
TOMORROW
starts here.
CISCO
Cisco live!

Building Simplified, Automated and Scalable Data Center Networks with Overlays (VXLAN/FabricPath)

BRKDCT-3378

Lukas Krattiger, Technical Marketing Engineer

Lukk@cisco.com

@CCIE21921

Cisco live!

Session Objectives

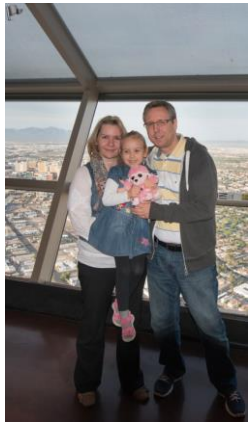
- Focus on Data Center Networks and Fabrics with Overlays
- Closer Look on Packet Encapsulation (VXLAN)
  - Underlay – the Transport for the Overlay
  - Control-Plane – Exchanging Information
  - Optimizing the Forwarding
- Overview on Frame Encapsulation (FabricPath)
  - Similarities to Packet Encapsulation
- Fabric Management & Automation approaches

Session Non-Objectives

- Deep-Dive into FabricPath
  - There are many Sessions and Recordings
- Comparison between different Orchestration and Management Tools
- Automation Workflows or Services Catalogs

Who am I?

- Born behind the Mountains
- Living in the land of E-Cars and Startups
- Not sure if I'm 100% Human
- Dad, Husband, Geek
- Other People call it Switzerland
- Aka San Francisco Bay Area
- US Immigration calls me "Legal Alien"



“We can NOT solve our Problems with the same Thinking we used when we Created them”

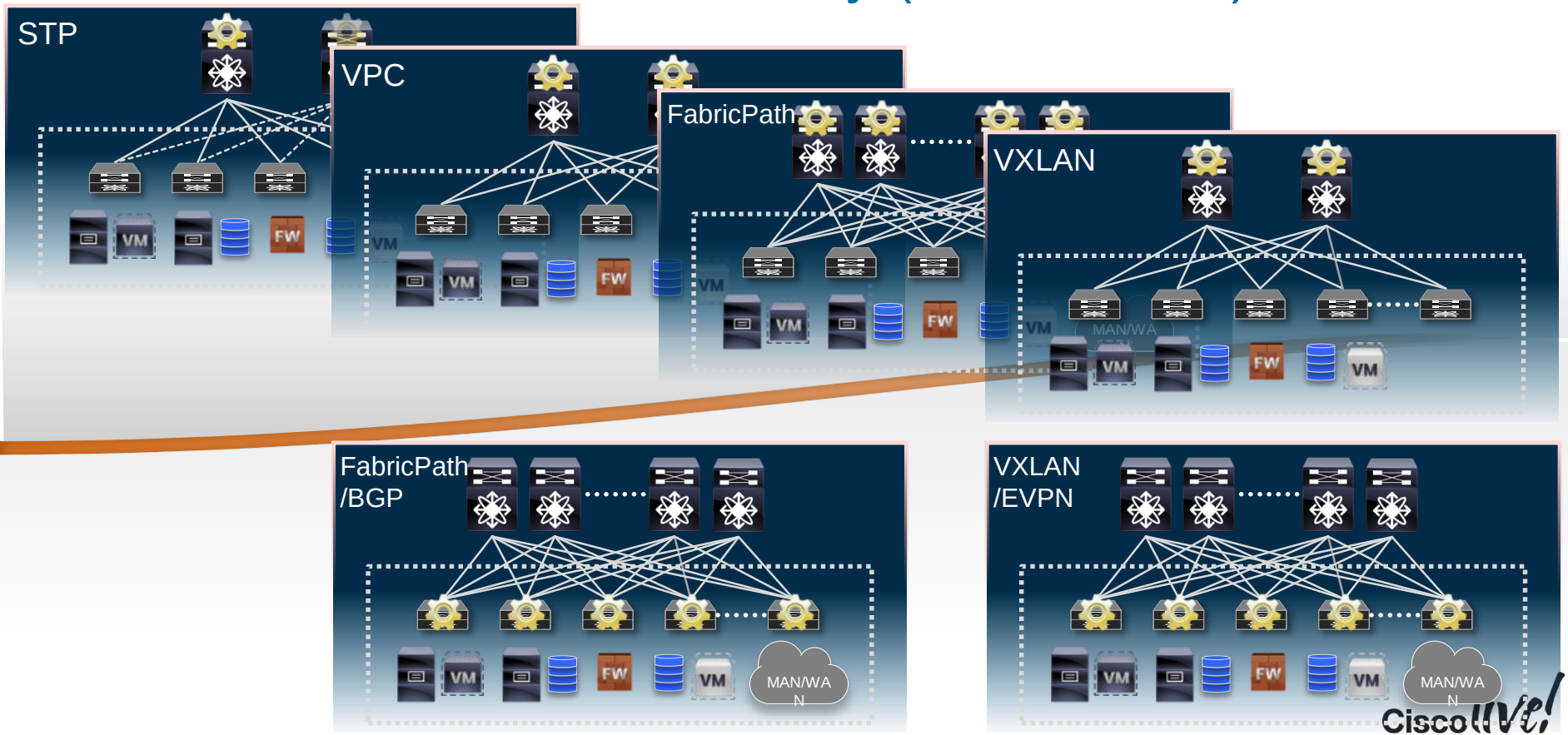
Albert Einstein

Agenda

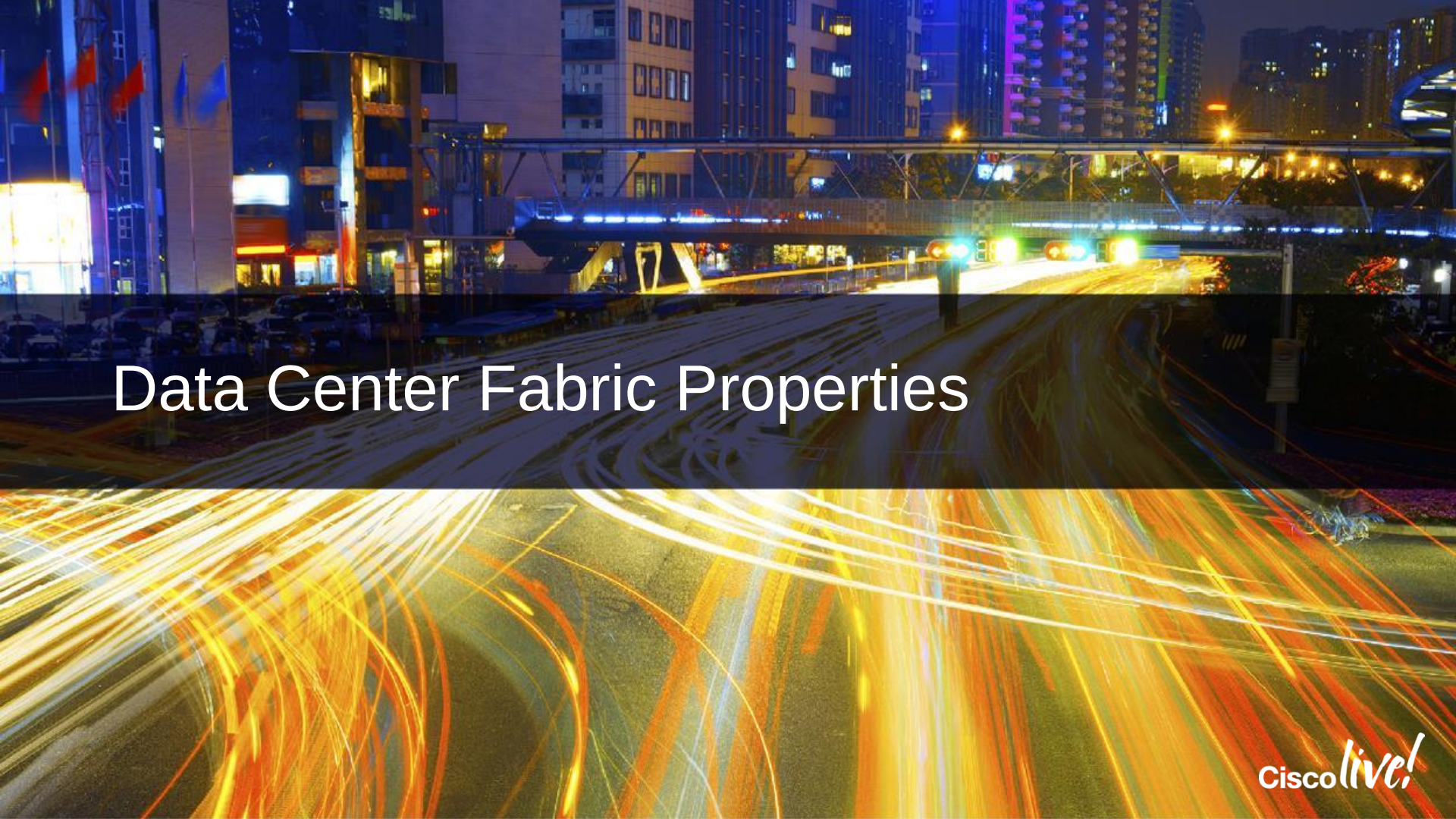
- Data Center Fabric Properties
- Optimized Networking with VXLAN
  - Overview
  - Underlay
  - Control & Data Plane
  - Multi-Tenancy
- Optimized Networking with FabricPath
- Fabric Management & Automation

Cisco live!

Data Center “Fabric” Journey (Standalone)



A Fork in the Road



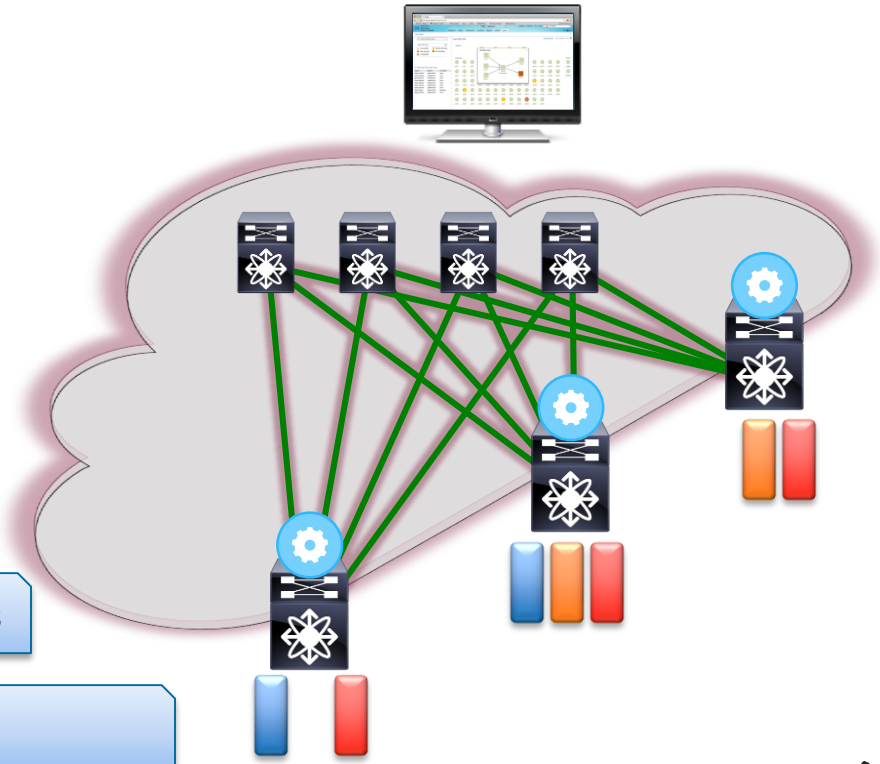
Data Center Fabric Properties

Cisco live!

Data Center Fabric Properties

Scale, Resiliency, Efficiency

- Any subnet, anywhere, rapidly
- Reduced Failure Domains
- Extensible Scale & Resiliency
- Profile Controlled Configuration



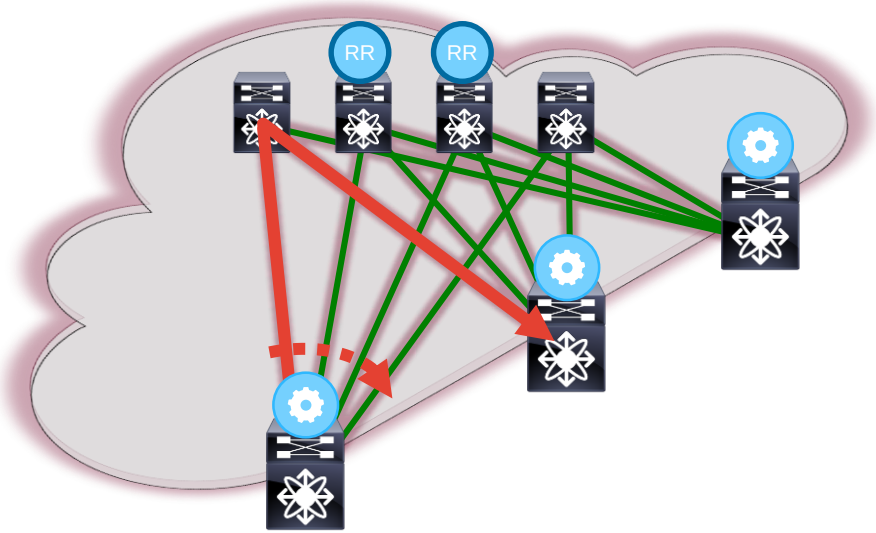
- Full Bi-Sectional Bandwidth (N Spines)
- Any/All Leaf Distributed Default Gateways
- Any/All Subnets on Any Leaf

Cisco live!

Spine/Leaf Topologies

Data Center Fabric Properties

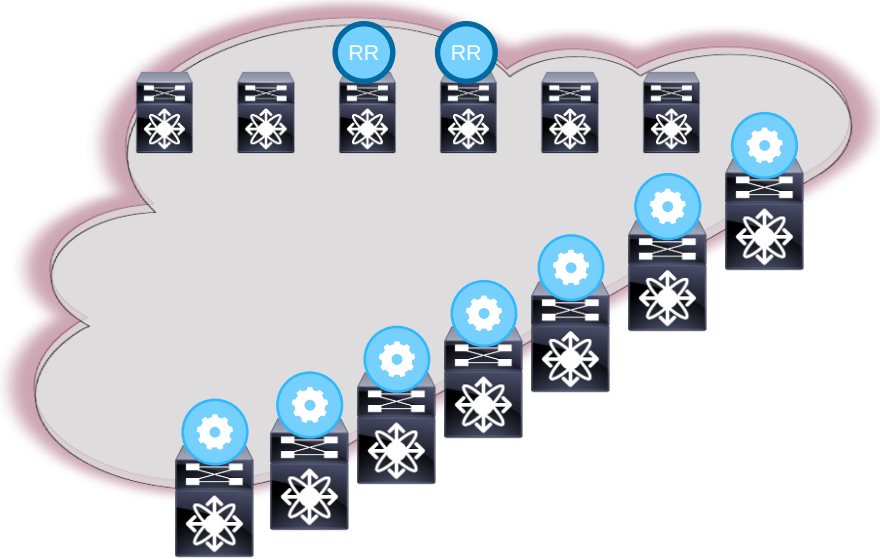
- High Bi-Sectional Bandwidth
- Wide ECMP: Unicast or Multicast
- Uniform Reachability, Deterministic Latency
- High Redundancy: Node/Link Failure
- Line rate, low latency, for all traffic



Variety of Fabric Sizes

Data Center Fabric Properties

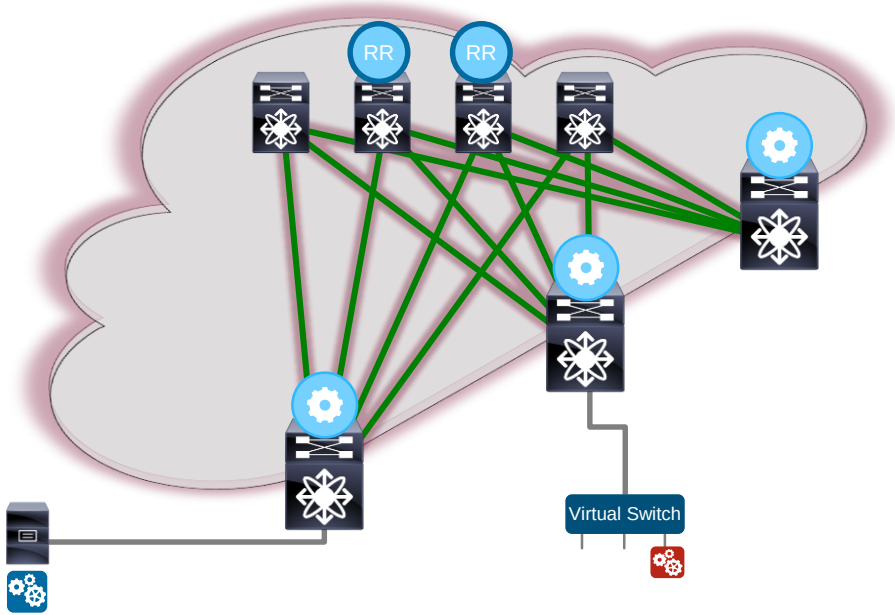
- Fabric size: Hundreds to 10s of Thousands of 10G ports
- Variety of Building Blocks:
  - Varying Size
  - Varying Capacity
  - Desired oversubscription
  - Modular and Fixed
- Scale Out Architecture
  - Add compute, service, external connectivity as the demand grows



