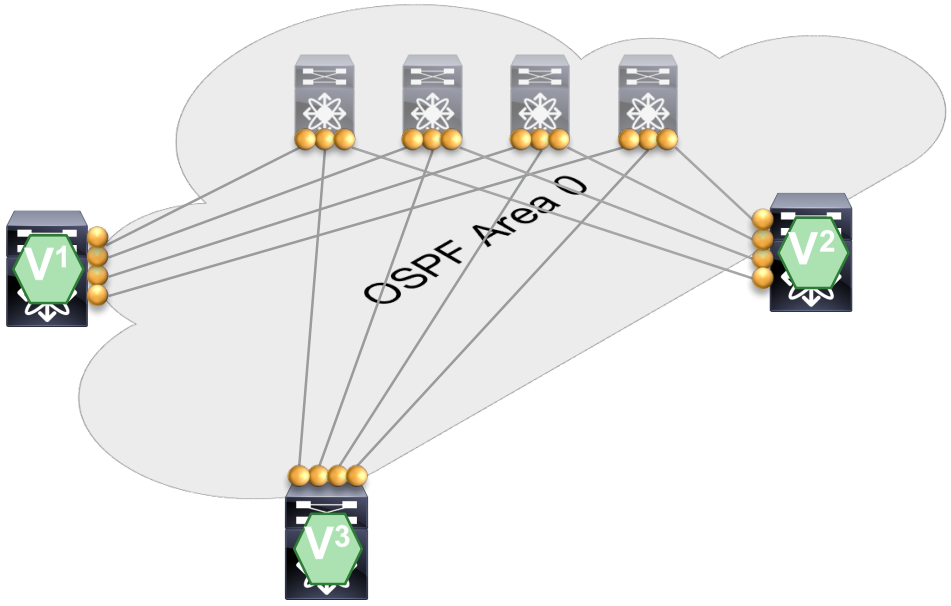
- Know your IP addressing and IP scale requirements
  - Best to use single Aggregate for all Underlay Links and Loopbacks
  - IPv4 only
  - For each Point-2-Point (P2P) connection, minimum /31 required
  - Loopback requires /32
- Routed Ports/Interfaces
  - Layer-3 Interfaces between Spine and Leaf (no switchport)
- VTEP uses Loopback as Source-Interface



Building your IP Network – Routing Protocols; OSPF

Underlay

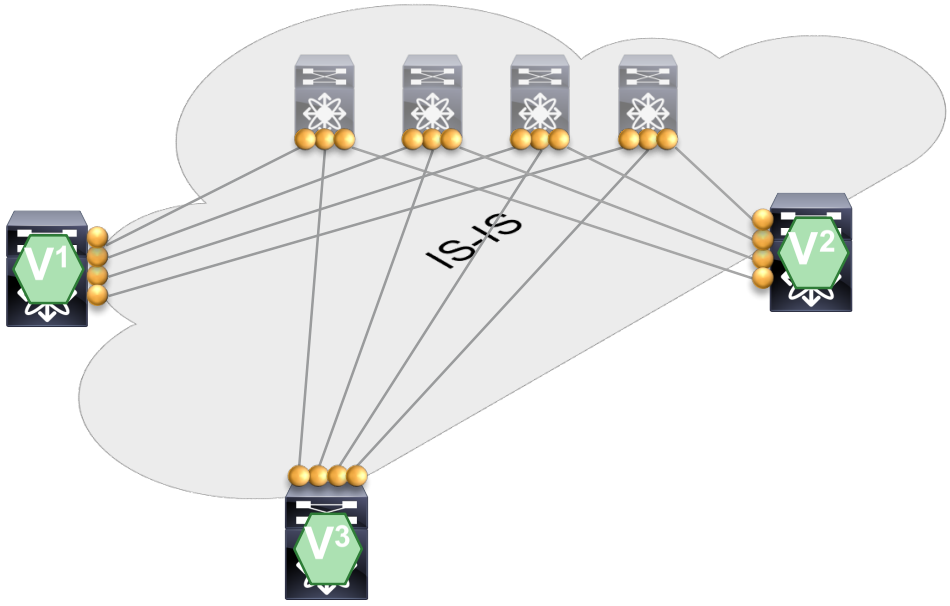
- OSPF – watch your Network type
  - Network Type Point-2-Point (P2P)
    - Preferred (only LSA type-1)
    - No DR/BDR election
    - Suits well for routed interfaces/ports (optimal from a LSA Database perspective)
    - Full SPF calculation on Link Change
  - Network Type Broadcast
    - Suboptimal from a LSA Database perspective (LSA type-1 & 2)
    - DR/BDR election
    - Additional election and Database Overhead



Building your IP Network – Routing Protocols; IS-IS

Underlay

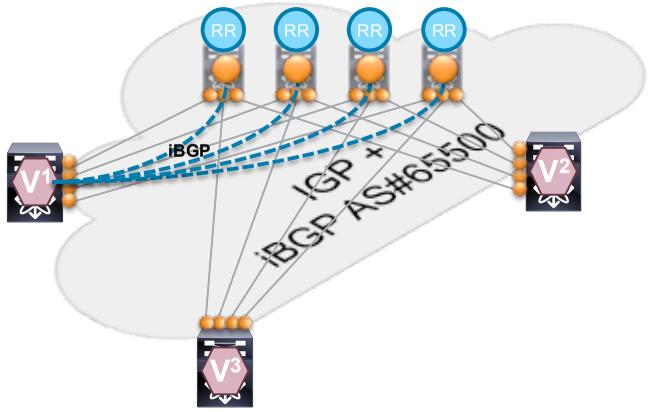
- IS-IS – what was this CLNS?
  - Independent of IP (CLNS)
  - Well suited for routed interfaces/ports
  - No SPF calculation on Link change; only if Topology changes
  - Fast Re-convergence
  - Not everyone is familiar with it



Building your IP Network – Routing Protocols; iBGP

Underlay

- iBGP + IGP = The Routing Protocol Combo
  - IGP for underlay topology & reachability (e.g. IS-IS, OSPF)
  - iBGP for VTEP (loopback) reachability
  - iBGP route-reflector for simplification and scale
  - Requires two routing protocols
  - Separates Links (IGP) from VTEPs (iBGP)
    - End-Host information are still in iBGP but different address-family



Multicast Enabled Underlay

- May use PIM-ASM or PIM-BiDir (Different hardware has different capabilities)

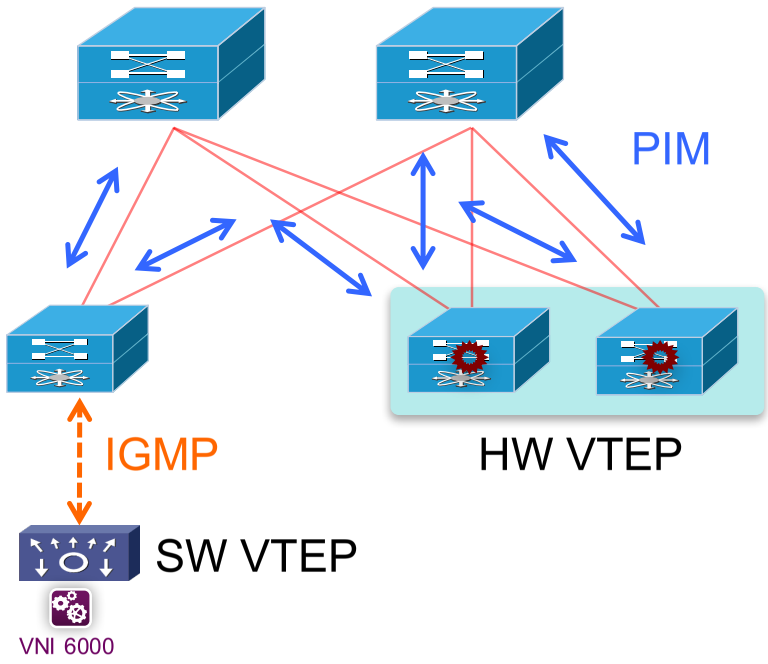
	N1KV	Nexus 7K with F3 LC	Nexus 3K	Nexus 5K/6K	Nexus 9K Standalone	CSR 1000V ASR1K	ASR9K
Mcast mode	IGMP v2/v3	PIM-ASM & Bidir-PIM	PIM-ASM (Bidir – Future)	Bidir-PIM	PIM-ASM (Bidir – Future)	Bidir-PIM (ASM –Future)	PIM-ASM & Bidir-PIM

- Cisco Nexus 9000 Series Switches (supports ASM/SSM/Ingress-Replication for VXLAN VTEP support) and Cisco Nexus 5600 Series Switches (supports only PIM BIDIR for VXLAN VTEP support) can be part of the same VXLAN EVPN fabric but not share the same Layer-2 VNI.
- Spine and Aggregation switches make good RP locations in clos and traditional topologies respectively
- Reserve a range of multicast channels to service the overlay and optimize for diverse VNIs
- In clos topologies with lean spine, using multiple RPs across the multiple spines and mapping different VNIs to different RPs will provide a simple load balancing measure
- Design a multicast underlay for a network overlay, host VTEPs will simply leverage this network.