Variety of South-bound Topological Connectivity

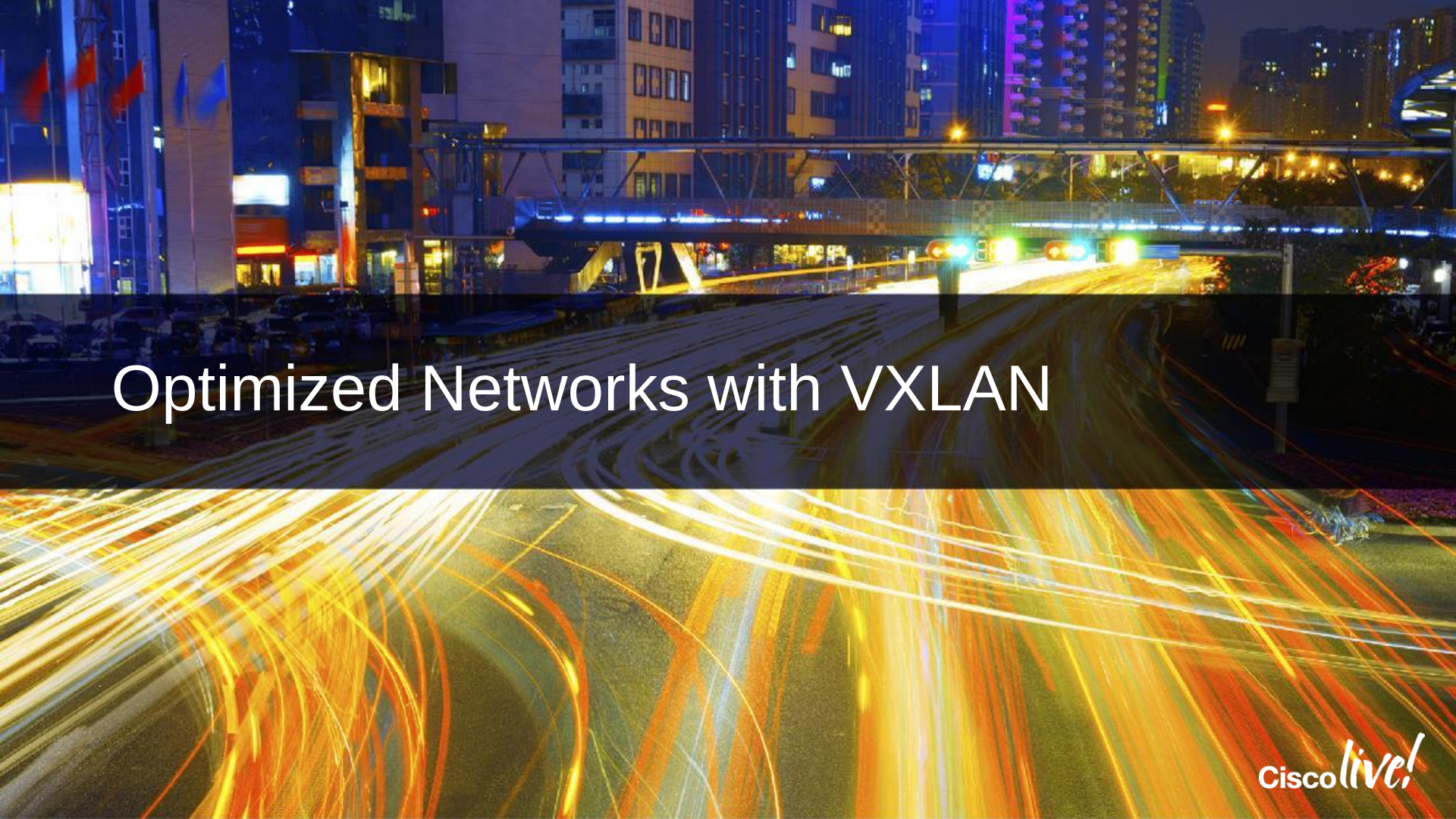
Data Center Fabric Properties

- Flexible connectivity options to the Leaf nodes
  - FEX in straight-through or dual-active mode (eVPC)
  - Blade Switches
  - Hypervisors or bare-metal servers attached in vPC mode



Data Center Fabric Properties

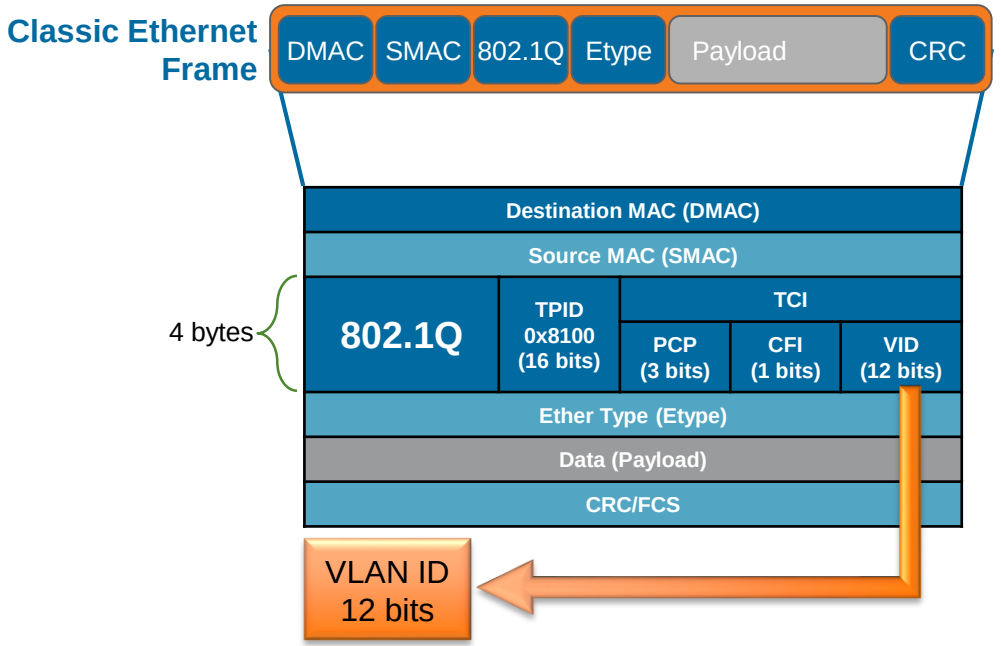
- Extended Namespace
- Scalable Layer-2 Domains
- Integrated Route and Bridge
- ~~Hybrid Overlays~~
- ~~Inter-Pod connectivity~~
- Multi-Tenancy


Optimized Networks with VXLAN
Cisco live!

Overview

Classic Ethernet IEEE 802.1Q Frame Format

- Traditionally VLAN is expressed over 12 bits (802.1Q tag)
  - Limits the maximum number of segments in a Data Center to 4096 VLANs

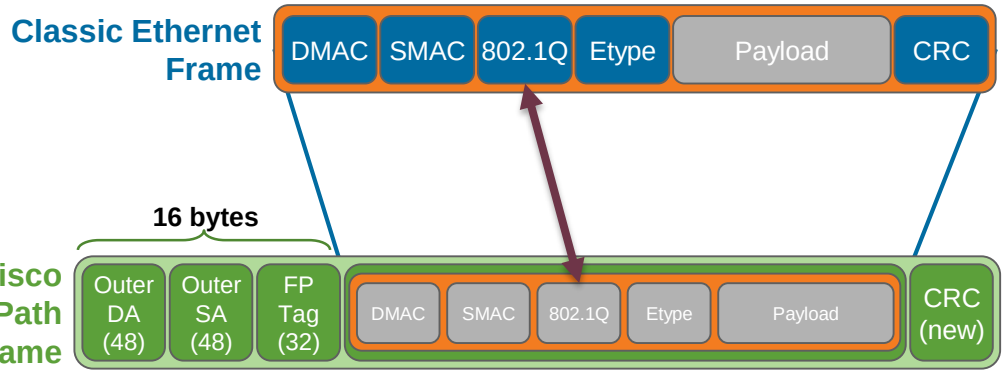


TPID = Tag Protocol Identifier, TCI = Tag Control Information, PCP = Priority Code Point, CFI = Canonical Format Indicator, VID = VLAN Identifier

Overview

FabricPath Encapsulation and Frame Format

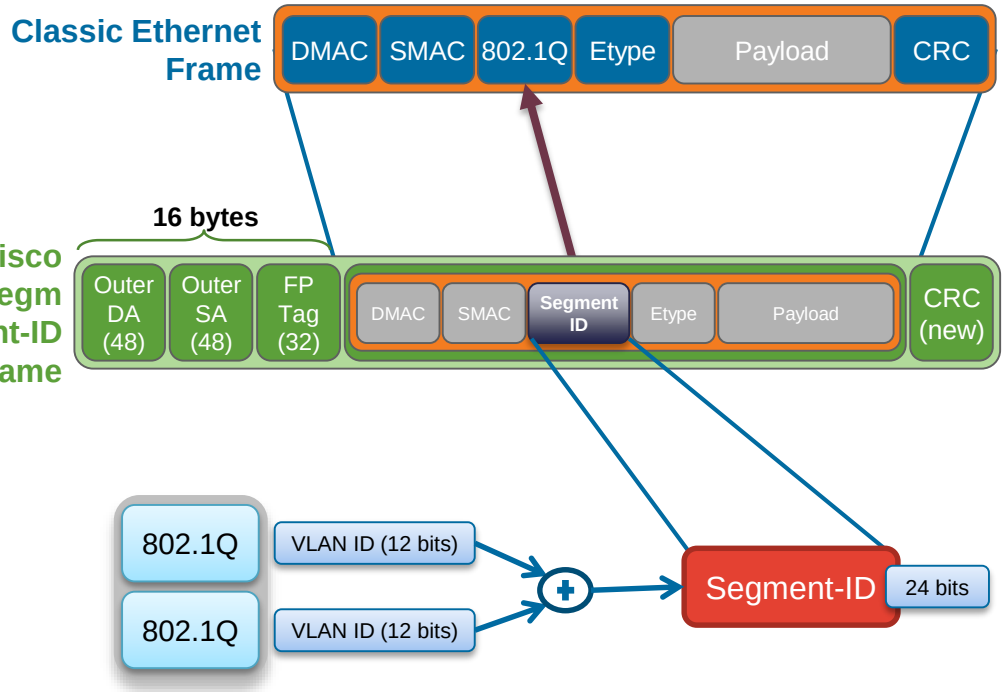
- Cisco FabricPath introduces Classic Ethernet Frame Encapsulation by adding 16 bytes
- Traditionally VLAN, expressed over IEEE 802.1Q tag, stays the same



Overview

Introducing FabricPath w/Segment-ID (1)

- Cisco FabricPath introduces Classic Ethernet Frame Encapsulation by adding 16 bytes
- Traditionally VLAN, expressed over IEEE 802.1Q tag, stays the same
- FabricPath w/Segment-ID leverages a two 802.1Q tags for a total address space of 24 bits
  - Support of ~16M segments



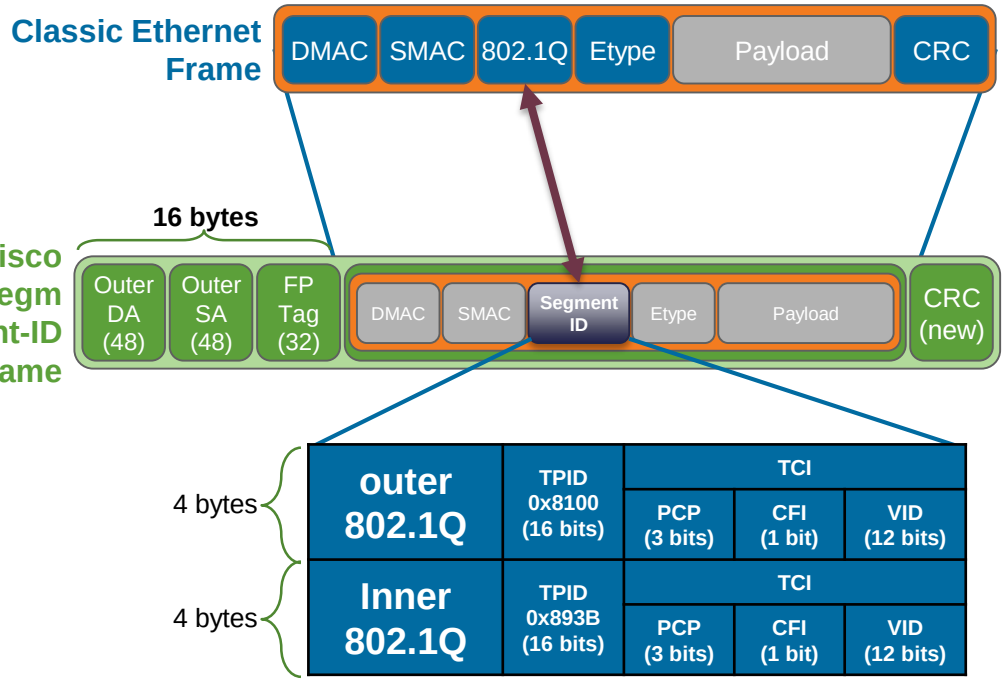
TPID 0x8100 = VLAN Tagged Frame with IEEE 802.1Q / TPID 0x893B = TRILL Fine Grained Labeling

FYI

Overview

Introducing FabricFath w/Segment-ID (2)

- Cisco FabricPath introduces Classic Ethernet Frame Encapsulation by adding 16 bytes
- Traditionally VLAN, expressed over IEEE 802.1Q tag, stays the same
- FabricPath w/Segment-ID leverages a double 802.1Q tag for a total address space of 24 bits
  - Support of ~16M segments

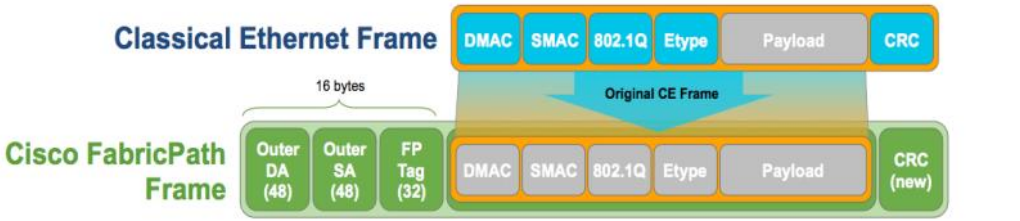


TPID 0x8100 = VLAN Tagged Frame with IEEE 802.1Q / TPID 0x893B = TRILL Fine Grained Labeling

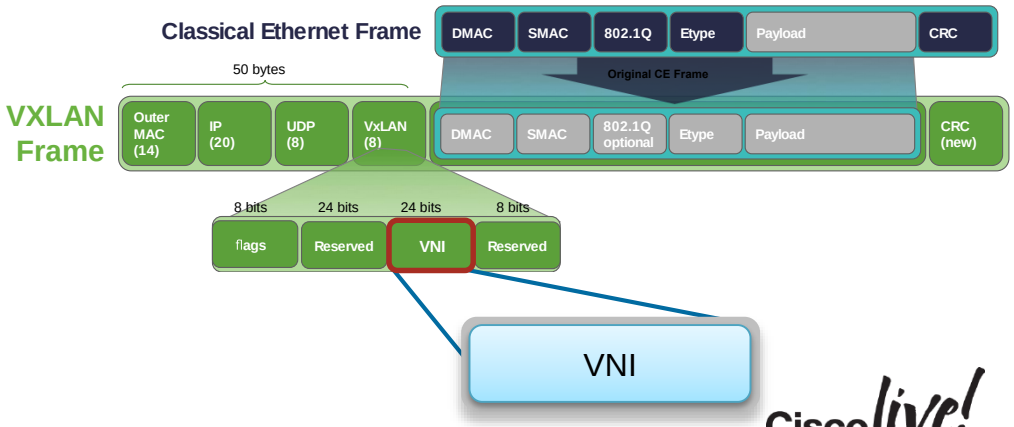
Overview

Introducing VXLAN

- Traditionally VLAN is expressed over 12 bits (802.1Q tag)
  - Limits the maximum number of segments in a Data Center to 4096 VLANs

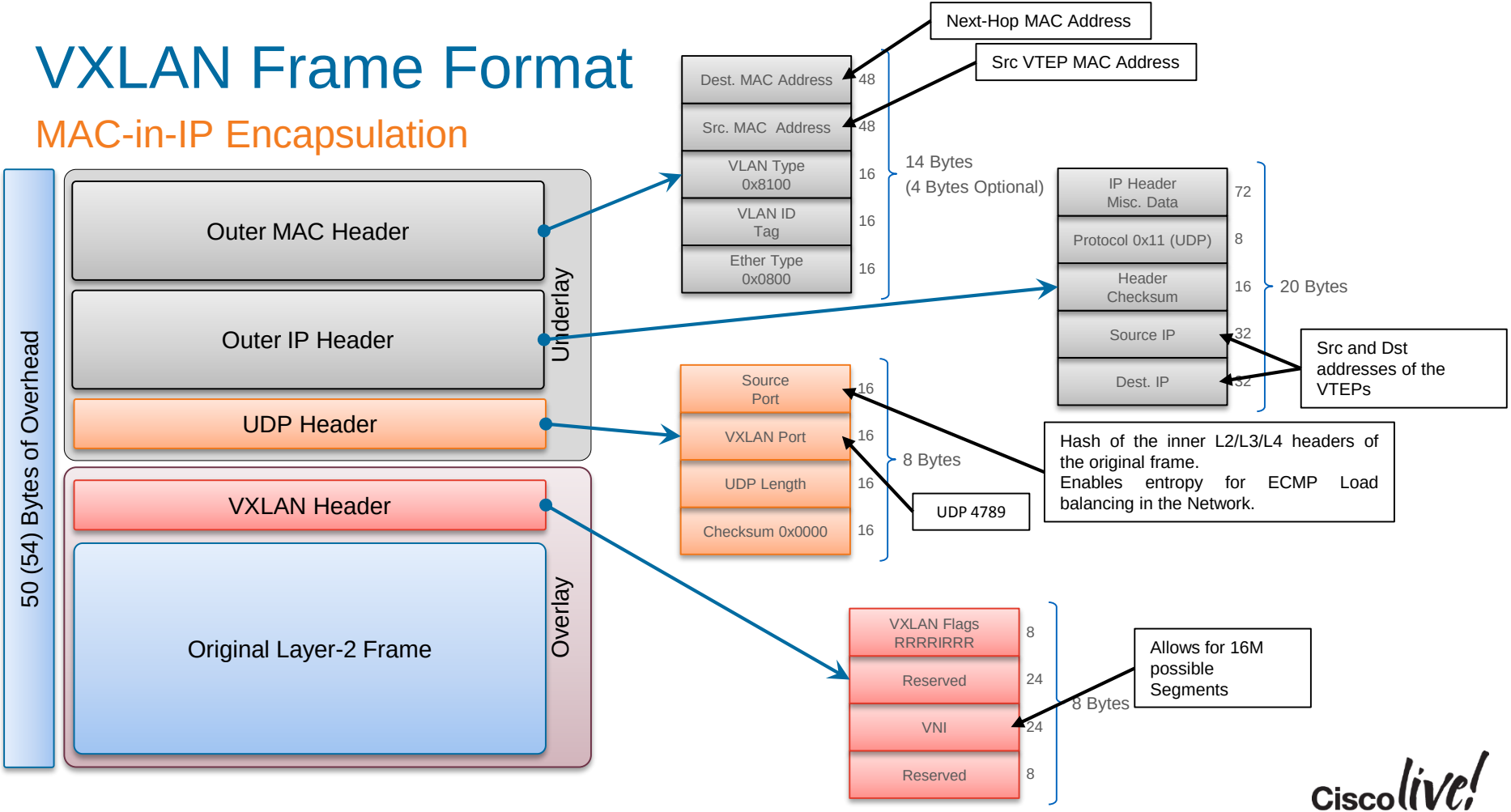


- VXLAN leverages the VNI field with a total address space of 24 bits
  - Support of ~16M segments
- The VXLAN Network Identifier (VNI/VNID) is part of the VXLAN Header



VXLAN Frame Format

MAC-in-IP Encapsulation



Optimized Networks with VXLAN

Data Center Fabric Properties

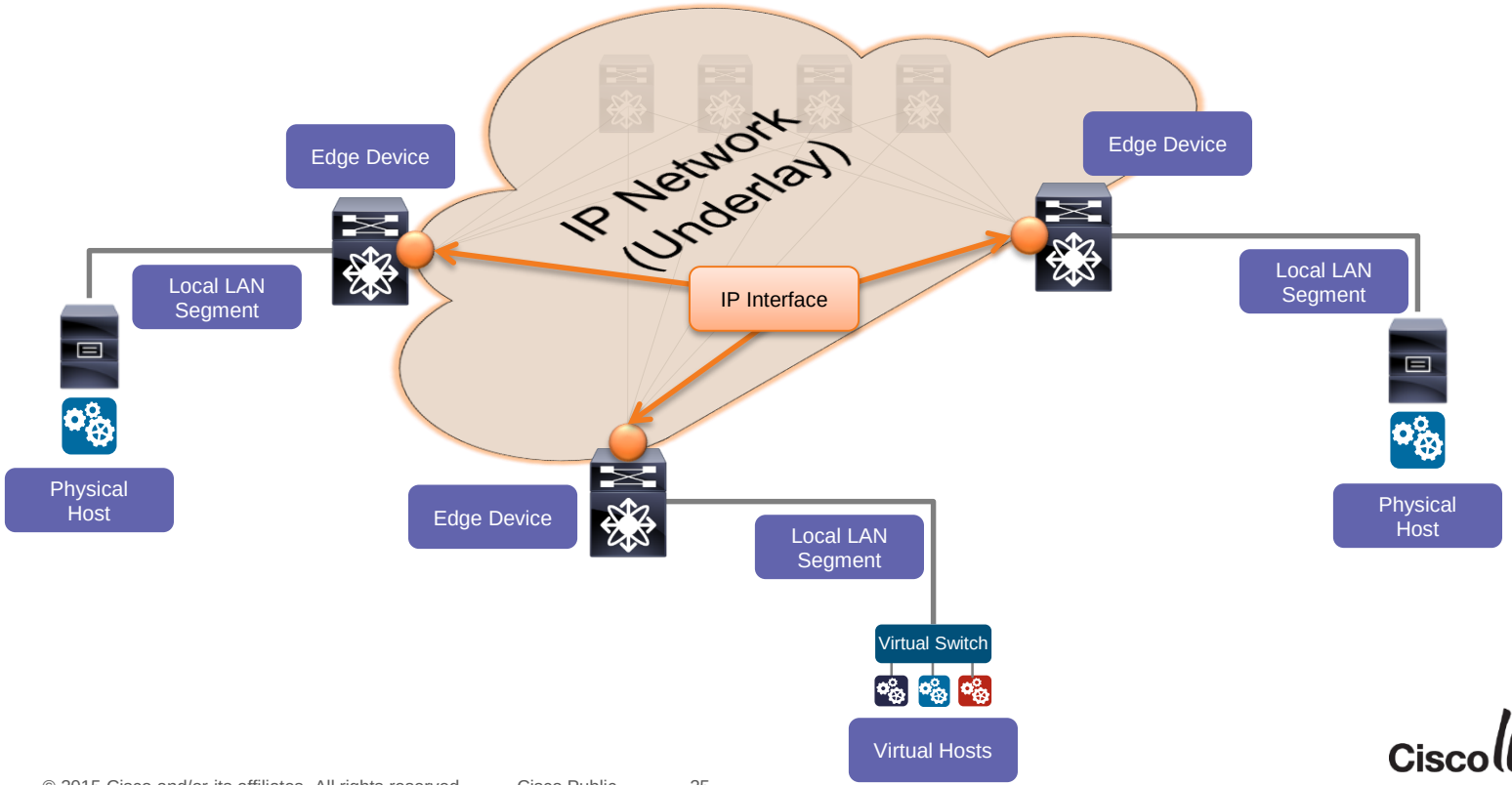
- ✓Extended Namespace
- ❑Scalable Layer-2 Domains
- ❑Integrated Route and Bridge
- ❑Multi-Tenancy

Why VXLAN?

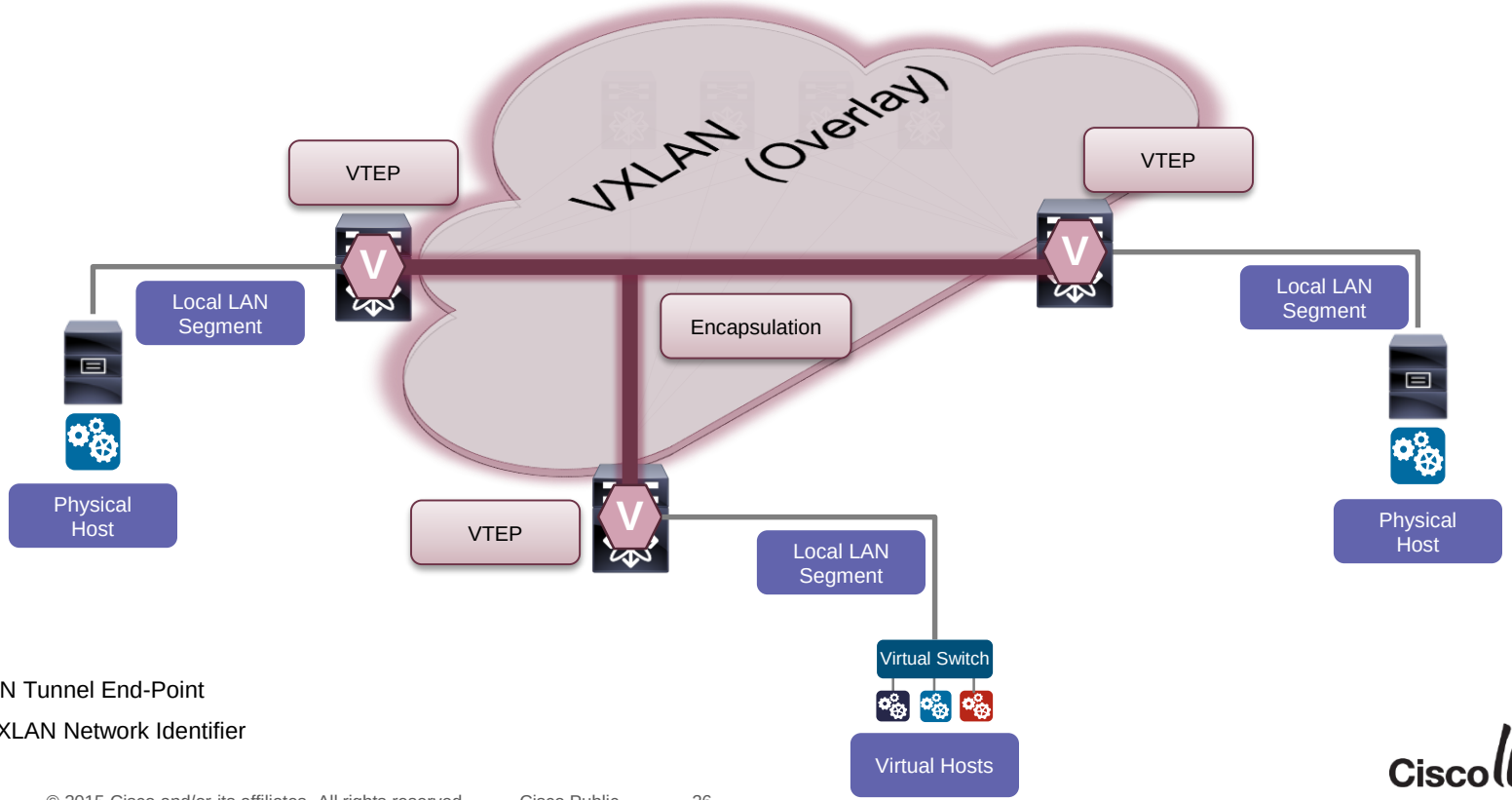
VXLAN provides a Network with Segmentation, IP Mobility, and Scale

- “Standards” based Overlay
- Leverages Layer-3 ECMP – all links forwarding
- Increased Name-Space to 16M identifier
- Integration of Physical and Virtual
- It’s SDN ☺

VXLAN Taxonomy (1)



VXLAN Taxonomy (2)



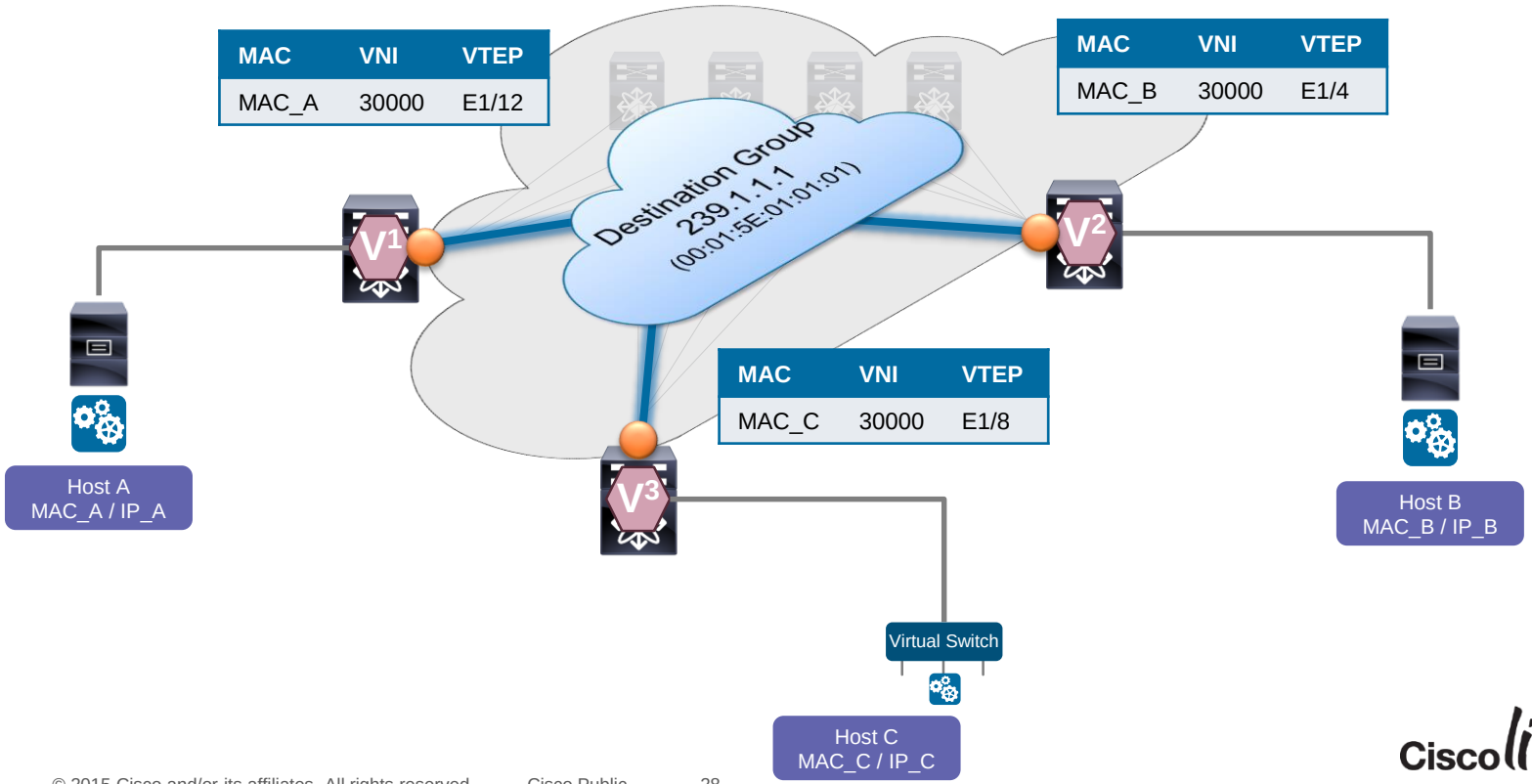
VTEP – VXLAN Tunnel End-Point

VNI/VNID – VXLAN Network Identifier

Session Marketing ☺

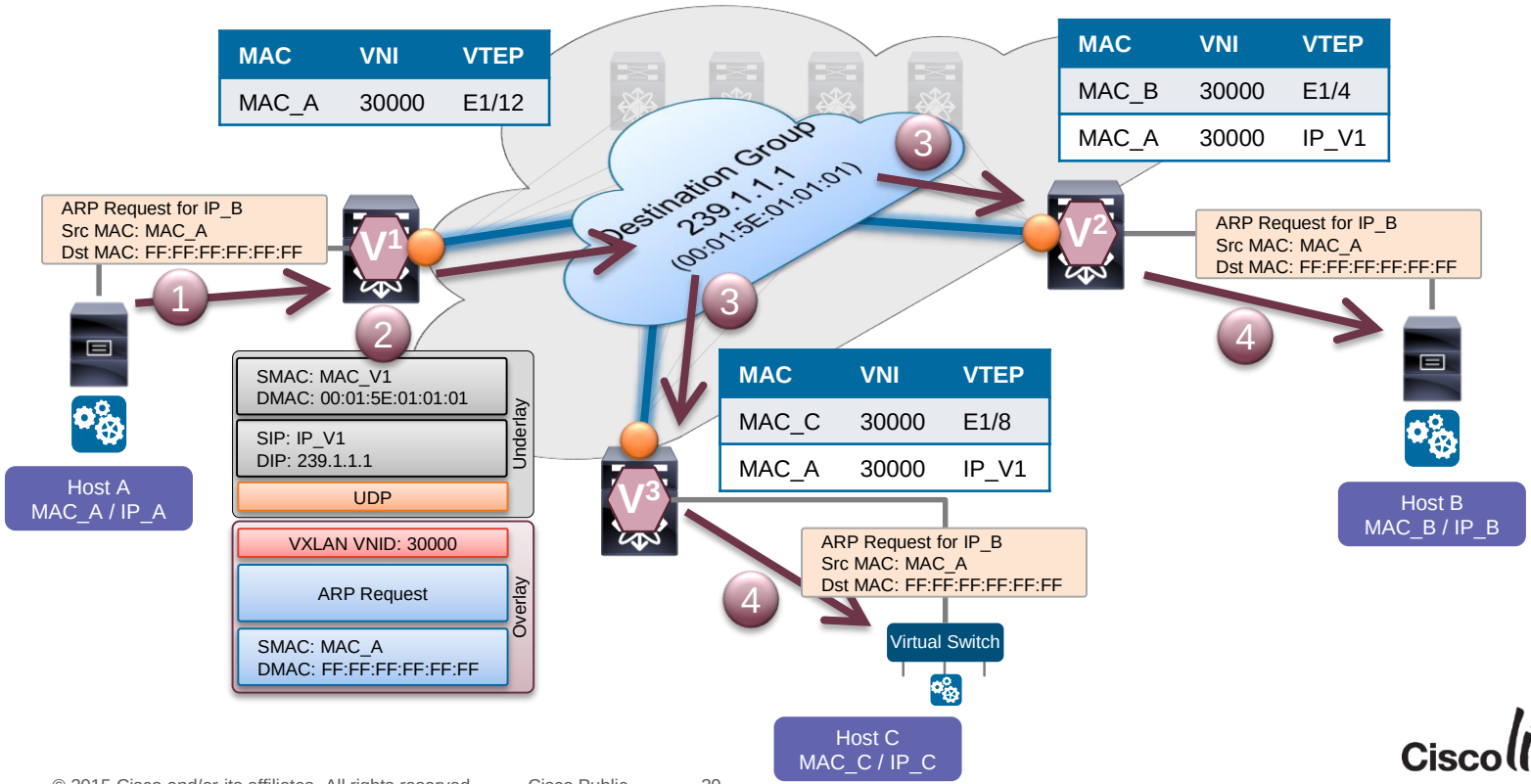
Session ID	Session Title	Speaker	Day / Time
LTRDCT-1224	Implementing VXLAN in Datacenter	Lilian Quan Technical Marketing Engineer	Tuesday, 9:30 Wednesday, 9:00
BRKAPP-9004	Data Center Mobility, VXLAN & ACI Fabric Architecture	Bradley Wong Principal Engineer	Tuesday, 11:15
PNLDCT-2001	Panel: Overlays in the Data Center - A Customer Perspective	Panel	Tuesday, 4:45p
BRKDCT-2328	Evolution of Network Overlays in Data Center Clouds	Victor Moreno Distinguished TME	Wednesday, 11:30
BRKDCT-2456	Building Multi-tenant DC using Cisco Nexus Switching	Elyor Khakimov Technical Marketing Engineer	Wednesday, 2:30p
BRKDCT-2404	VXLAN deployment models - A practical perspective	Victor Moreno Distinguished TME	Thursday, 9:00
BRKACI-2001	Integration and Interoperation of existing Nexus networks into an ACI architecture	Mike Herbert Principal Engineer	Thursday, 11:30
BRKDCT-2049	Overlay Transport Virtualization	Brian Farnham Technical Marketing Engineer	Thursday, 2:30p