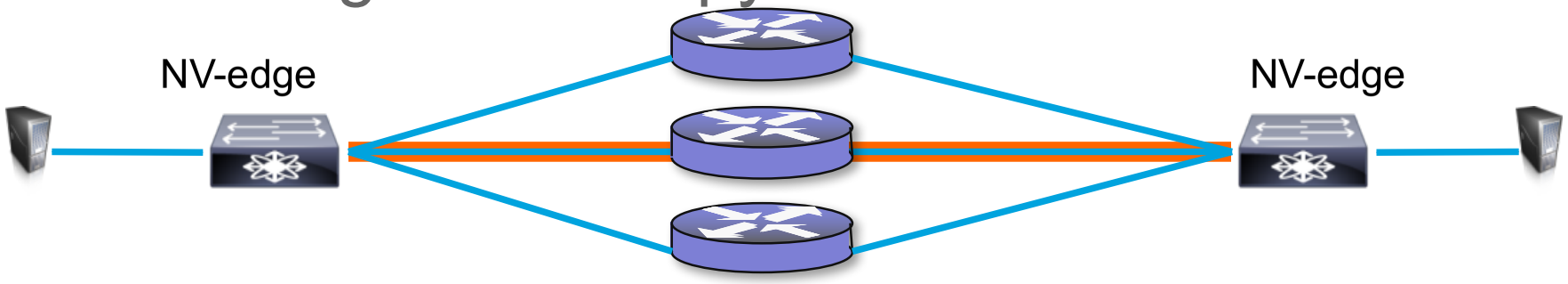
Multicast Enabled Underlay

Host Overlay to Hybrid Overlay

- Host Overlay VTEPs join multicast groups as hosts using IGMP reports
- Host overlays will work over an L2 underlay, ensure IGMP snooping is in place to scope the reach of multicast
- A multicast enabled L3 underlay is the better option as it enables a hybrid overlay (host and network VTEPs)
  - Ensure that the first hop router for the host in the underlay is configured to service the IGMP reports from the host VTEP



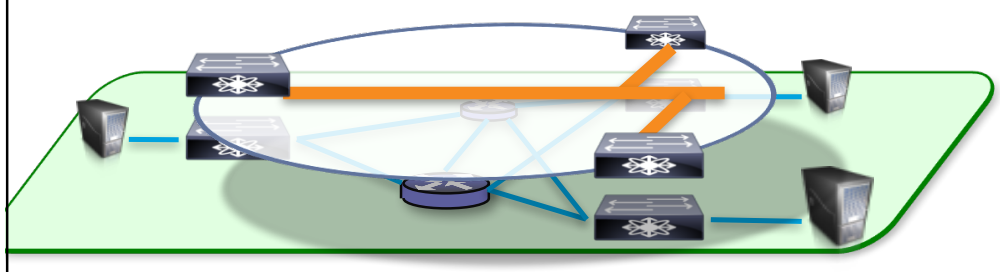
Multi-Pathing and Entropy

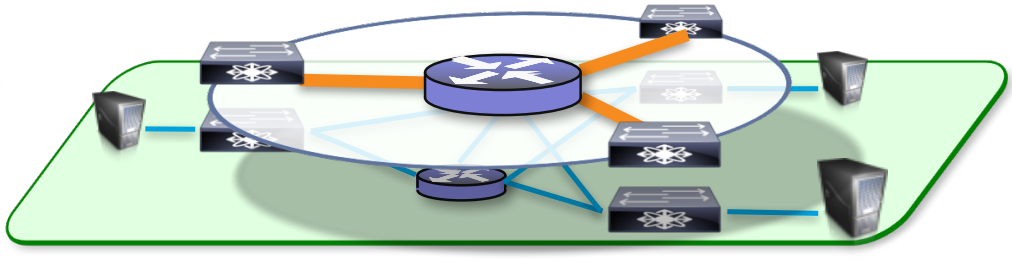


- Symmetric Underlay Network topologies facilitate ECMP routing:
  - Multi-path load balancing
  - Fast Re-convergence on link Failures
- Polarization: Encapsulated flows appear as a single flow which hashes to a single path
- Entropy in the encapsulation header to depolarize tunnels
  - Variable UDP source port in VXLAN outer header
  - Underlay must support ECMP hashing on L4 port numbers

Overlay Deployment Considerations

Type of Overlay Service





Layer 2 Overlays

- Emulate a LAN segment
- Transport Ethernet Frames (IP and non-IP)
- Single subnet mobility (L2 domain)
- Exposure to open L2 flooding
- Useful in emulating physical topologies

Layer 3 Overlays

- Abstract IP based connectivity
- Transport IP Packets
- Full mobility regardless of subnets
- Contain network related failures (floods)
- Useful in abstracting connectivity and policy

Hybrid L2/L3 Overlays offer the best of both domains

How to configure?

Building your VTEP (VXLAN Tunnel End-Point)

Overlay

Configuration Example

```
# Features & Globals
feature bgp
feature nv overlay
nv overlay evpn

# Spine (S1)

# Leaf (V1)
interface nve1
  source-interface loopback0
  host-reachability protocol bgp
```

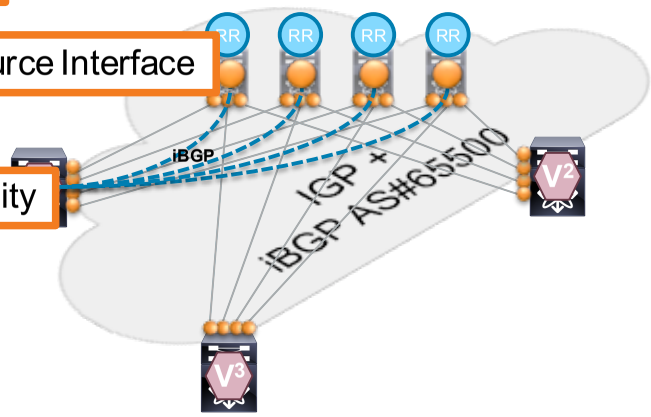
Enables VTEP (only required on Leaf or Border)

Enables EVPN Control-Plane in BGP

Configure the VTEP interface

Use a Loopback for Source Interface

Enable BGP for Host reachability



*Simplified BGP configuration; would have 4 BGP peers (RR)
IGP not shown

Building your Overlay Control-Plane

Overlay

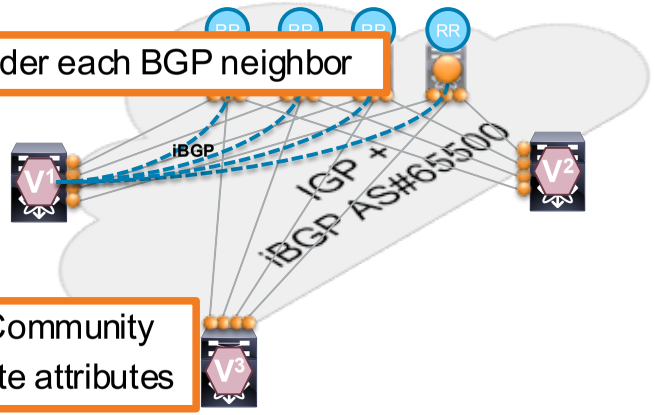
Configuration Example

```
# Features & Globals
feature bgp
feature nv overlay
nv overlay evpn

# Spine (S1)
router bgp 65500
  router-id 10.10.10.S1
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.10.10.V1 remote-as 65500
    update-source loopback0
    address-family l2vpn evpn
      send-community both
      route-reflector-client

# Leaf (V1)
router bgp 65500
  router-id 10.10.10.V1
  address-family ipv4 unicast
  neighbor 10.10.10.S1 remote-as 65500
    update-source loopback0
    address-family l2vpn evpn
      send-community both
```

Enables EVPN Control-Plane in BGP

Activate L2VPN EVPN under each BGP neighbor

Send Extended BGP Community to distribute EVPN route attributes

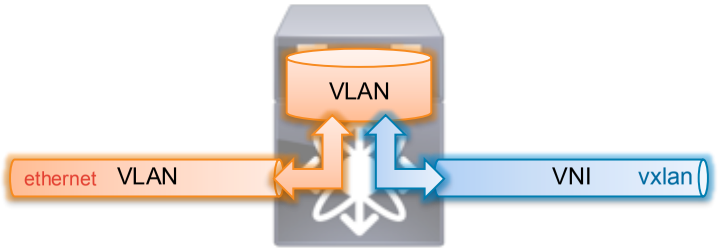


*Simplified BGP configuration; would have 4 BGP peers (RR)
IGP not shown

Extend your VLAN to VXLAN

Overlay

- Mapping a IEEE 802.1Q VLAN ID to a VXLAN VNI
  - VLAN to VNI configuration on a per-Switch based
  - VLAN becomes “Switch Local Identifier”
  - VNI becomes “Network Global Identifier”
  - 4k VLAN limitation per-Switch does still apply
  - 4k Network limitation has been removed
  - Dependent on VLAN Space!



Configuration Example

```
# Features
feature vn-segment-vlan-based

# VLAN to VNI mapping (MT-Lite)
vlan 43
 vn-segment 30000

# Activate Layer-2 VNI for EVPN
evpn
  vni 30000 l2
    rd auto
    route-target import auto
    route-target export auto

# Activate Layer-2 VNI on VTEP
interface nve1
  source-interface loopback0
  host-reachability protocol bgp
  member vni 30000
    mcast-group 239.239.239.100
    suppress-arp
```

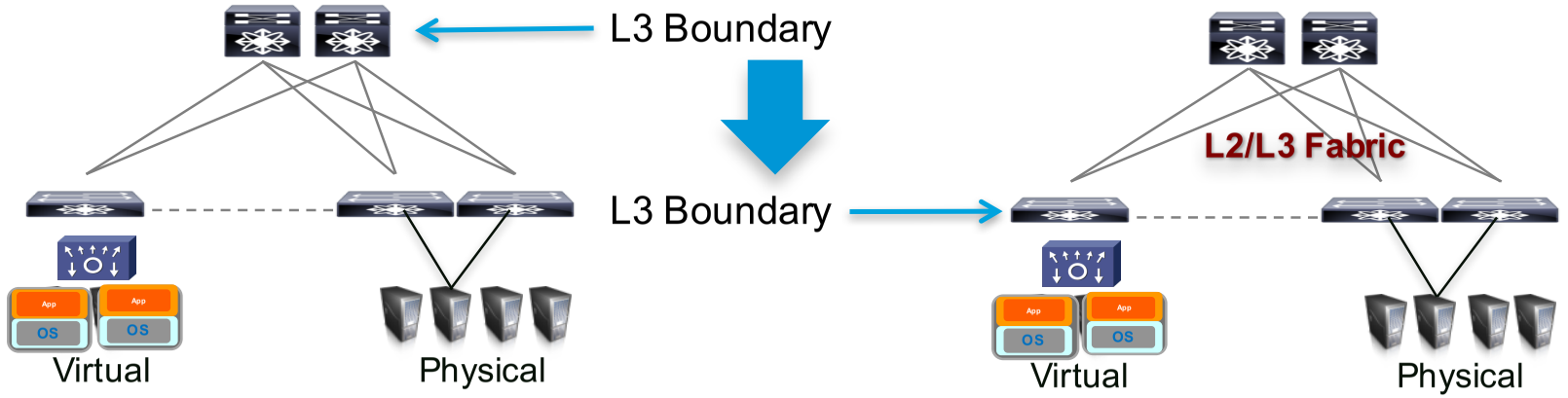
VLAN to Layer-2 VNI mapping

Enables EVPN Control-Plane for Layer-2 Services

Alternative is to use “`ingress-replication protocol bgp`”

Enables Layer-2 VNI on VTEP and suppress ARP

Distributed Gateway Function in L3 Overlays



Traditional L2 - centralised L2/L3 boundary

- Always bridge, route only at an aggregation point
- Large amounts of state converge
- Scale problem for large# of L2 segments
- Traditional L2 and L2 overlays

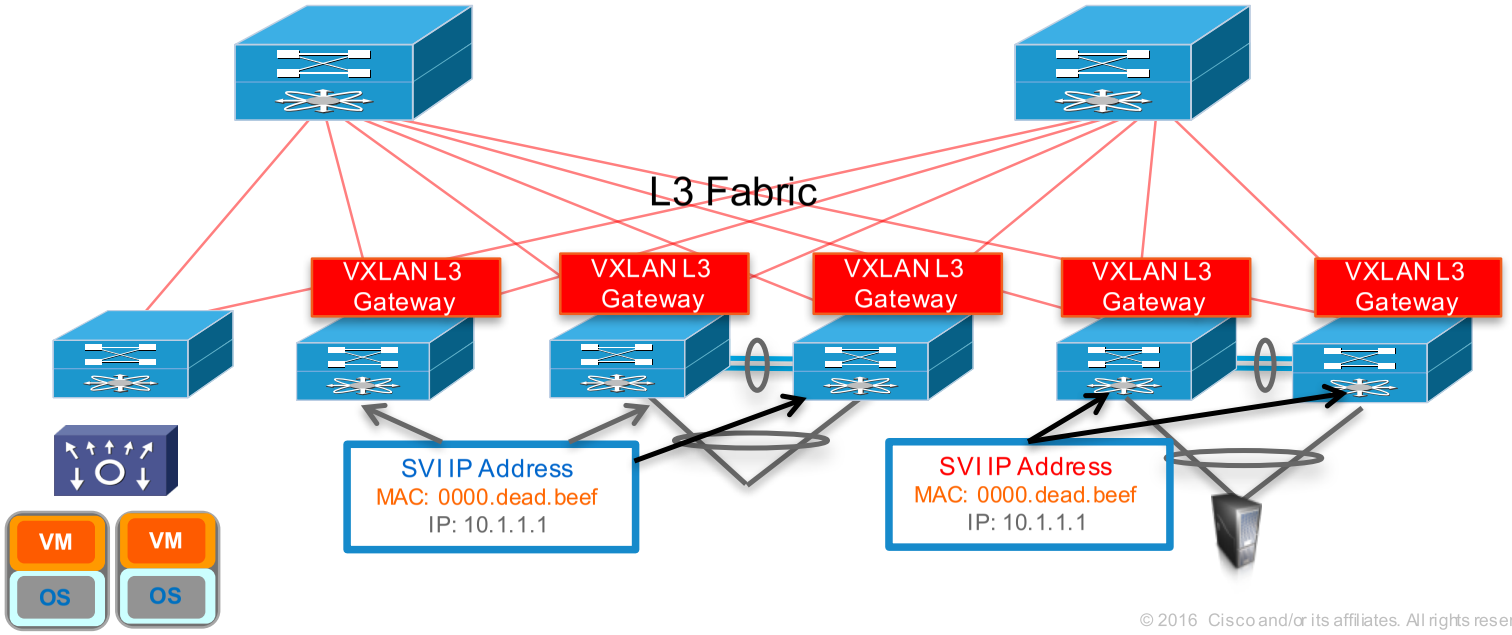
L2/L3 fabric (or overlay)

- Always route (at the leaves), bridge when necessary
- Distribute and disaggregate necessary state
- Optimal scalability
- Enhanced forwarding and L3 overlays

Distributed IP Anycast Gateway

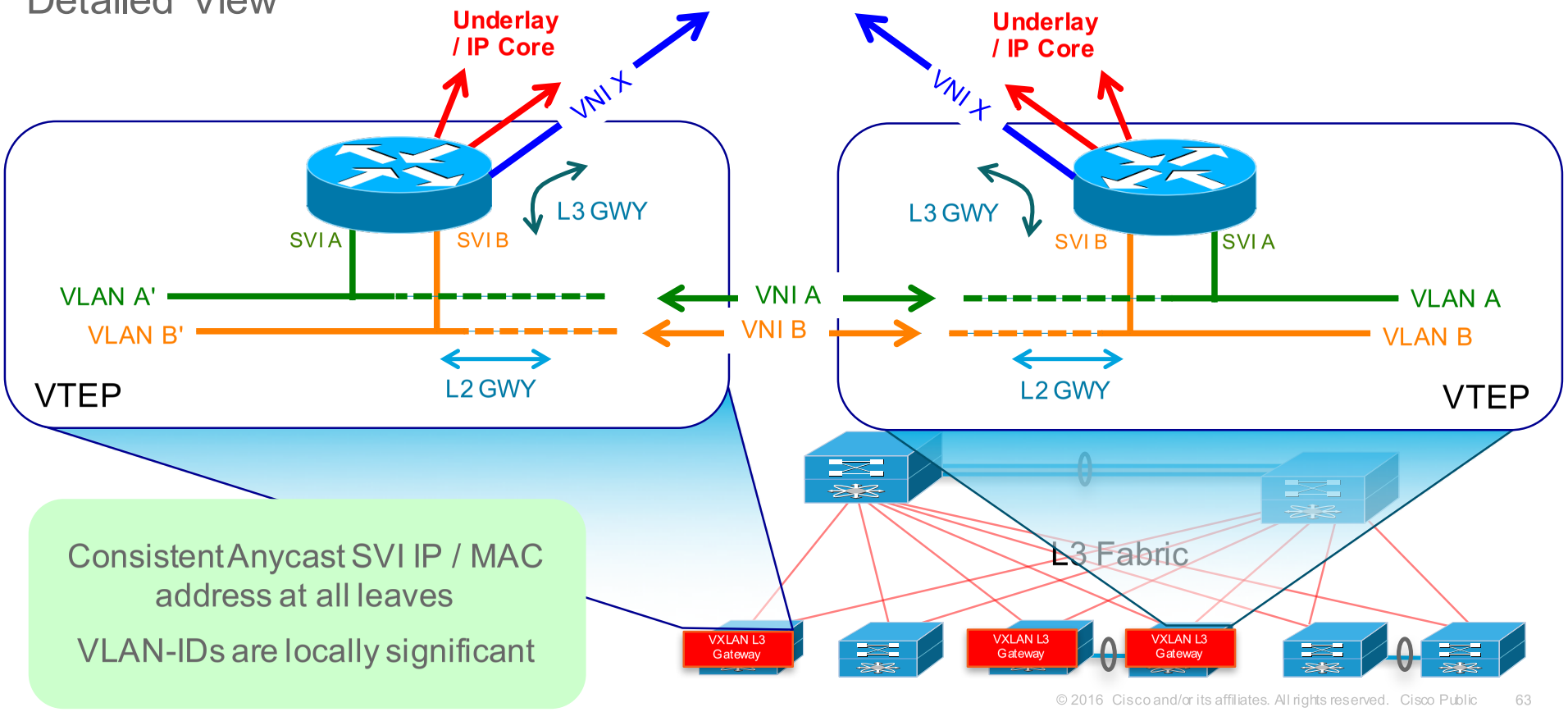
The same “Anycast” SVI IP/MAC is used at all VTEPs/ToRs