VXLAN Flood & Learn



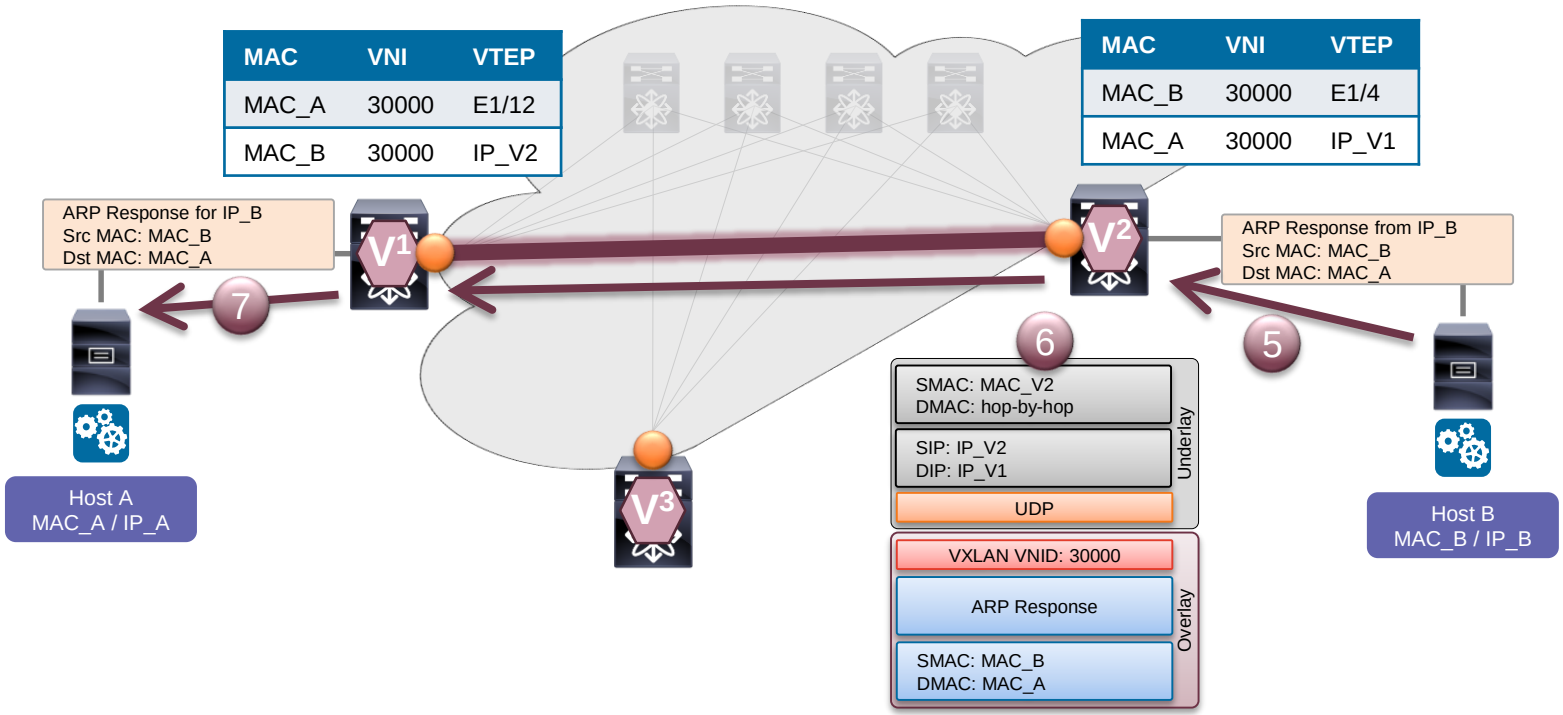
VTEP Peer Discovery & Address Learning (1)

VXLAN Flood & Learn



VTEP Peer Discovery & Address Learning (2)

VXLAN Flood & Learn



VXLAN Flood & Learn and Cisco FabricPath

Compared

		VXLAN	FabricPath
Encapsulation		Packet Encapsulation (PE)	Frame Encapsulation (FE)
Transport Medium Requirement		Layer-3	Layer-1 (mandatory)
End-Host Reachability and Distribution		Flood & Learn	Flood & Learn (+Conversational Learning)
End-Host Detection		Flood & Learn	Flood & Learn
Multi-Destination Traffic (BUM*) forwarding		Multicast (PIM)	FabricPath IS-IS
Underlay Control-Plane		Any Unicast Routing Protocol (static, OSPF, IS-IS, eBGP)	FabricPath IS-IS
Unique Node Identifier		VTEP IP	SwitchID
Standard Reference		RFC 7348	TRILL based (Cisco Proprietary)

*BUM: Broadcast, Unknown Unicast, Multicast

Agenda

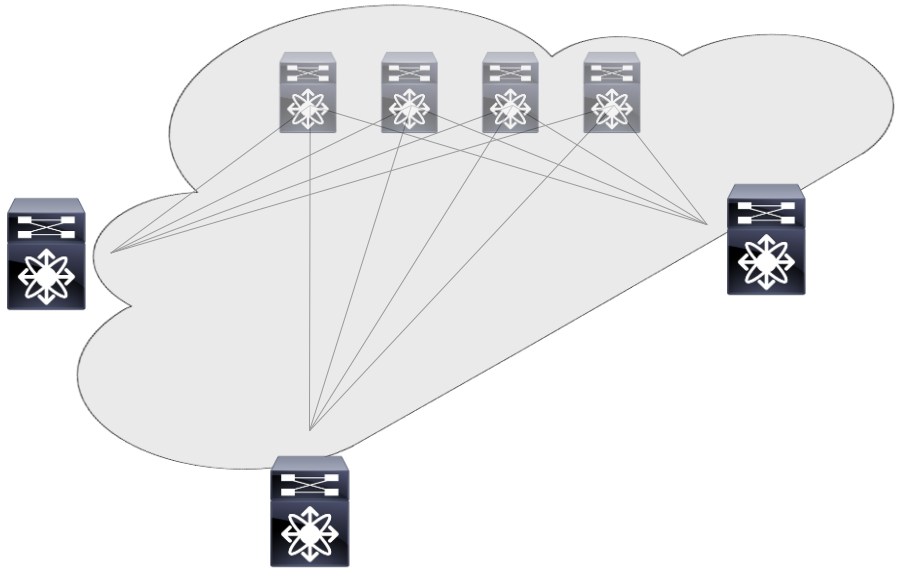
- Data Center Fabric Properties
- Optimized Networks with VXLAN
  - Overview
  - **Underlay**
  - Control & Data Plane
  - Multi-Tenancy
- Optimized Networks with FabricPath
- Fabric Management & Automation

Cisco live!

Deployment Considerations

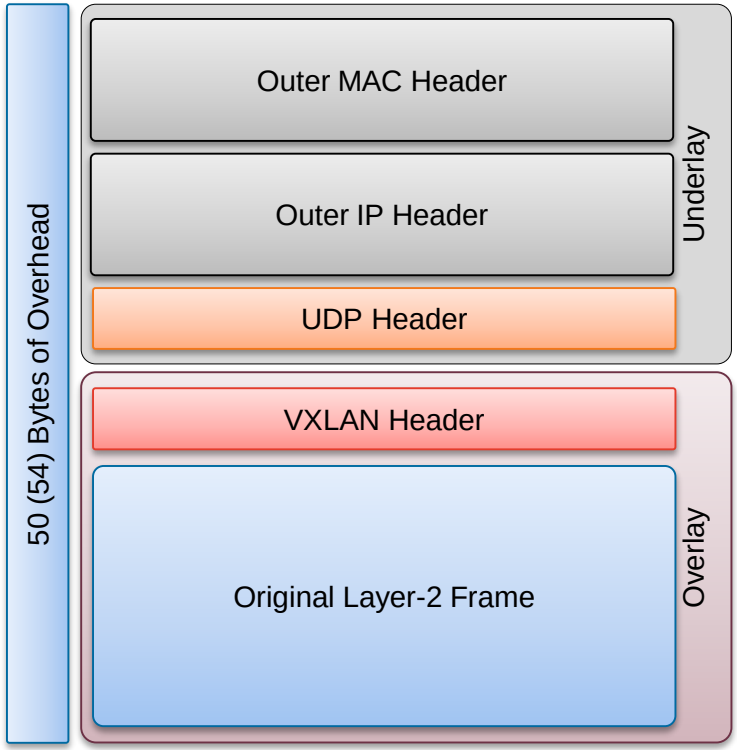
Underlay

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



MTU and VXLAN

Underlay



No Fragmentation Needed

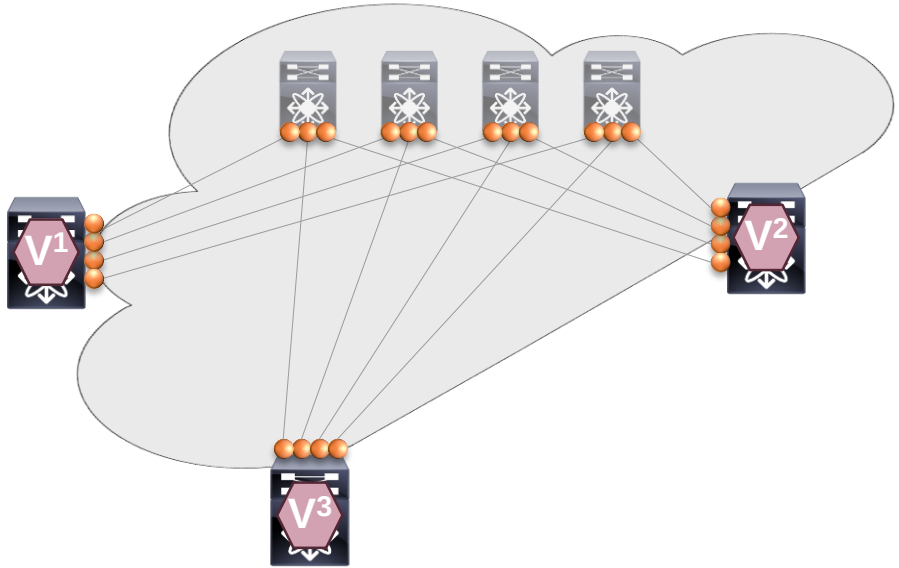
- VXLAN adds 50 Bytes to the Original Ethernet Frame
- Avoid Fragmentation by adjusting the IP Networks MTU
- Data Centers often require Jumbo MTU; most Server NIC do support up to 9000 Bytes
- Using a MTU of 9216* Bytes accommodates VXLAN Overhead plus Server max. MTU

*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



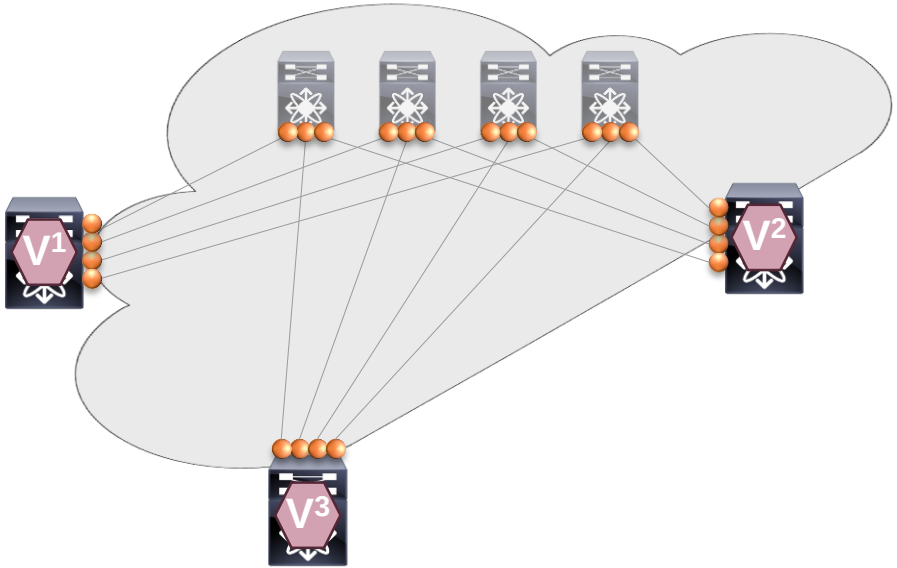
FYI

Building your IP Network – Interface Configuration

Underlay

Interface Configuration Example for (V^1)

```
# Loopback Interface Configuration (VTEP)
interface loopback 0
 ip address 10.10.10.V1/32
 mtu 9192

# Point-2-Point (P2P) Interface Configuration
interface Ethernet 2/1
 no switchport
 ip address 192.168.1.1/31
 mtu 9192

interface Ethernet 2/2
 no switchport
 ip address 192.168.1.3/31
 mtu 9192
.
.
```



Cisco live!

Building your IP Network – Some Math

Underlay

Example from depicted topology:
4 Spine * 3 Leaf = 12 Point-2-Point (P2P) Links
12 Links * 2 (/31) + 3 VTEP + 2 RR

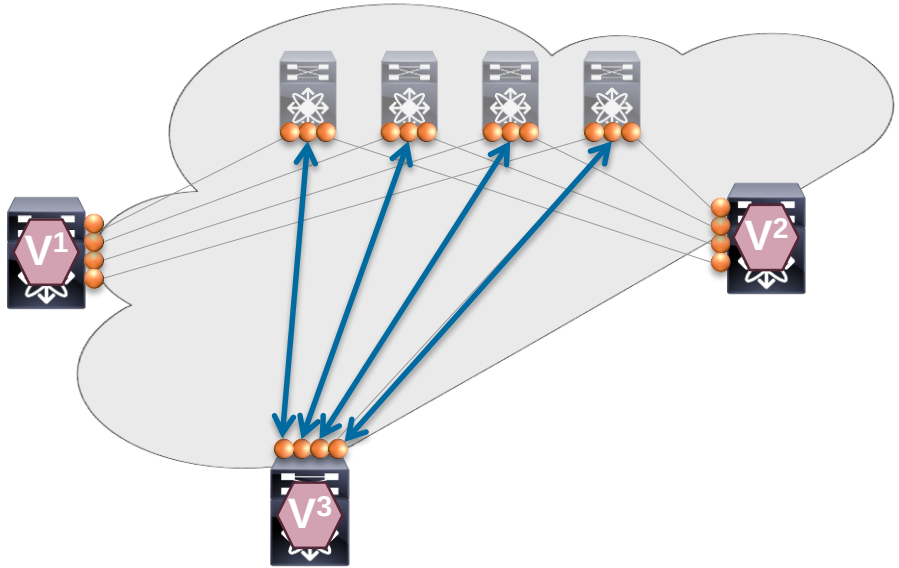
= 24 IP Addresses for P2P Links
= 5 IP Addresses for Loopback Interfaces

29 IP Addresses required == /27 Prefix

A More Realistic Scenario:
4 Spine * 40 Leaf = 160 Point-2-Point (P2P) Link
160 Links * 4 (/30) + 40 VTEP + 2 RR

= 640 IP Addresses for P2P Links
= 42 IP Addresses for Loopback Interface

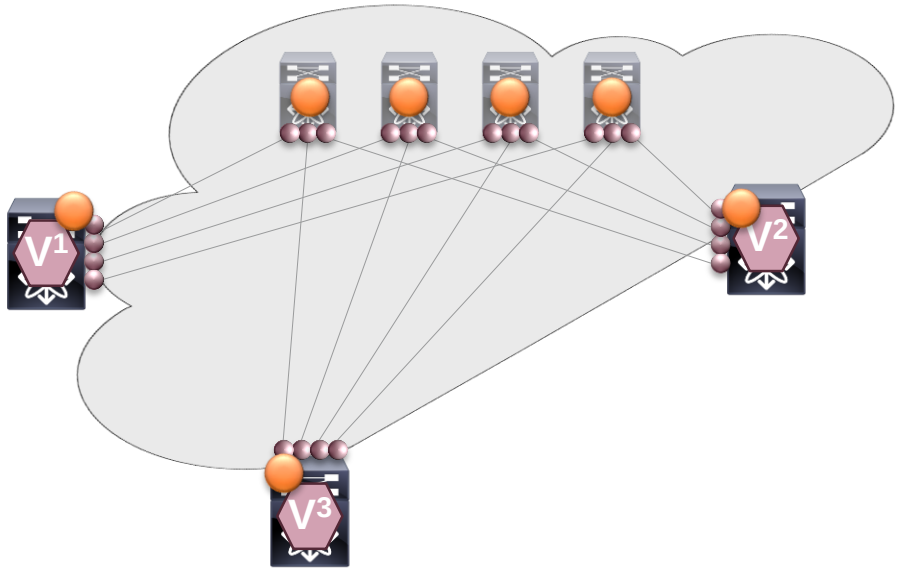
682 IP Addresses required == /22 Prefix



IP Unnumbered – Simplifying the Principals

Underlay

- **IP Unnumbered** – Single IP Address for multiple Interfaces
- Well-Known from Serial Interfaces (back in time)
- Used for Layer-3 Interfaces between Spine and Leaf (no switchport)
- For each Switch in the fabric, **single IP address** is sufficient
  - Loopback for VTEP
  - IP Unnumbered from Loopback for routed Interfaces



Check Platform & Release Support for Ethernet IP Unnumbered

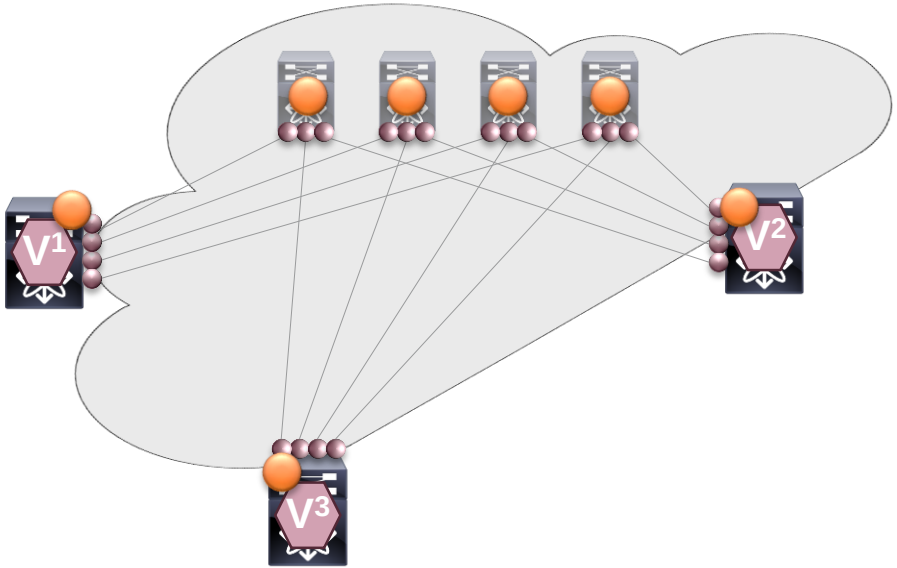
FYI

IP Unnumbered – Interface Configuration

Underlay

Interface Configuration Example for (V^1)

```
# Loopback Interface Configuration (VTEP & IP
Unnumbered)
interface loopback 0
 ip address 10.10.10.V1/32
 mtu 9192

# Point-2-Point (P2P) Interface Configration
interface Ethernet 2/1
 no switchport
 ip unnnumbered loopback 0
 mtu 9192

interface Ethernet 2/2
 no switchport
 ip unnnumbered loopback 0
 mtu 9192
.
.
```



Check Platform & Rlease Support for Ethernet IP Unnumbered

IP Unnumbered– Simplifying the Math

Underlay

Example from showed Topology:
4 Spine + 3 Leaf = 7 Individual Devices

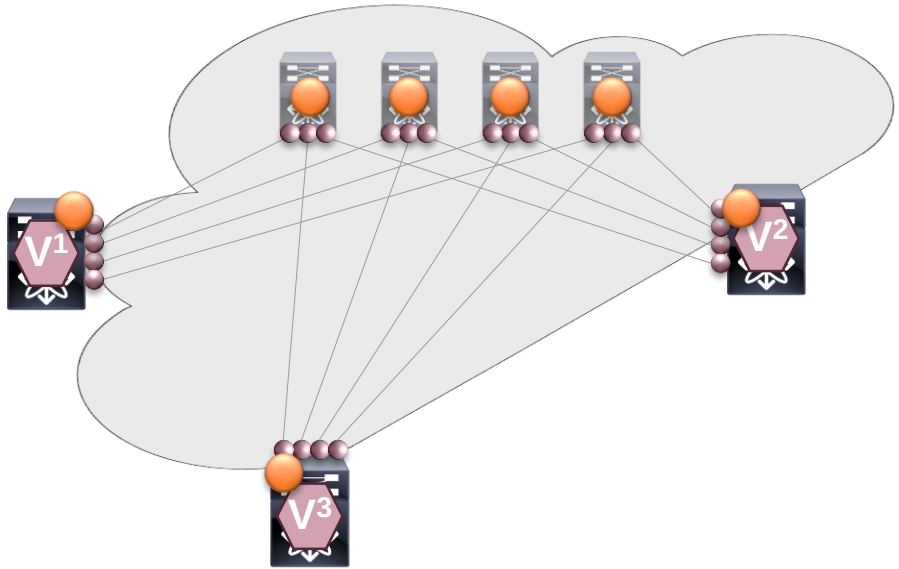
= 7 IP Addresses for Loopback Interface
(Used for VTEP & Routed Interfaces; IP Unnumbered)

7 IP Addresses required == /29 Prefix

A More Realistic Scenario:
4 Spine + 40 Leaf = 44 Individual Devices

= 44 IP Addresses for Loopback Interface
(Used for VTEP & Routed Interfaces; IP Unnumbered)

44 IP Addresses required == /26 Prefix

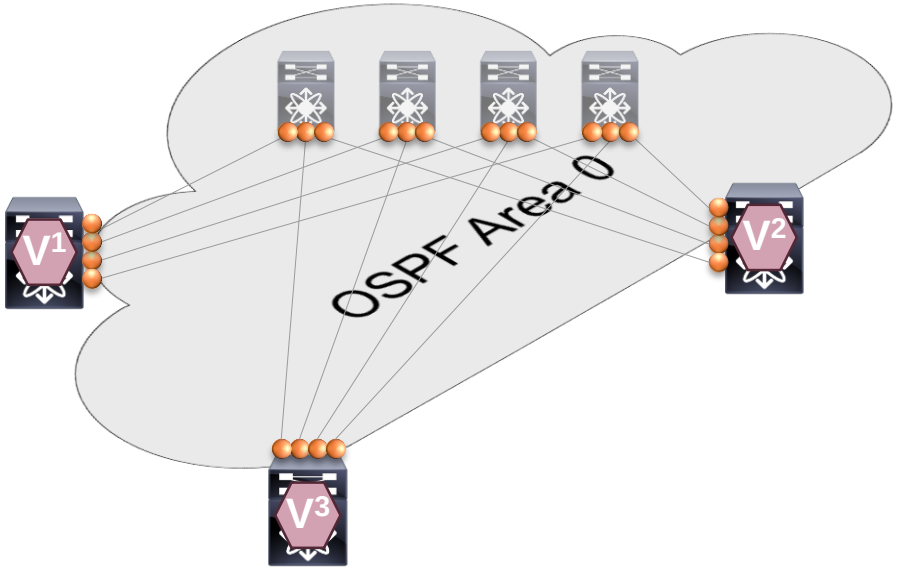


Check Platform & Release Support for Ethernet IP Unnumbered

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



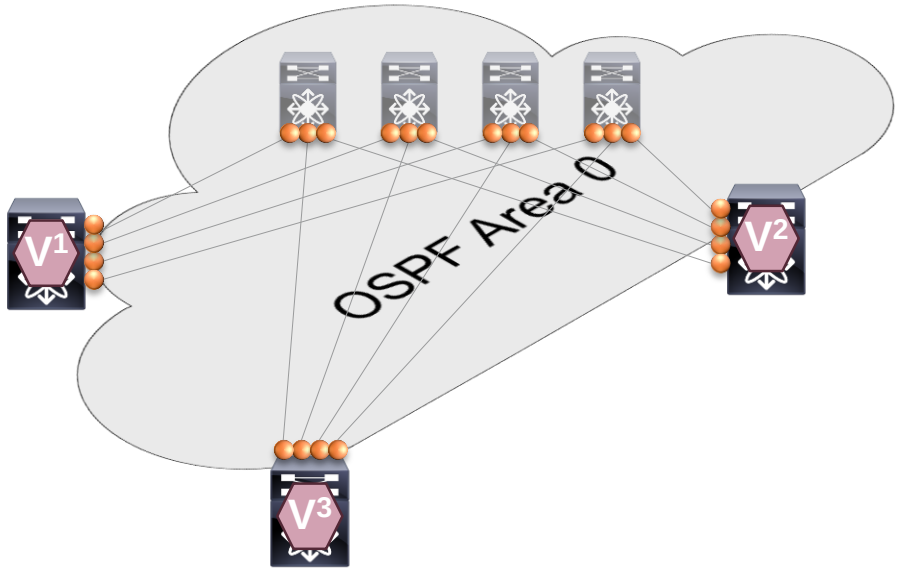
FYI

Building your IP Network – Routing Protocols; OSPF

Underlay

Configuration Example for (V^1)

```
# Loopback Interface Configuration (VTEP)
interface loopback 0
 ip address 10.10.10.V1/32
 mtu 9192
 ip router ospf 1 area 0.0.0.0
 ip ospf network point-to-point

# Point-2-Point (P2P) Interface Configuration
interface Ethernet 2/1
 no switchport
 ip address 192.168.1.1/31
 mtu 9192
 ip router ospf 1 area 0.0.0.0
 ip ospf network point-to-point
*
```



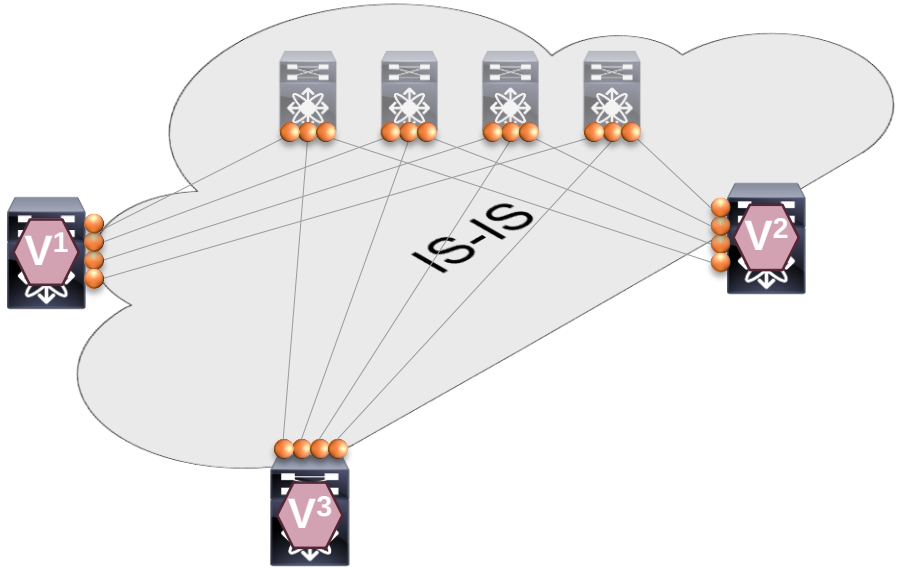
*Additional OSPF Configuration like Routing Process, Authentication or BFD are not shown

Cisco live!

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



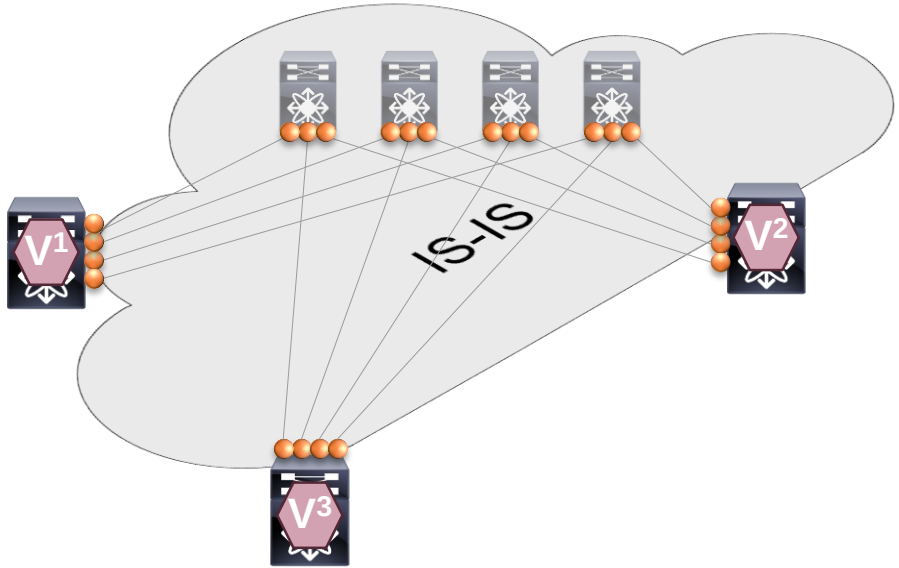
FYI

Building your IP Network – Routing Protocols; IS-IS

Underlay

Configuration Example for (V^1)

```
# Loopback Interface Configuration (VTEP)
interface loopback 0
 ip address 10.10.10.V1/32
 mtu 9192
 ip router isis 1
 medium p2p

# Point-2-Point (P2P) Interface Configuration
interface Ethernet 2/1
 no switchport
 ip address 192.168.1.1/31
 mtu 9192
 ip router isis 1
 medium p2p
*
```

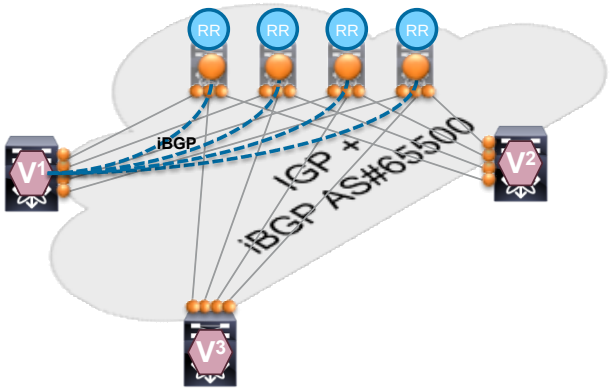


*Additional ISIS Configuration like Routing Process, Authentication or BFD are not shown

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



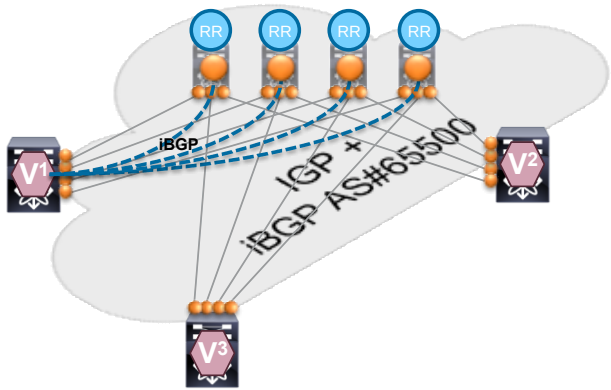
FYI

Building your IP Network – Routing Protocols; iBGP

Underlay

Configuration Example

```
# Spine (S1)
router bgp 65500
 neighbor 10.10.10.V1 remote-as 65500
  description S1-Loopback-to-V1-Loopback
  address-family ipv4 unicast
   route-reflector-client
 neighbor 10.10.10.V2 remote-as 65500
  description S1-Loopback-to-V2-Loopback
  address-family ipv4 unicast
   route-reflector-client

# Leaf (V1)
router bgp 65500
 neighbor 10.10.10.S1 remote-as 65500
  description V1-Loopback-to-S1-Loopback
 neighbor 10.10.10.S2 remote-as 65500
  description V1-Loopback-to-S2-Loopback
*
```



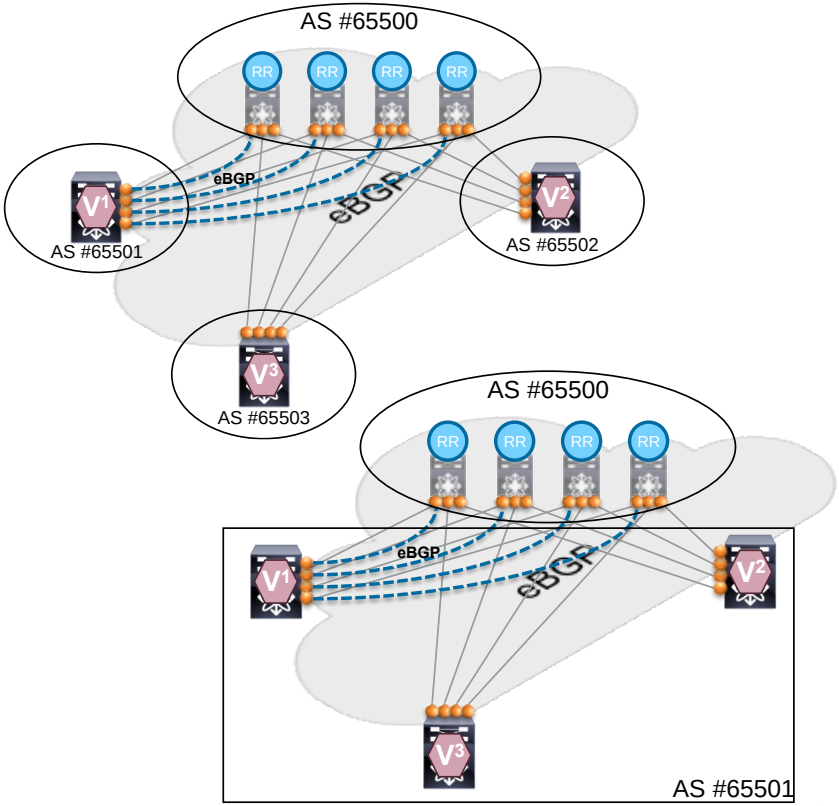
*Simplified BGP configuration; IGP not shown

Cisco live!