A host will always find its SVI anywhere it moves



Distributed IP Anycast Gateway

Detailed View

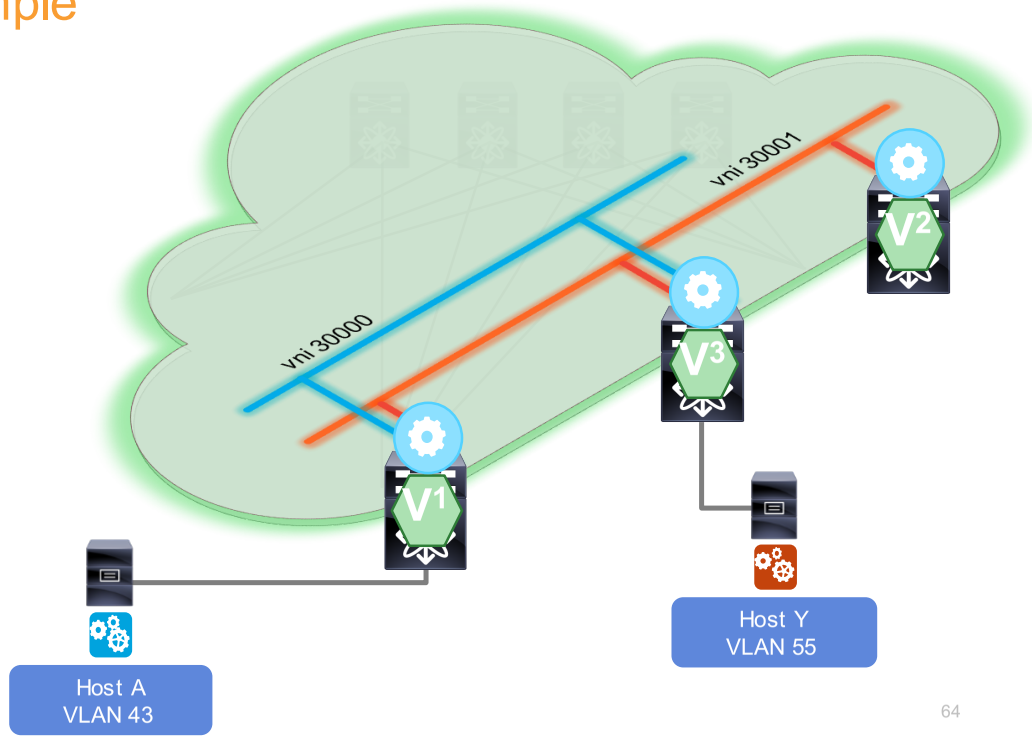


Consistent Anycast SVI IP / MAC address at all leaves

VLAN-IDs are locally significant

Distributed IP Anycast Gateway*

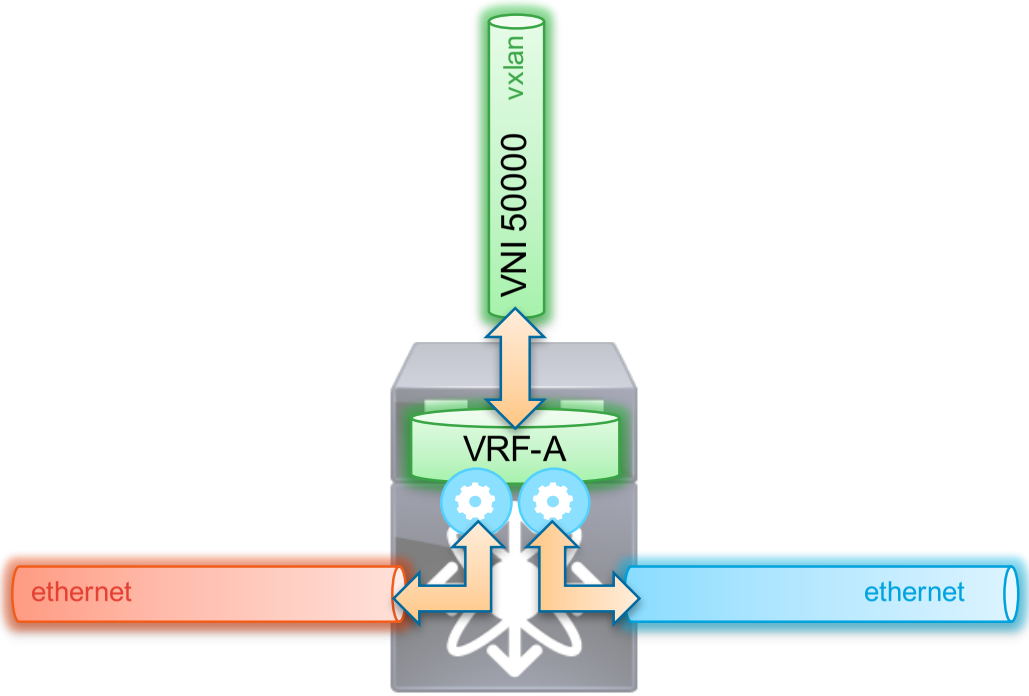
Configuration example



*Requires EVPN Control-Plane.

Routing in VXLAN – define the resources

Overlay



Configuration Example for VRF-A

```
# Define VLAN for VRF routing instance
vlan 2500
  vn-segment 50000

# Define SVI for VRF routing instance
interface Vlan2500
  no shutdown
  mtu 9216
  vrf member VRF-A
  ip forward

# VRF configuration for "customer" VRF
vrf context VRF-A
  vni 50000
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
```

VLAN to Layer-3 VNI mapping

VLAN to Layer-3 VNI mapping
- ip forward required for prefix-based routing

VRF context definition
- VNI
- Route-Distinguisher
- Route-Targets
- IPv4 and/or IPv6

FYI

Distributed IP Anycast Gateway*

Overlay

Configuration Example for "BLUE" (V^1 & V^3)

```
# Features & Globals
feature interface-vlan
fabric forwarding anycast-gateway-mac 2020.DEAD.BEEF

# VLAN to VNI mapping (MT-Lite)
vlan 43
 vn-segment 30000

# Anycast Gateway MAC, inherited by any interface
(SVI) using "fabric forwarding"
fabric forwarding anycast-gateway-mac 0002.0002.0002

# Distributed IP Anycast Gateway (SVI)
interface vlan 43
 no shutdown
 vrf member VRF-A
 ip address 11.11.11.1/24 tag 12345
 fabric forwarding mode anycast-gateway
```

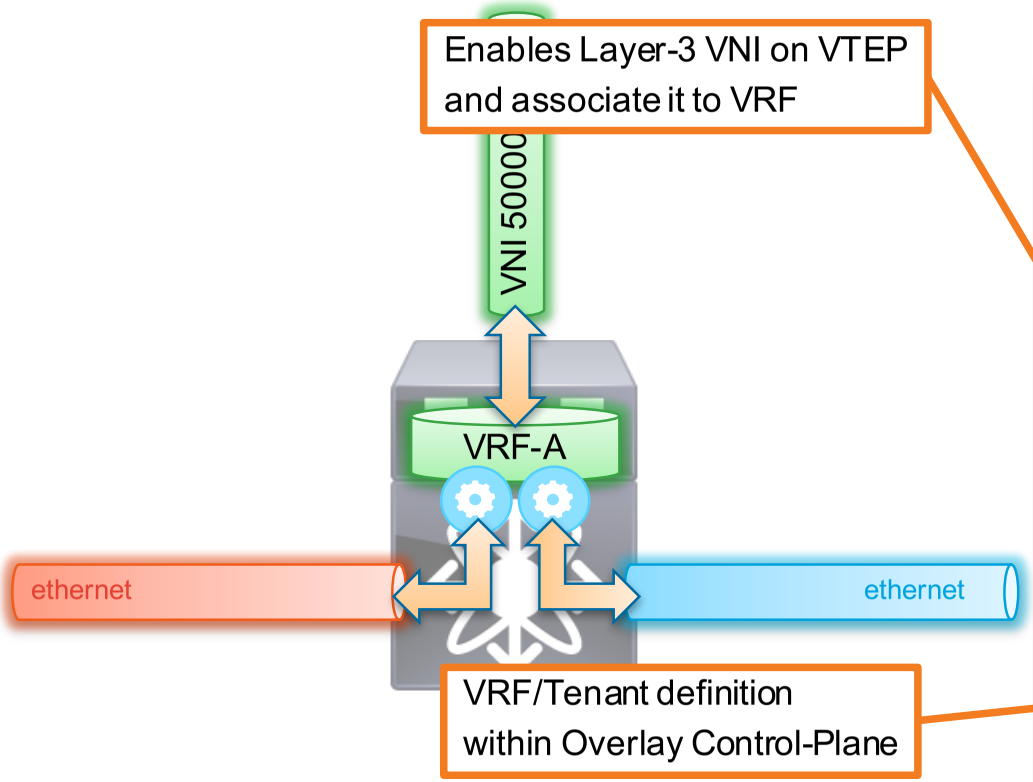
Configuration Example for "RED" (V^{1-3})

```
# Features & Globals
feature interface-vlan
fabric forwarding anycast-gateway-mac 2020.DEAD.BEEF

# VLAN to VNI mapping (MT-Lite)
vlan 55
 vn-segment 30001

# Anycast Gateway MAC, inherited by any interface
(SVI) using "fabric forwarding"
fabric forwarding anycast-gateway-mac 0002.0002.0002

# Distributed IP Anycast Gateway (SVI)
interface vlan 55
 no shutdown
 vrf member VRF-A
 ip address 98.98.98.1/24 tag 12345
 fabric forwarding mode anycast-gateway
```

*Requires EVPN Control-Plane. VRF and BGP configuration not shown

Routing in VXLAN – configure the routing

Overlay



Configuration Example for VRF-A

```
# Activate Layer-3 VNI on VTEP
interface nve1
  source-interface loopback0
  host-reachability protocol bgp
  member vni 30000
    mcast-group 239.239.239.100
    suppress-arp
  member vni 50000 associate-vrf

# Route-Map for Redistribute Subnet
route-map REDIST-SUBNET permit 10
  match tag 12345

# Control-Plane configuration for VRF (Tenant)
router bgp 65500
…
  vrf VRF-A
    address-family ipv4 unicast
      advertise l2vpn evpn
      redistribute direct route-map REDIST-SUBNET
      maximum-paths ibgp 2
```

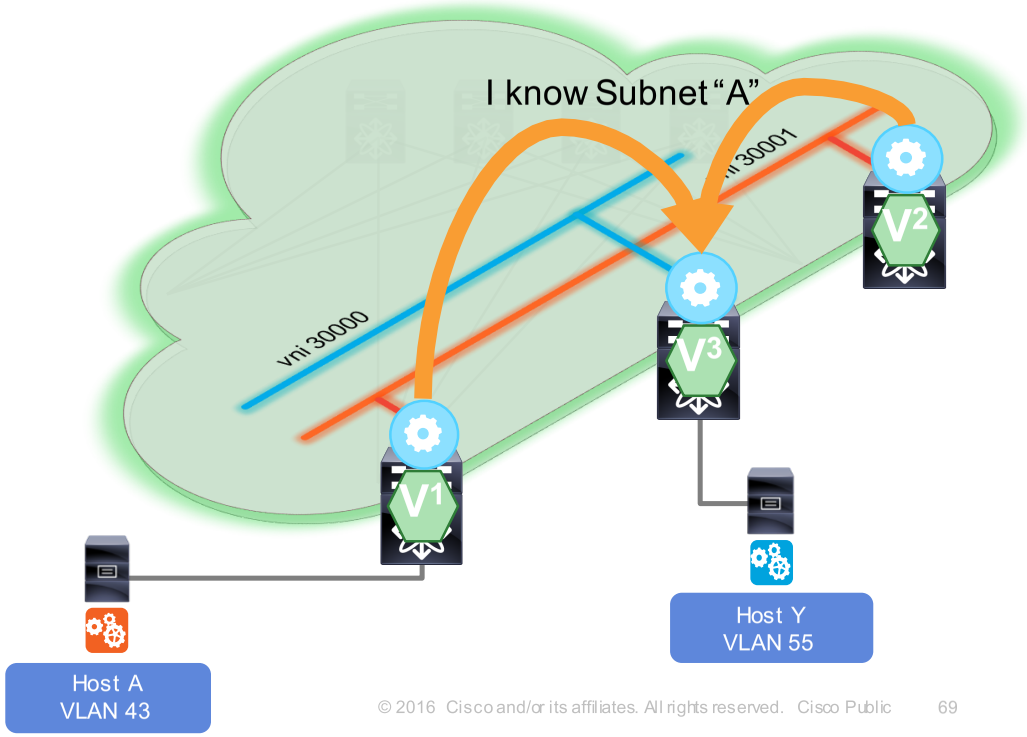
Why redistribution of the Gateway Subnet (Direct)?

Isn't the Host-Route enough?

Host Subnet Redistribution

Overlay

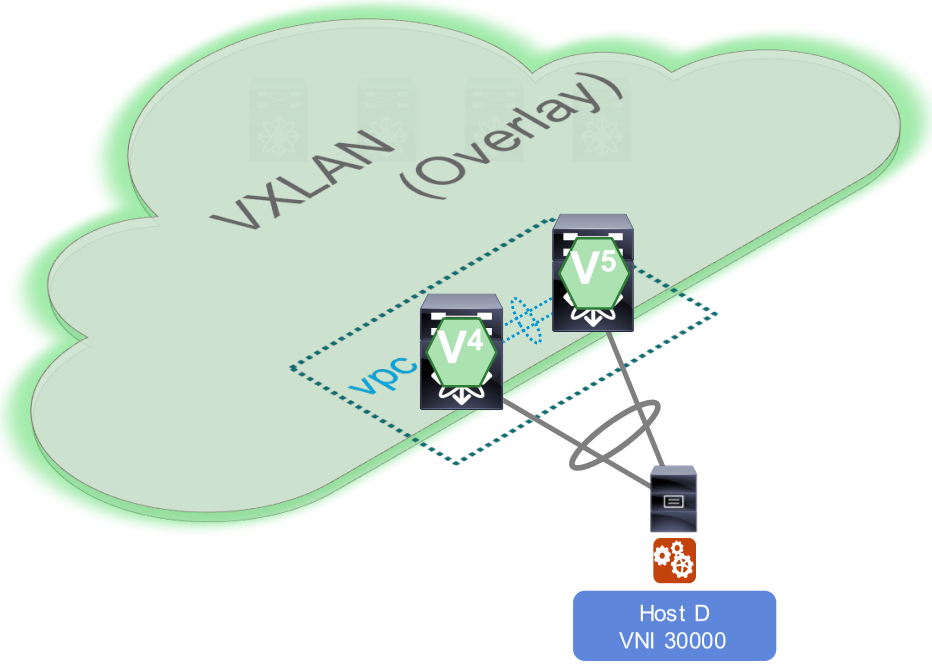
- Host “A” is a silent Host
  - Not known via ARP/IP
- How can Host “Y” reach Host “A”
  - Host “A” and “Y” are in different VLAN/Subnet
- Route for Host “A”-Subnet will be advertised by V^1 and V^2
- Host “Y” will reach either V^1 or V^2 based on ECMP
- From V^1 or V^2, Host “A” can be reached via Layer-2 Segment.



VXLAN Hardware VTEP Redundancy (vPC)

Southbound Connectivity

- VXLAN vPC Domain Configuration Classic Ethernet
- Configure VXLAN specific vPC Peer-Link Configuration
- Extend the IP Interface (Loopback) configuration for the VTEP
  - Secondary IP address (anycast) is used as the anycast VTEP address
  - Both vPC VTEP switches need to have the identical secondary IP address configured under the loopback interface