Building your IP Network – Routing Protocols; eBGP

Underlay

- eBGP
  - eBGP Peer is IP interface
    - Loopback would require additional IGP and eBGP multi-hop
  - Multiple Autonomous-Systems (AS)
    - Minimum amount of AS is two
  - Many BGP Neighbors
    - For each neighboring p2p interface
  - AS Path
    - Src and Dst AS might be same
  - No Route-Reflector
    - But next-hop needs to be unchanged



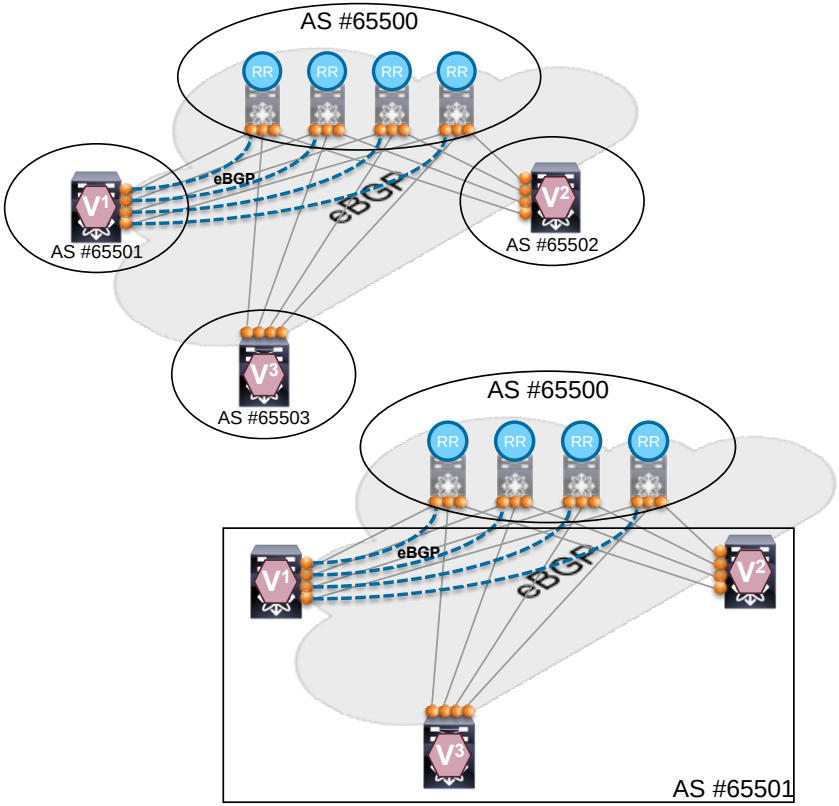
FYI

Building your IP Network – Routing Protocols; eBGP

Underlay

Configuration Example

```
# Spine (S1)
router bgp 65500
 neighbor 192.168.1.1 remote-as 65501
  description Spine1-Interface-to-V1
 neighbor 192.168.1.5 remote-as 65501
  description Spine1-Interface-to-V2

# Leaf (V1)
router bgp 65501
 neighbor 192.168.1.2 remote-as 65500
  description V1-Interface-to-S1
  address-family ipv4 unicast
   allowas-in
 neighbor 192.168.1.6 remote-as 65500
  description V1-Interface-to-S2
  address-family ipv4 unicast
   allowas-in
*
```



*Simplified BGP configuration

Cisco live!

FYI

Multicast Enabled Underlay

Underlay

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

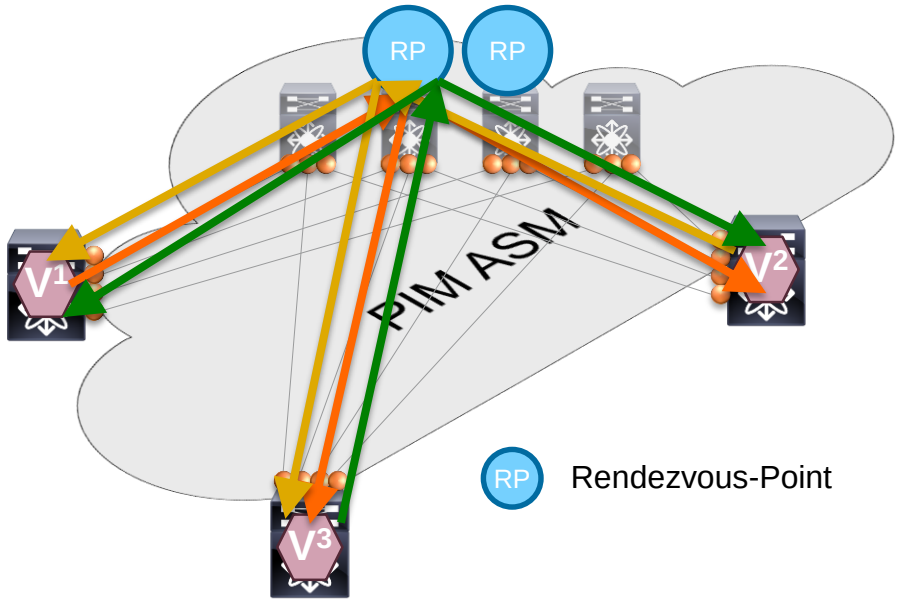
	Nexus 1000v	Nexus 3000	Nexus 5600	Nexus 7000/F3	Nexus 9000	ASR 1000 CSR 1000	ASR 9000
Multicast Mode	IGMP v2/v3	PIM ASM	PIM BiDir	PIM ASM / PIM BiDir	PIM ASM	PIM BiDir	PIM ASM / PIM BiDir

- Spine and Aggregation Switches make good Rendezvous-Point (RP) Locations in Topologies
- Reserve a range of Multicast Groups (Destination Groups/DGroups) to service the Overlay and optimize for diverse VNIs
- In Spine/Leaf topologies with lean Spine
  - Use multiple Rendezvous-Point across the multiple Spines
  - Map different VNIs to different Rendezvous-Point for simple load balancing measure
  - Use Redundant Rendezvous-Pint
- Design a Multicast Underlay for a Network Overlay, Host VTEPs will leverage this Network

Multicast Enabled Underlay – PIM ASM

Underlay

- PIM Sparse-Mode (ASM)
- Redundant Rendezvous-Point using PIM Anycast-RP or MSDP
- Source-Tree or Unidirectional Shared-Tree (Source-Tree shown)
  - Shared-Tree will always use RP for forwarding
- 1 Source-Tree per Multicast-Group per VTEP (each VTEP is Source & Receiver)
- Example from depicted topology
  - 3 VTEPs sharing same VNI and Multicast-Group mapping (single Multicast-Group)
  - 3x Source-Tree (1 per VTEP per Multicast-Group)



FYI

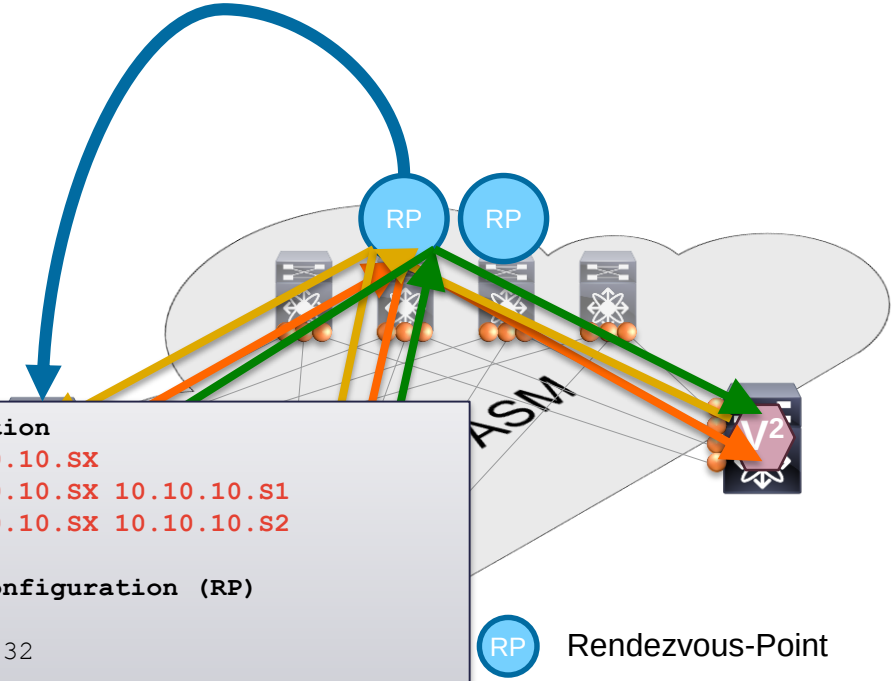
Multicast Enabled Underlay – PIM ASM

Underlay

Configuration Example for (V^1)

```
# Using Anycast Rendezvous-Point
ip pim rp-address 10.10.10.SX

# Loopback Interface Configuration (VTEP)
interface loopback 0
 ip address 10.10.10.V1/32
 mtu 9192
 ip pim sparse-mode

# Point-2-Point (P2P) Interface
interface Ethernet 2/1
 no switchport
 ip address 192.168.1.1/31
 mtu 9192
 ip pim sparse-mode
```

```
# Anycast-RP Configuration
ip pim rp-address 10.10.10.SX
ip pim anycast-rp 10.10.10.SX 10.10.10.S1
ip pim anycast-rp 10.10.10.SX 10.10.10.S2

# Loopback Interface Configuration (RP)
interface loopback 0
ip address 10.10.10.S1/32
ip pim sparse-mode

# Loopback Interface Configuration (Anycast RP)
interface loopback 1
ip address 10.10.10.SX/32
ip pim sparse-mode
```

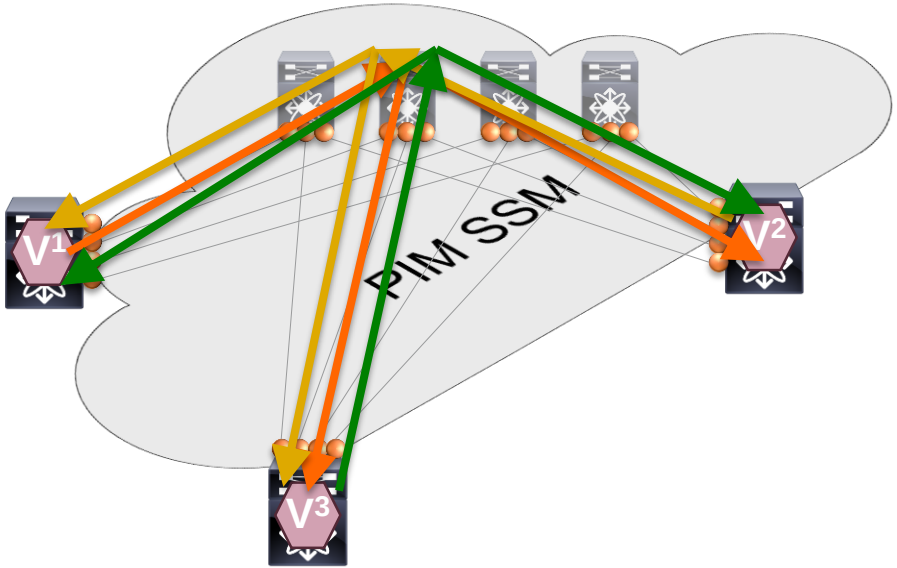


Cisco live!

Multicast Enabled Underlay – PIM SSM

Underlay

- PIM Source Specific Multicast (SSM)
- No Rendezvous-Point required
- Source-Tree or Unidirectional Shared-Tree
- 1 Source-Tree per Multicast-Group per VTEP (each VTEP is Source & Receiver)
- Example from showed Topology
  - 3 VTEPs sharing same VNI and Multicast-Group mapping (single Multicast-Group)
  - 3x Source-Tree (1 per VTEP per Multicast-Group)



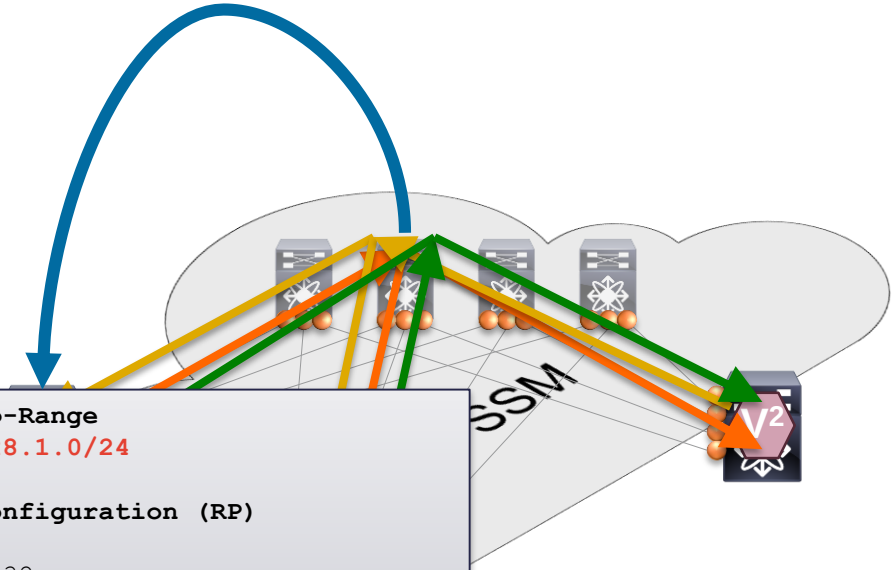
FYI

Multicast Enabled Underlay – PIM SSM

Underlay

Configuration Example for (V^1)

```
# Specificing PIM Group-Range
ip pim ssm range 239.128.1.0/24

# Loopback Interface Configuration (VTEP)
interface loopback 0
 ip address 10.10.10.V1/32
 mtu 9192
 ip pim sparse-mode

# Point-2-Point (P2P) Interface
interface Ethernet 2/1
 no switchport
 ip address 192.168.1.1/31
 mtu 9192
 ip pim sparse-mode
```



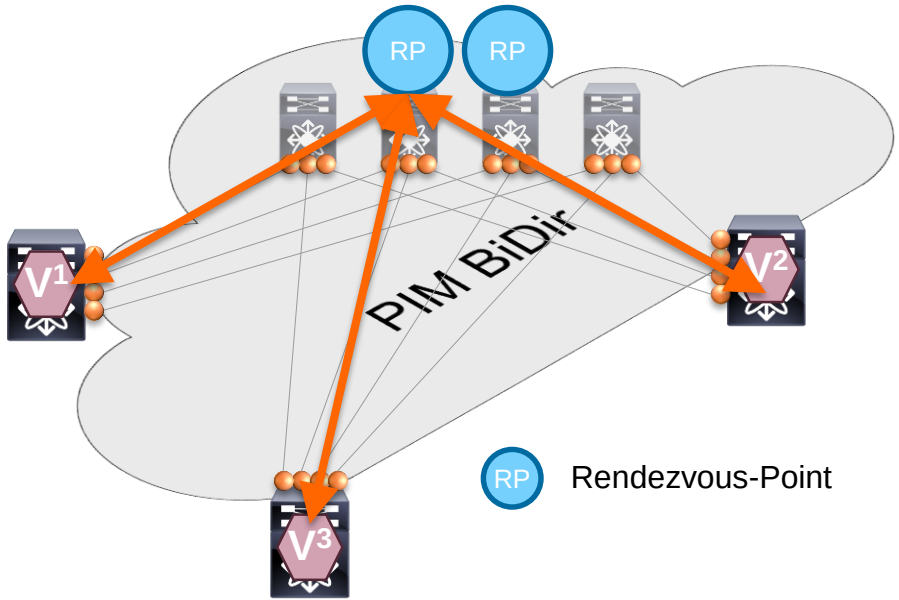

```
# Specificing PIM Group-Range
ip pim ssm range 239.128.1.0/24

# Loopback Interface Configuration (RP)
interface loopback 0
ip address 10.10.10.S1/32
ip pim sparse-mode

# Point-2-Point (P2P) Interface Configuration
interface Ethernet 2/1
 no switchport
 ip address 192.168.1.1/31
 mtu 9192
 ip pim sparse-mode
```


Cisco live!