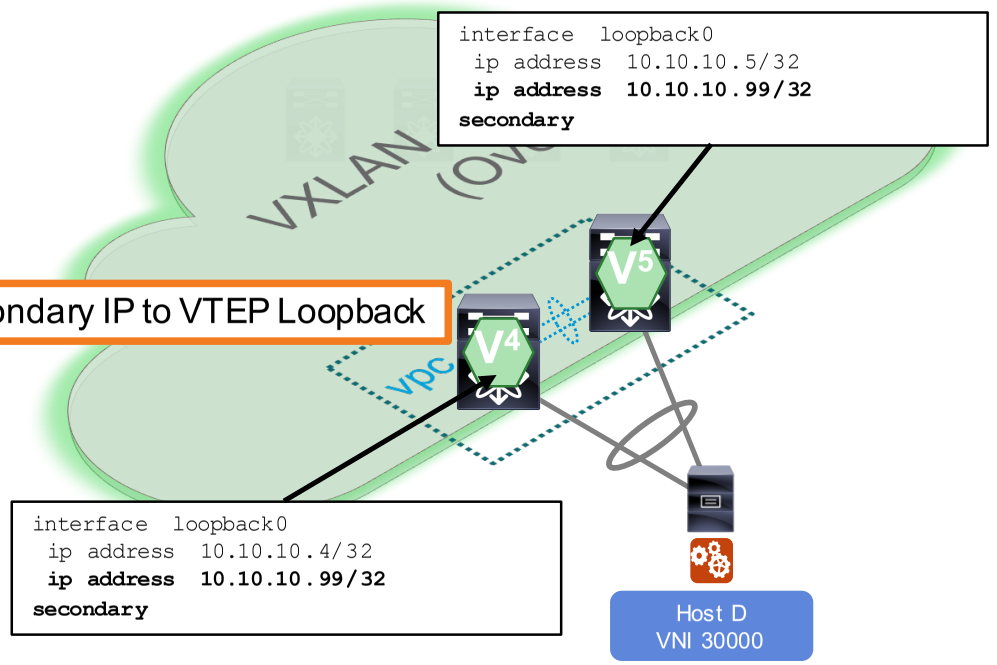
VXLAN Hardware Gateway Redundancy (vPC)

Southbound Connectivity

vPC VTEP Configuration Example for (V4-5)

```
# VLAN to VNI mapping (MT-Lite)
vlan 55
 vn-segment 30000

# VTEP IP Interface; Source/Destination for all
VXLAN Encapsulated Traffic.
▪ Primary IP address is used for Orphan Hosts
▪ Secondary IP is for vPC Hosts (same IP on both
  vPC Peers)
interface loopback0
 ip address 10.10.10.V/32
 ip address 10.10.10.VAnycast/32 secondary

# VTEP configuration using Loopback as source.
Destination Group for VNI 30001 is "239.1.1.2"
interface nve1
  source-interface loopback0
  host-reachability protocol bgp
  member vni 30000
    mcast-group 239.239.239.100
    suppress-arp
  member vni 50000 associate-vrf
```

Add Secondary IP to VTEP Loopback



VXLAN Hardware Gateway Redundancy (vPC)

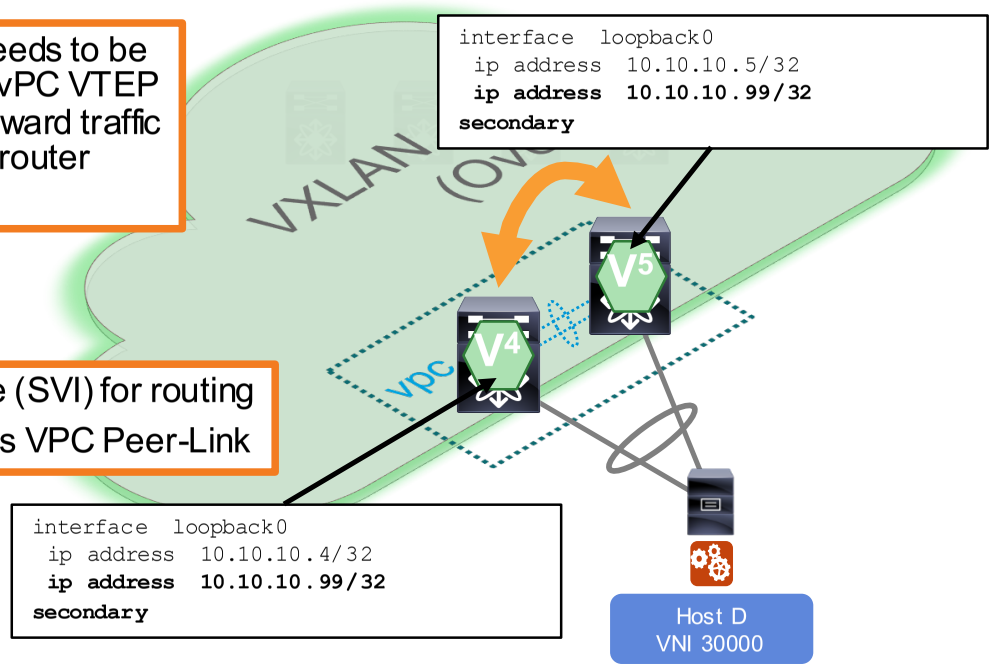
Not to Forget!

vPC VTEP Configuration Example for (V^{4-5})

```
# VPC Domain Configuration
vpc domain 99
  peer-switch
  peer-keepalive destination V4-mgmt sour
  peer-gateway
  ip arp synchronize

# VPC Peer-Link
interface port-channelXX
  switchport mode trunk
  vpc peer-link

# VPC Domain Routing Adjacency
interface Vlan3999
  no shutdown
  ip address 10.254.254.1/30
  ip router ospf 1 area 0.0.0.0
  ip ospf network point-to-point
  ip pim sparse-mode
```

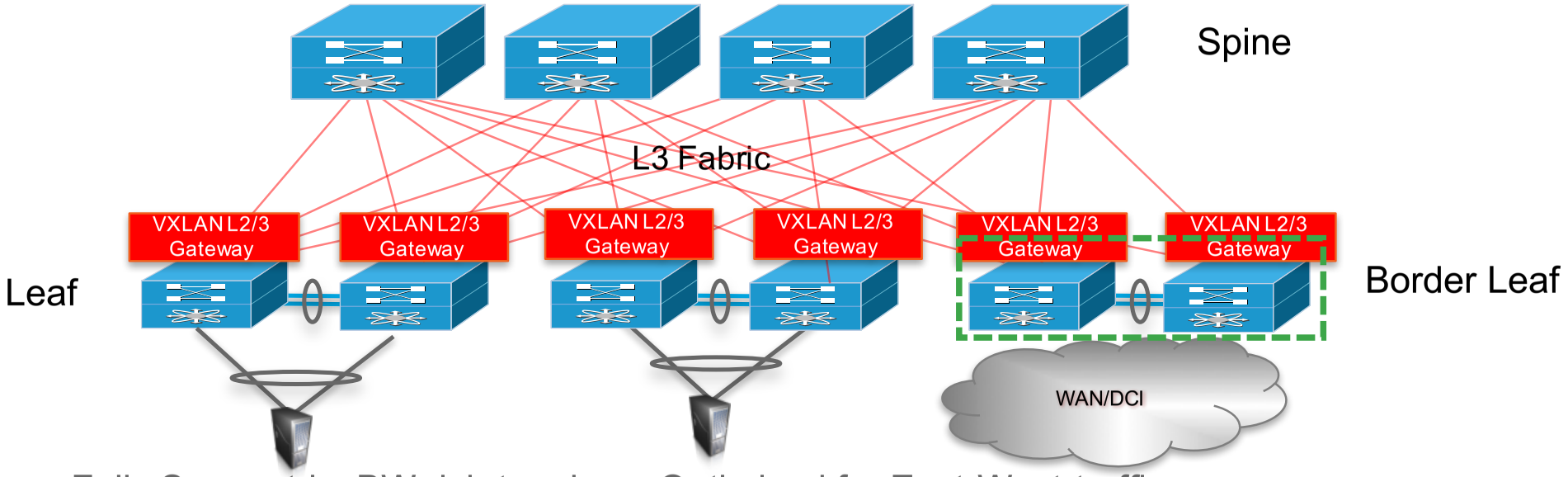
peer-gateway needs to be enabled so that vPC VTEP switches can forward traffic for each other's router MAC address

Routed Interface (SVI) for routing adjacency across VPC Peer-Link



Folded Clos Topology

Providing Topology Symmetry

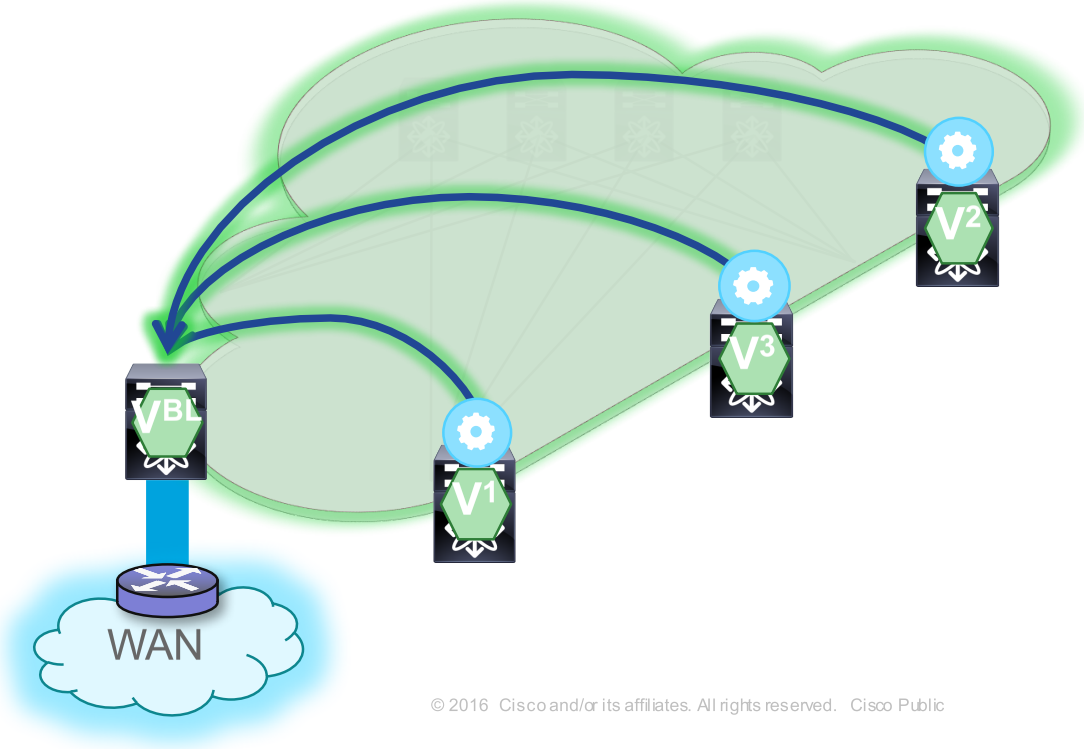


- Fully Symmetric, BW rich topology, Optimized for East-West traffic
- Lean Spine does not do any VXLAN termination/gateway
- Access to other networks through border leaf block

VXLAN/EVPN Fabric External Routing

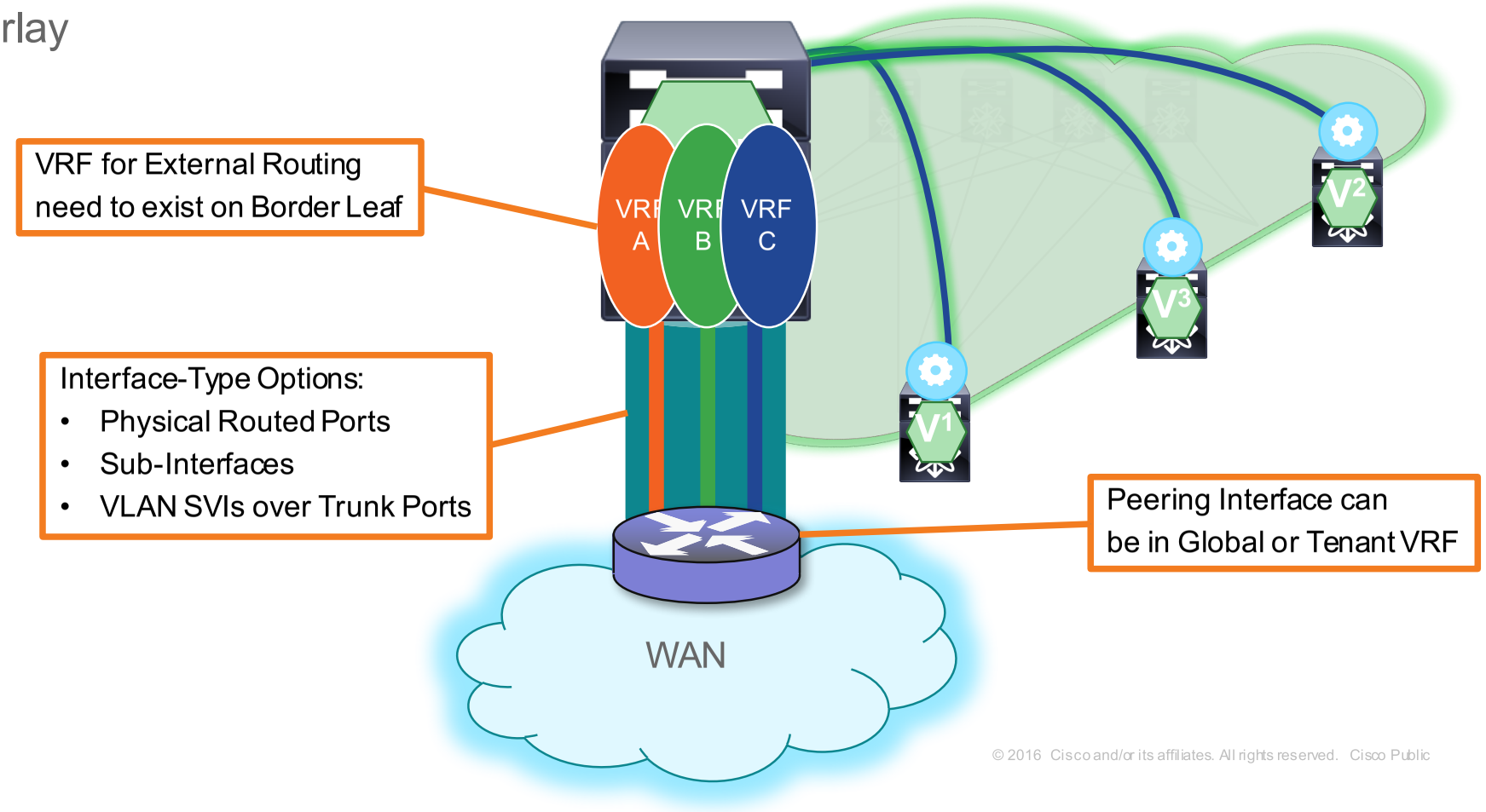
Overlay

- The Border Leaf/Spine provides Layer-2 and Layer-3 connectivity to external Network
- Flexible routing protocol options for for external routing
- Today, VRF-lite allows to extend VRF outside of the fabric
- With Nexus 7000/7700 and F3, other options are coming



VXLAN/EVPN Fabric External Routing

Overlay



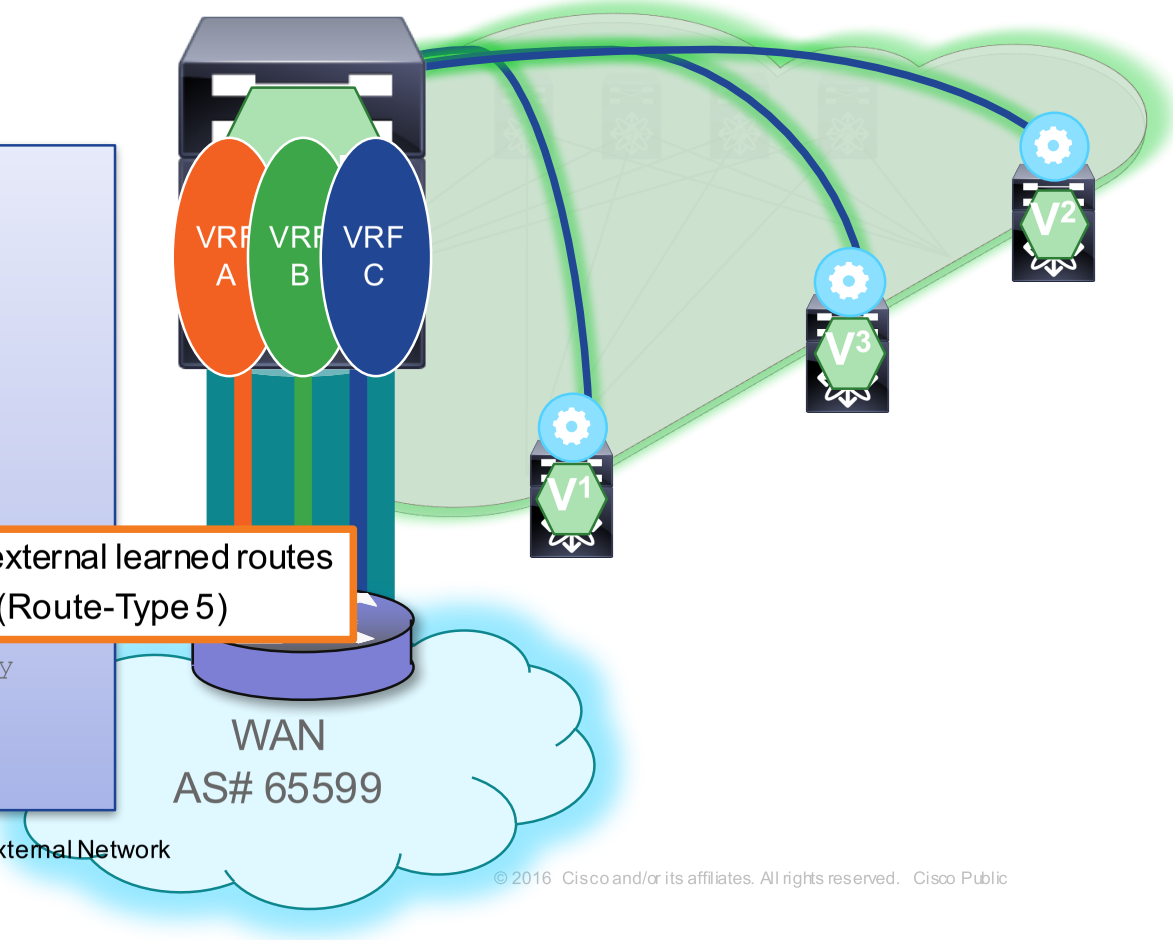
VXLAN/EVPN Fabric External Routing (eBGP)