Multicast Enabled Underlay – PIM BiDir

Underlay

- Bidirectional PIM (BiDir)
- Redundant Rendezvous-Point using Phantom-RP
- Building Bi-Directional Shared-Tree
  - Uses shortest path between Source and Receiver with RP as routing-vector
- 1 Shared-Tree per Multicast-Group
- Example from depicted topology
  - 3 VTEPs sharing same VNI and Multicast-Group mapping (single Multicast-Group)
  - 1x Shared-Tree (1 per Multicast-Group)



FYI

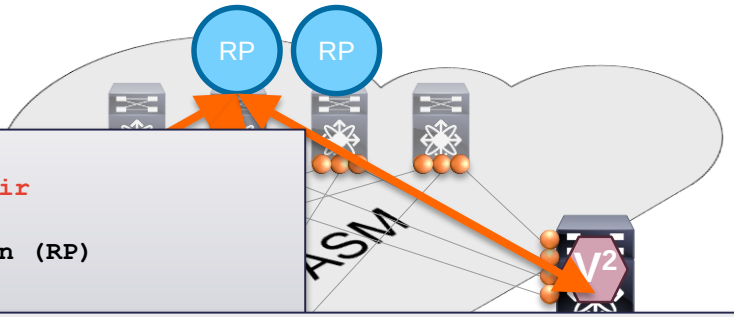
Multicast Enabled Underlay – PIM BiDir

Underlay

Configuration Example for (V1)

```
# Using Phantom Rendezvous-Point
ip pim rp-address 10.10.10.SX bidir

# Loopback Interface Configuration (VTEP)
interface loopback 0
 ip address 10.10.10.V1/32
 mtu 9192
 ip pim sparse-mode

# Point-2-Point (P2P) Interface
interface Ethernet 2/1
 no switchport
 ip address 192.168.1.1/31
 mtu 9192
 ip pim sparse-mode
```

```
# Using Phantom Rendezvous-Point
ip pim rp-address 10.10.10.SX bidir

# Loopback Interface Configuration (RP)
interface loopback 0
ip address 10.10.10.S1/32
ip pim sparse-mode

# Loopback Interface Config
interface loopback 1
ip address 10.10.10.SX/32
ip pim sparse-mode
```

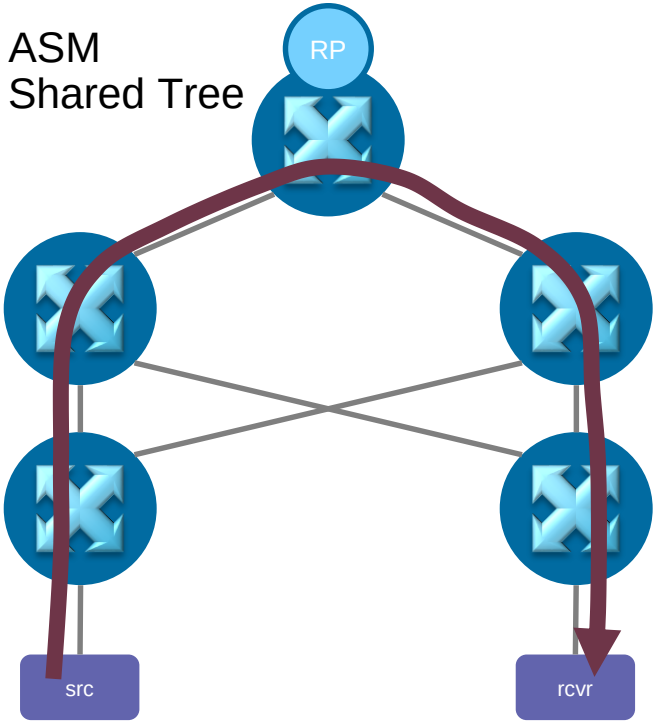



```
# Using Phantom Rendezvous-Point
ip pim rp-address 10.10.10.SX bidir

# Loopback Interface Configuration (RP)
interface loopback 0
ip address 10.10.10.S2/32
ip pim sparse-mode

# Loopback Interface Configuration (Anycast RP)
interface loopback 1
ip address 10.10.10.SX/31
ip pim sparse-mode
```

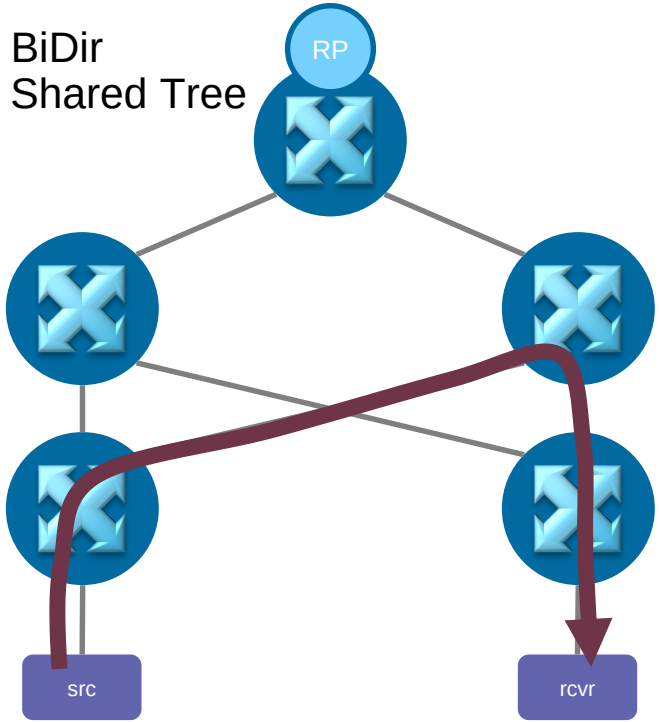
Multicast Enabled Underlay – PIM ASM vs. BiDir

Underlay - Flow Diagram Shared Tree

PIM ASM
with Shared Tree



PIM BiDir
with Shared Tree



FYI

To Remember

Multicast Enabled Underlay

- Multi-Destination Traffic (Broadcast, Unknown Unicast, etc.) needs to be replicated to ALL VTEPs serving a given VNI
  - Each VTEP is Multicast Source & Receiver
- For a given VNI, all VTEPs act as a Sender and a Receiver
- Head-End Replication will depend on hardware scale/capability
- Resilient, efficient, and scalable Multicast Forwarding is highly desirable
  - Choose the right Multicast Routing Protocol for your need (type/mode)
  - Use redundant Multicast Rendezvous Points (Spine/Aggregation generally preferred)
  - 99% percent of Overlay problems are in the Underlay (OTV experience)

Keep in Mind

Overlay Convergence = Underlay Convergence!

Cisco live!

Simplified Underlay

- No IP addressing on Point-2-Point connections*
- No ARP flooding on Underlay
- Optimized Multi-Destination Topology for Scale and Convergence

Optimized Overlay

Automated Configuration

*Related to VXLAN Layer-3 Underlay (IP Un-Numbered)

Cisco live!

Optimized Networks with VXLAN

Data Center Fabric Properties

- ✓Extended Namespace
- ✓Scalable Layer-2 Domains
- ❑Integrated Route and Bridge
- ❑Multi-Tenancy

Cisco live!

Agenda

- Data Center Fabric Properties
- Optimized Networks with VXLAN
  - Overview
  - Underlay
  - **Control & Data Plane**
  - Multi-Tenancy
- Optimized Networks with FabricPath
- Fabric Management & Automation

Cisco live!

VXLAN Flood & Learn and Cisco FabricPath

Compared

		VXLAN	FabricPath
Encapsulation		Packet Encapsulation (PE)	Frame Encapsulation (FE)
Transport Medium Requirement		Layer-3	Layer-1 (mandatory)
End-Host Reachability and Distribution		Flood & Learn	Flood & Learn (+Conversational Learning)
End-Host Detection		Flood & Learn	Flood & Learn
Multi-Destination Traffic (BUM*) forwarding		Multicast (PIM)	FabricPath IS-IS
Underlay Control-Plane		Any Unicast Routing Protocol (static, OSPF, IS-IS, eBGP)	FabricPath IS-IS
Unique Node Identifier		VTEP IP	SwitchID
Standard Reference		RFC 7348	TRILL based (Cisco Proprietary)

*BUM: Broadcast, Unknown Unicast, Multicast

Getting the Puzzle Together!

Driving Standards based Overlay-Evolution with

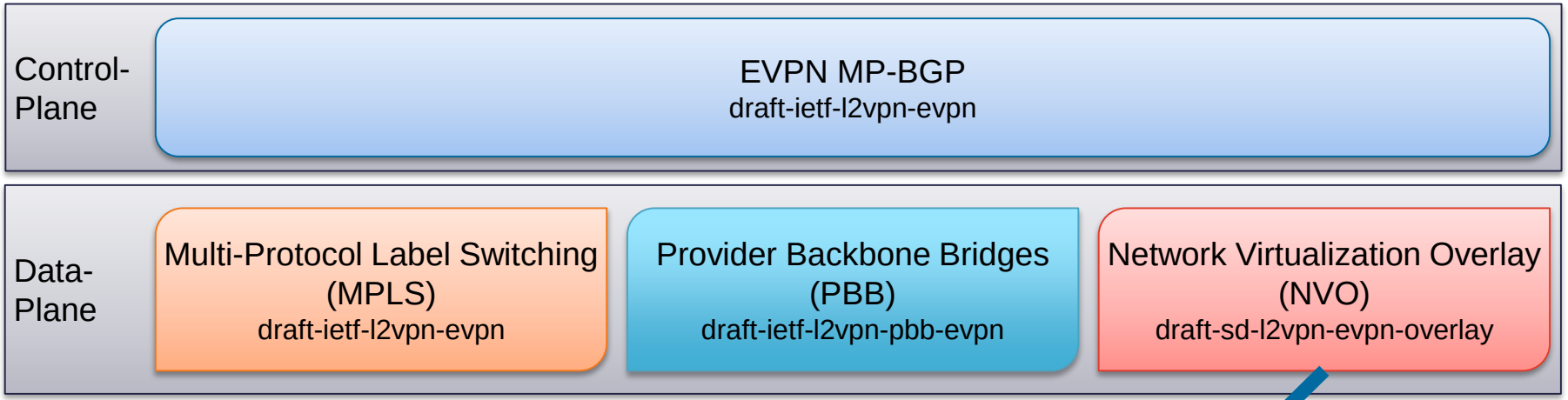
VXLAN/EVPN

Cisco live!

FYI

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

Cisco live!

FYI

EVPN – Ethernet VPN

Additional Information

- IETF Layer-2 Virtual Private Network (l2vpn) Working Group
  - http://datatracker.ietf.org/wg/l2vpn/
- RFC 7209: Requirements for Ethernet VPN (EVPN)
  - http://tools.ietf.org/html/rfc7209
- Based Specifications: draft-ietf-l2vpn-evpn
  - http://tools.ietf.org/html/draft-ietf-l2vpn-evpn

Cisco live!

VXLAN Evolution

Protocol Learning

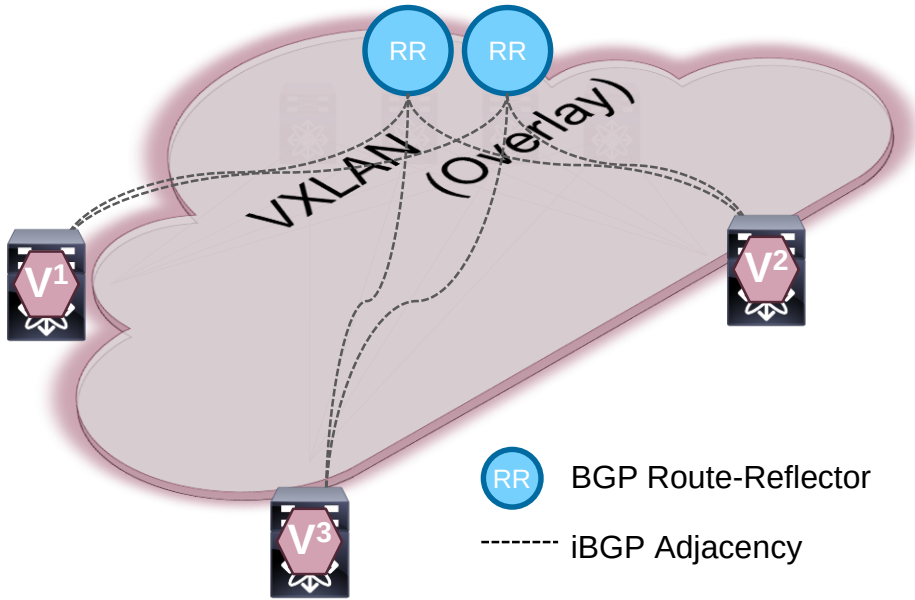
- Workload MAC and IP Addresses learnt by VXLAN Edge Devices (NVEs)
- Advertises Layer-2 and Layer-3 Address-to-VTEP Association (Overlay Control-Plane)
- Flood Prevention
- Optimized ARP forwarding

- 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); Integrated Route/Bridge (IRB)
- Reduce Flooding
- Reduce impact of ARP on the Network
- Standards Based (IETF draft)

Host and Subnet Route Distribution

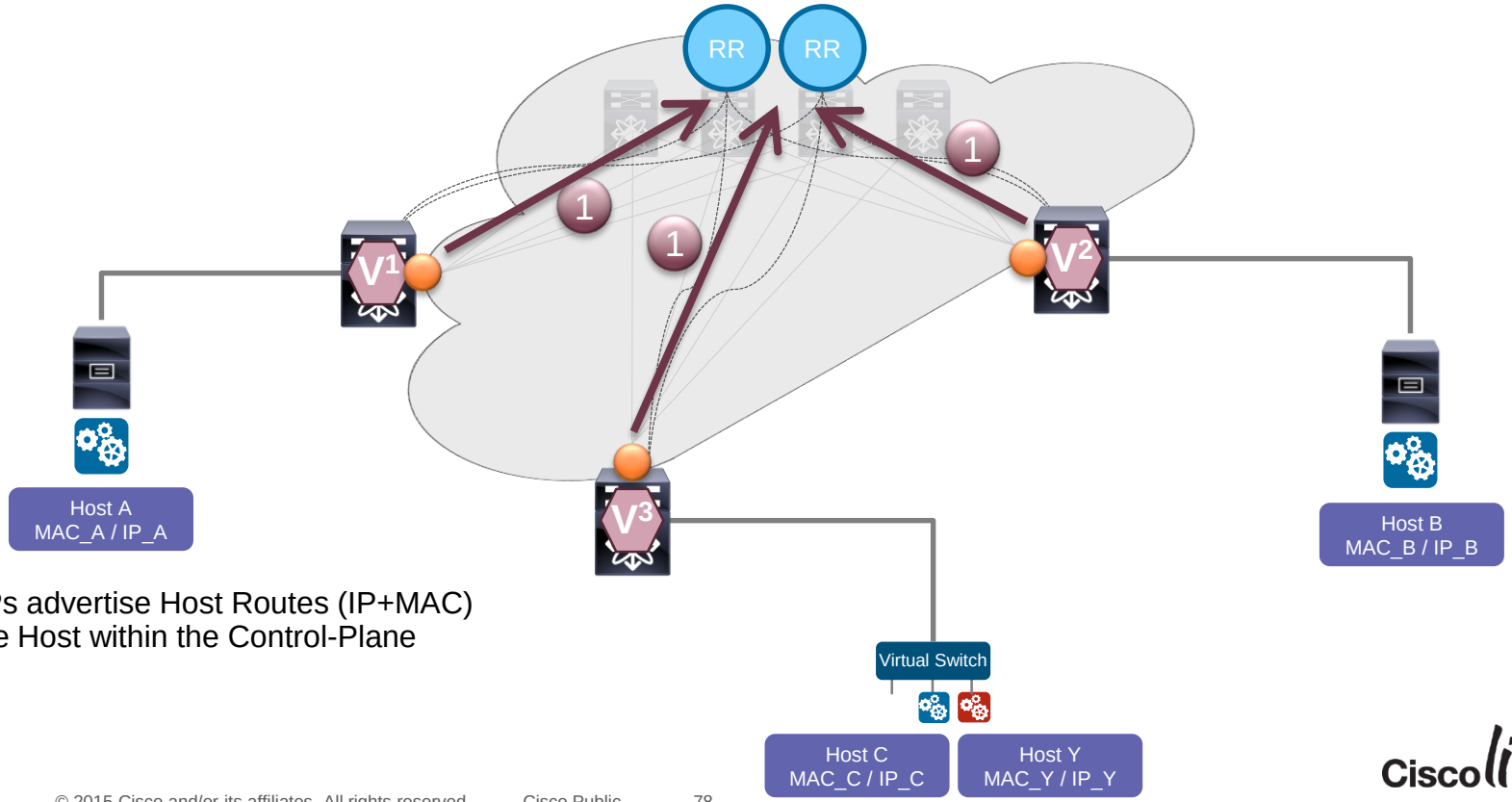
VXLAN/EVPN

- 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



Protocol Learning & Distribution (1)