Overlay

Border Leaf Configuration Example

```
# Sub-Interface Configuration
interface Ethernet1/1
  no switchport
interface Ethernet1/1.10
  mtu 9216
  encapsulation dot1q 10
  vrf member VRF-A
  ip address 10.254.254.1/30

# eBGP Configuration
router bgp 65500
…
  vrf VRF-A
    address-family ipv4 unicast
      advertise l2vpn evpn
      aggregate-address 10.0.0.0/8 summary-only
    neighbor 10.254.254.2 remote-as 65599
      update-source Ethernet1/1.10
      address-family ipv4 unicast
*
```

Advertise external learned routes into EVPN (Route-Type 5)



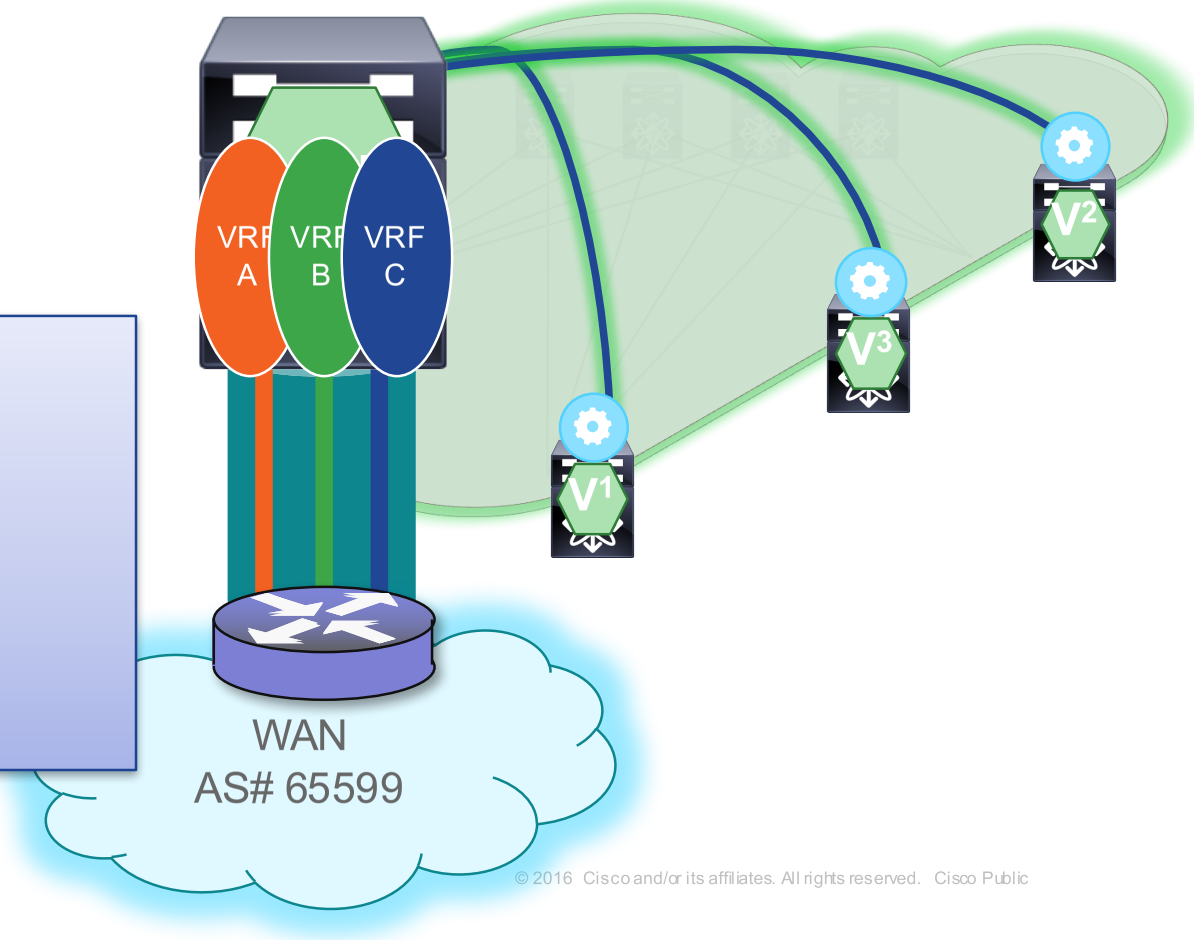
*Ensure that non-necessary routes are not advertised towards the External Network

VXLAN/EVPN Fabric External Routing (eBGP)

Overlay

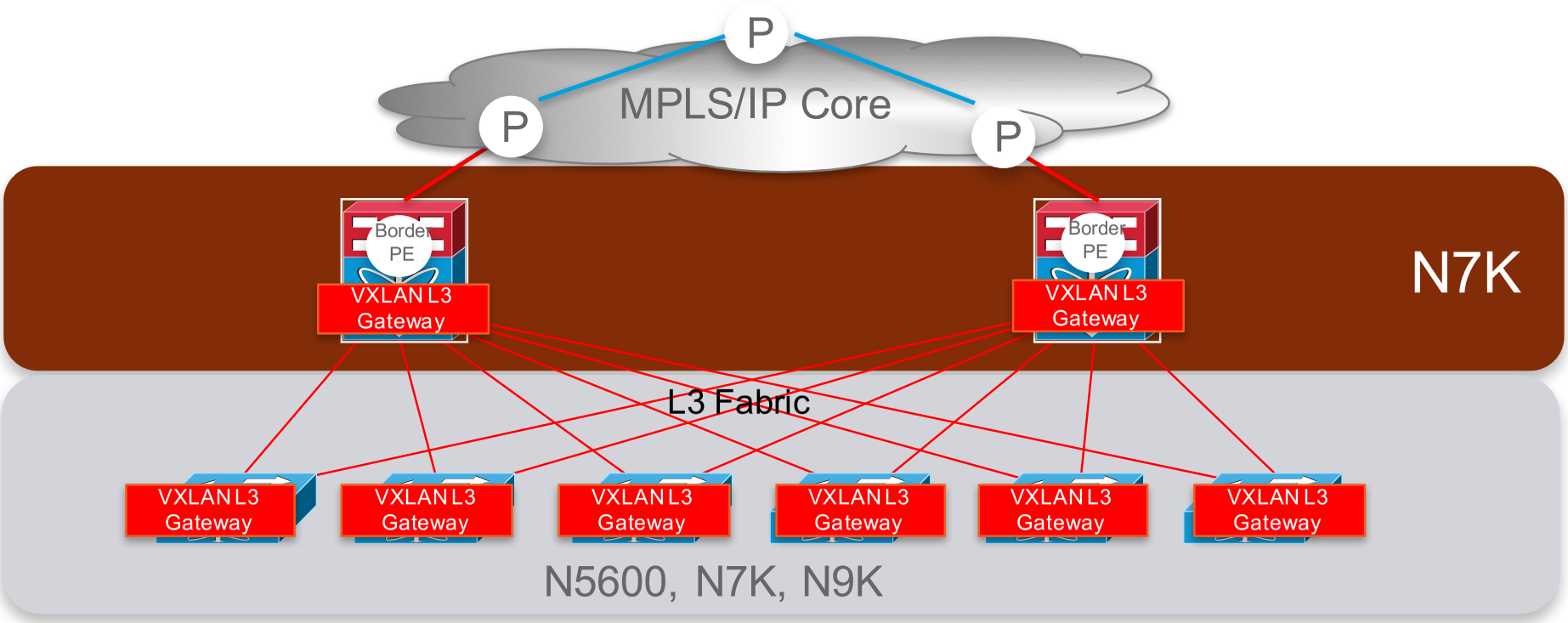
Edge Router Configuration Example

```
# Interface Configuration
interface Ethernet1/1
  vrf member VRF-A
  ip address 10.254.254.2/30

# eBGP Configuration
router bgp 65599
…
  vrf VRF-A
    address-family ipv4 unicast
    neighbor 10.254.254.2 remote-as 65500
      update-source Ethernet1/1
      address-family ipv4 unicast
```



MPLS IP-VPN & VXLAN – in NX-OS 7.3

L3 Handoff – Border PE



Summary and Conclusion

Summary recommendations & takeaways

- Optimize the location of L2 and L3 GWYs to optimize routing and minimize failure exposure
- Leverage L3 VXLAN services enabled by control protocols as the main service and L2 extensions as the exception
- Design the underlay with the VXLAN overlay in mind
- Design the network hierarchically: both the underlay as well as the overlay
- L3 Gateways are key to a sound overlay design
- A combination of pull protocols and push protocols may render optimal scale and resiliency
- Link the provisioning of the overlay and scoping of VNIs to the host orchestration system for optimal scale

Thank you

We’re ready. Are you?