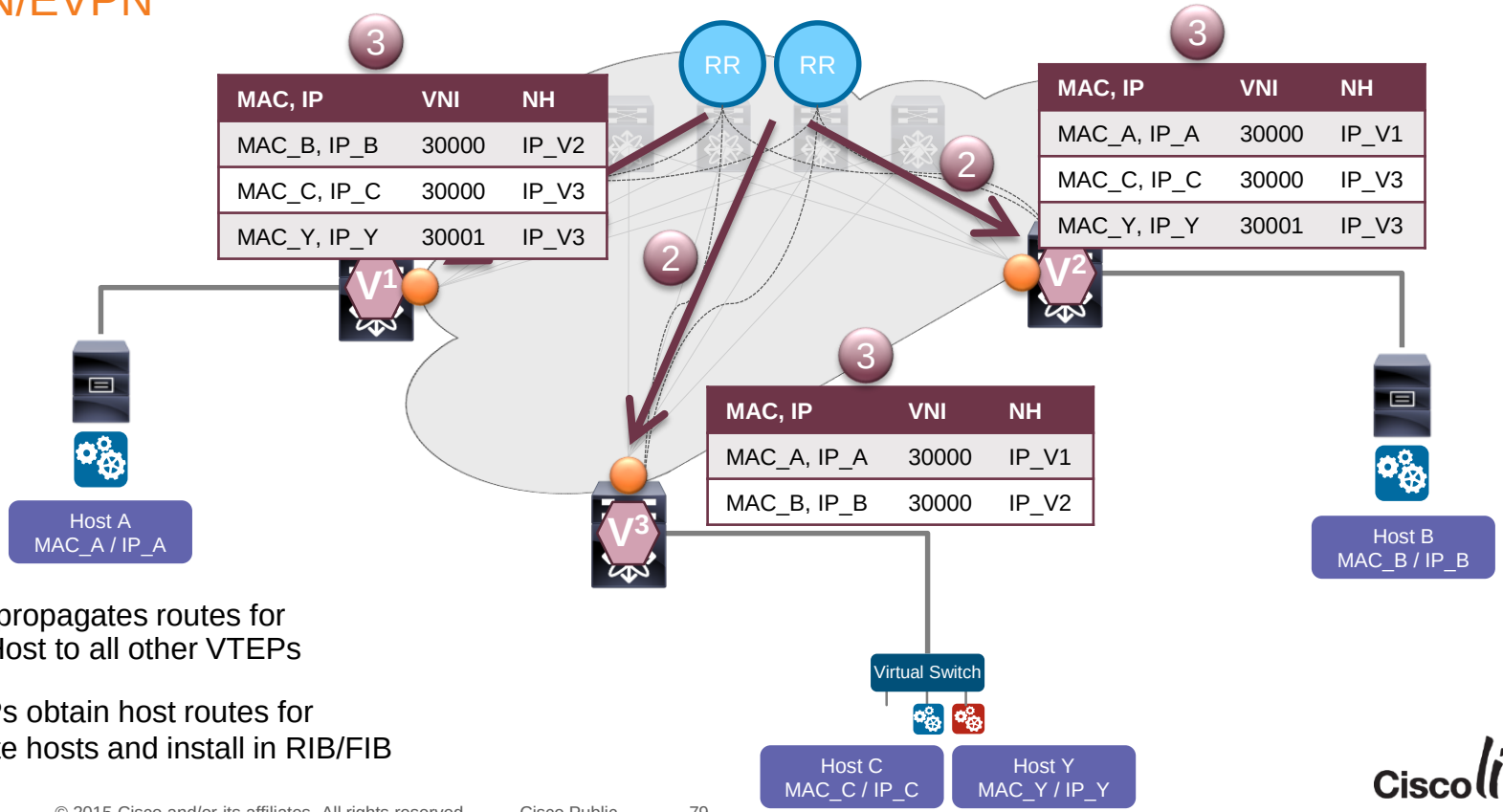
VXLAN/EVPN



1 VTEPs advertise Host Routes (IP+MAC) for the Host within the Control-Plane

Protocol Learning & Distribution (2)

VXLAN/EVPN



2 BGP propagates routes for The Host to all other VTEPs

3 VTEPs obtain host routes for remote hosts and install in RIB/FIB

Host Advertisement

VXLAN/EVPN

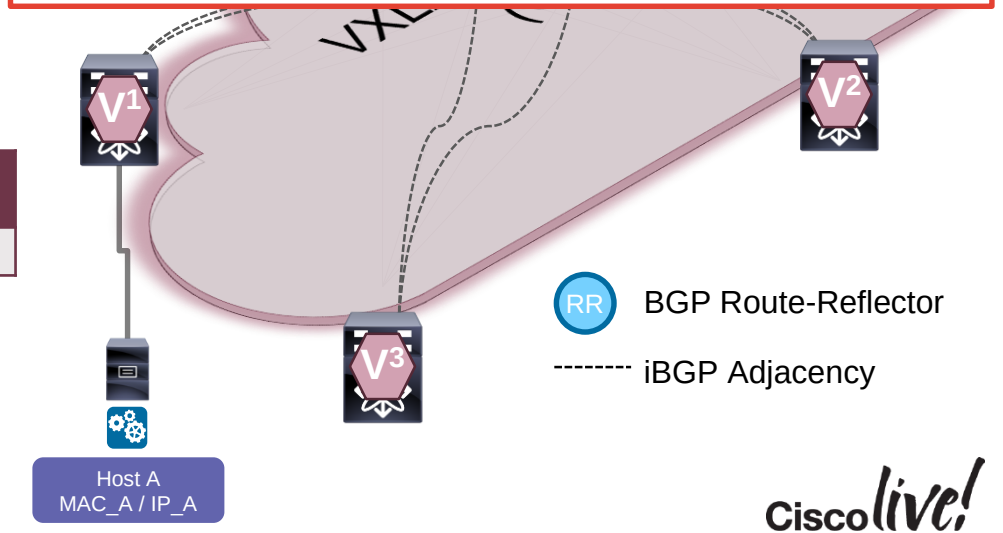
1. Host Attaches
2. VTEP V1 advertises Host A MAC (+IP) through BGP RR
3. Choice of Encapsulation is also advertised

```
V1# sh bgp l2vpn evpn IP_A
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 30000:V1
BGP routing table entry for [2]:[0]:[0]:[48]:[MAC_A]:[32]:[IP_A]/272, version 28838
Paths: (1 available, best #1)
Flags: (0x000202) on xmit-list, is not in l2rib/evpn

  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop
  AS-Path: NONE, path sourced internal to AS
    IP_V1 (metric 3) from RR (RR)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 30000 50000
      Extcommunity: RT:1000:30000 RT:1000:50000 ENCAP:3
      Originator: IP_V1 Cluster list: RR
      Remote Next-hop Attribute: IP_V1
        encapsulation VXLAN VNID 50000 MAC MAC_V1
```

48, MAC, 32, IP

ENCAP:3 = VXLAN

MAC, IP	VNI (L2)	VNI (L3)	NH	Encap	Seq
MAC_A, IP_A	30000	50000	IP_V1	3:VXLAN	0



```
V1# sh bgp l2vpn evpn IP_A
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 30000:V1
BGP routing table entry for [2]:[0]:[0]:[48]:[MAC_A]:[32]:[IP_A]/272, version 28838
Paths: (1 available, best #1)
Flags: (0x000202) on xmit-list, is not in l2rib/evpn

  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop
  AS-Path: NONE, path sourced internal to AS
    IP_V1 (metric 3) from RR (RR)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 30000 50000
      Extcommunity: RT:1000:30000 RT:1000:50000 ENCAP:3
      Originator: IP_V1 Cluster list: RR
      Remote Next-hop Attribute: IP_V1
        encapsulation VXLAN VNID 50000 MAC MAC_V1
```

48, MAC, 32, IP

ENCAP:3 = VXLAN

Cisco live!

Host Moves

VXLAN/EVPN

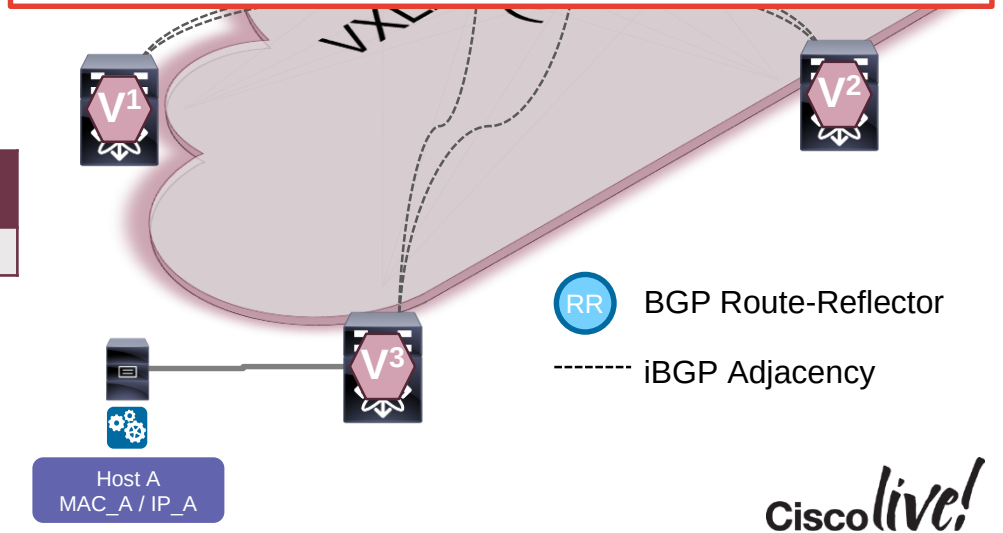
1. Host Moves to V3
2. V3 detects Host A and advertises it with Seq #1
3. V1 sees more recent route and withdraws its advertisement

MAC, IP	VNI (L2)	VNI (L3)	NH	Encap	Seq
MAC_A, IP_A	30000	50000	**IP_V3**	3:VXLAN	**1**

```
V1# sh bgp l2vpn evpn IP_A
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 30000:V3
BGP routing table entry for [2]:[0]:[0]:[48]:[MAC_A]:[32]:[IP_A]/272, version 28839
Paths: (1 available, best #1)
Flags: (0x000202) on xmit-list, is not in l2rib/evpn

  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop
  AS-Path: NONE, path sourced internal to AS
    IP_V3 (metric 3) from RR (RR)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 30000 50000
      Extcommunity: RT:1000:30000 RT:1000:50000 ENCAP:3
      Originator: IP_V3 Cluster list: RR
      Remote Next-hop Attribute: IP_V3
        encapsulation VXLAN VNID 50000 MAC MAC_V3
```

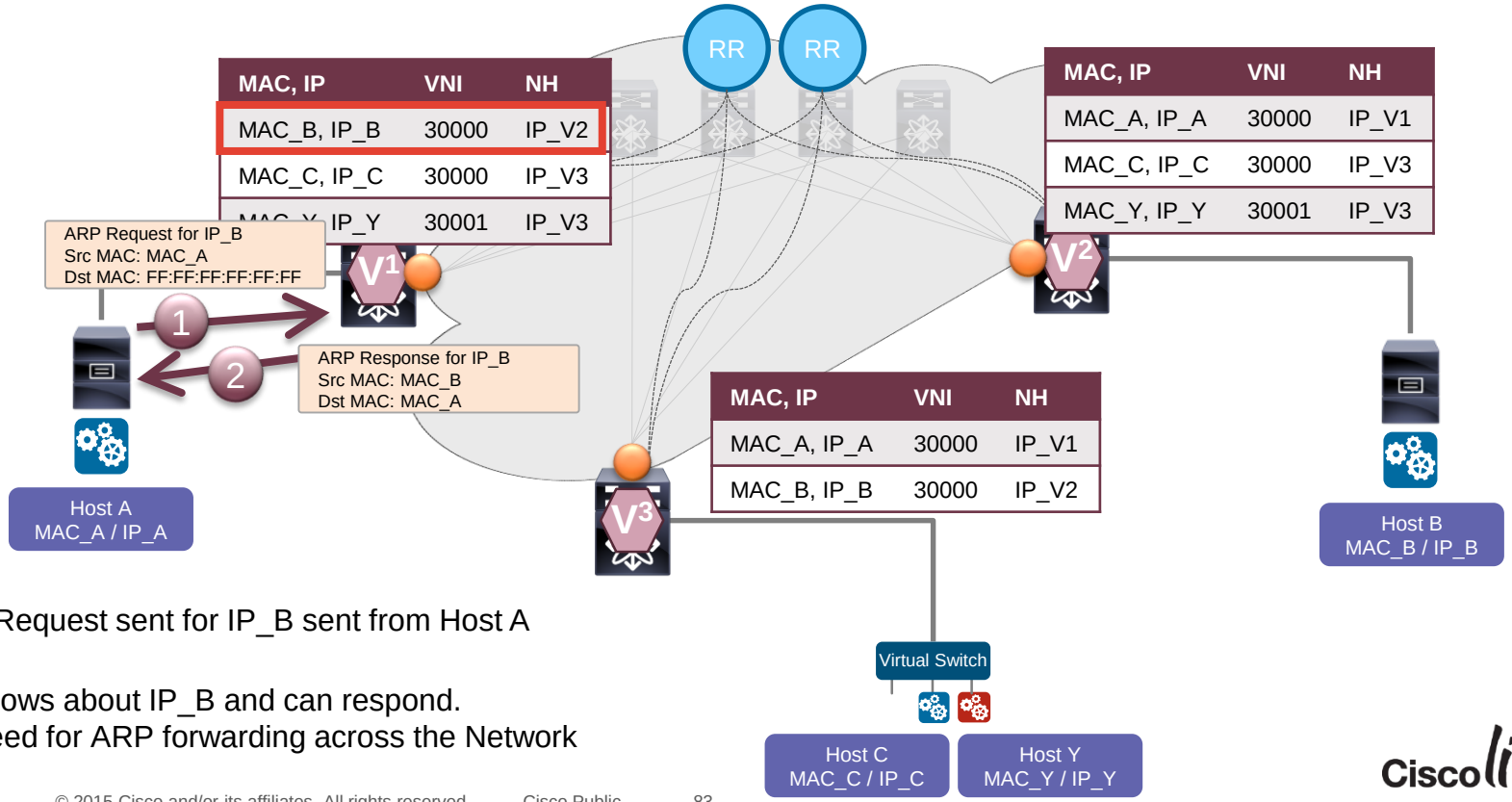
48, MAC, 32, IP

ENCAP:3 = VXLAN



ARP Suppression

VXLAN/EVPN



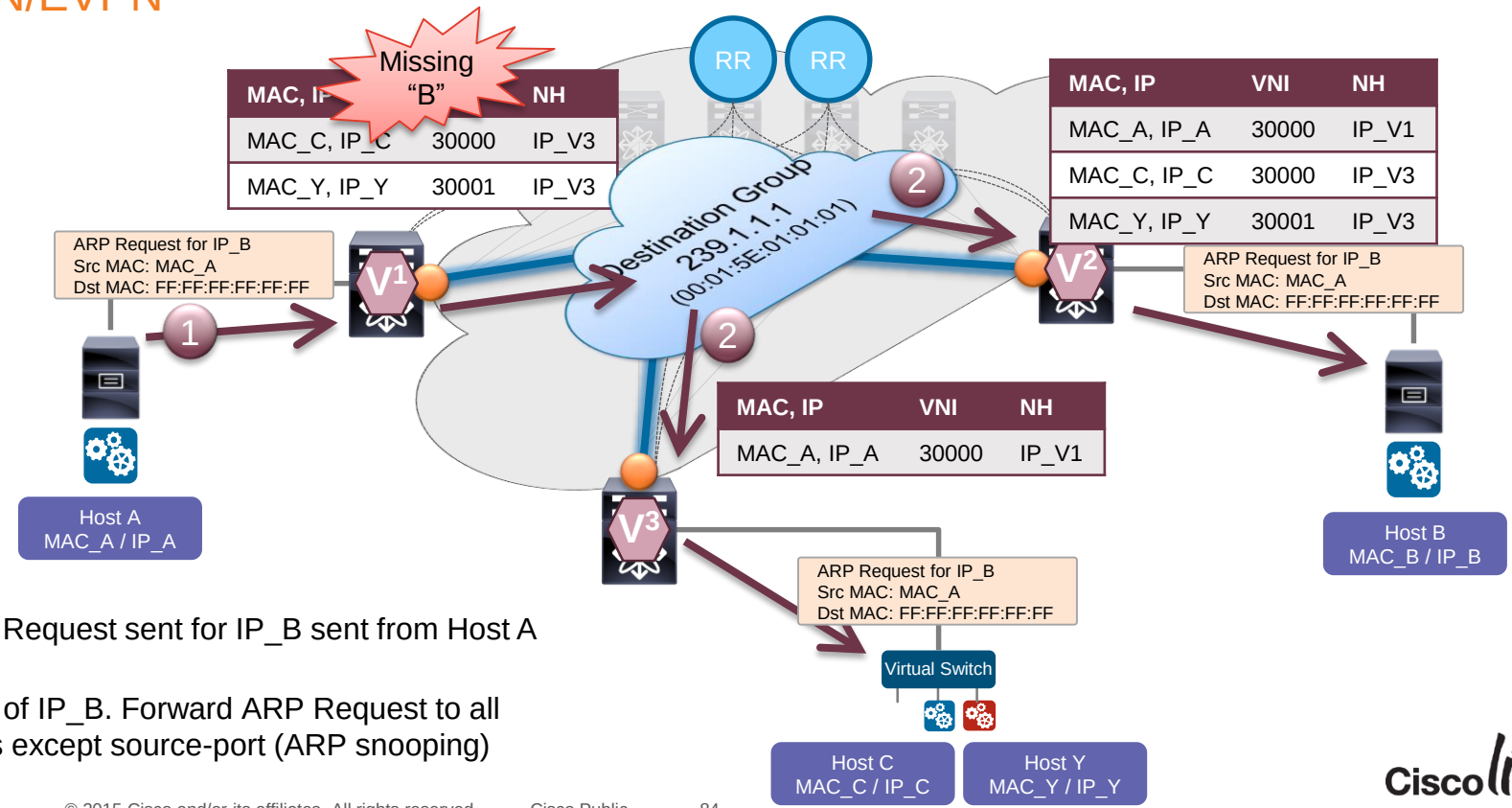
1. ARP Request sent for IP_B sent from Host A

2. V1 knows about IP_B and can respond.
   No need for ARP forwarding across the Network

Cisco live!

ARP Handling on Lookup “Miss” (1)

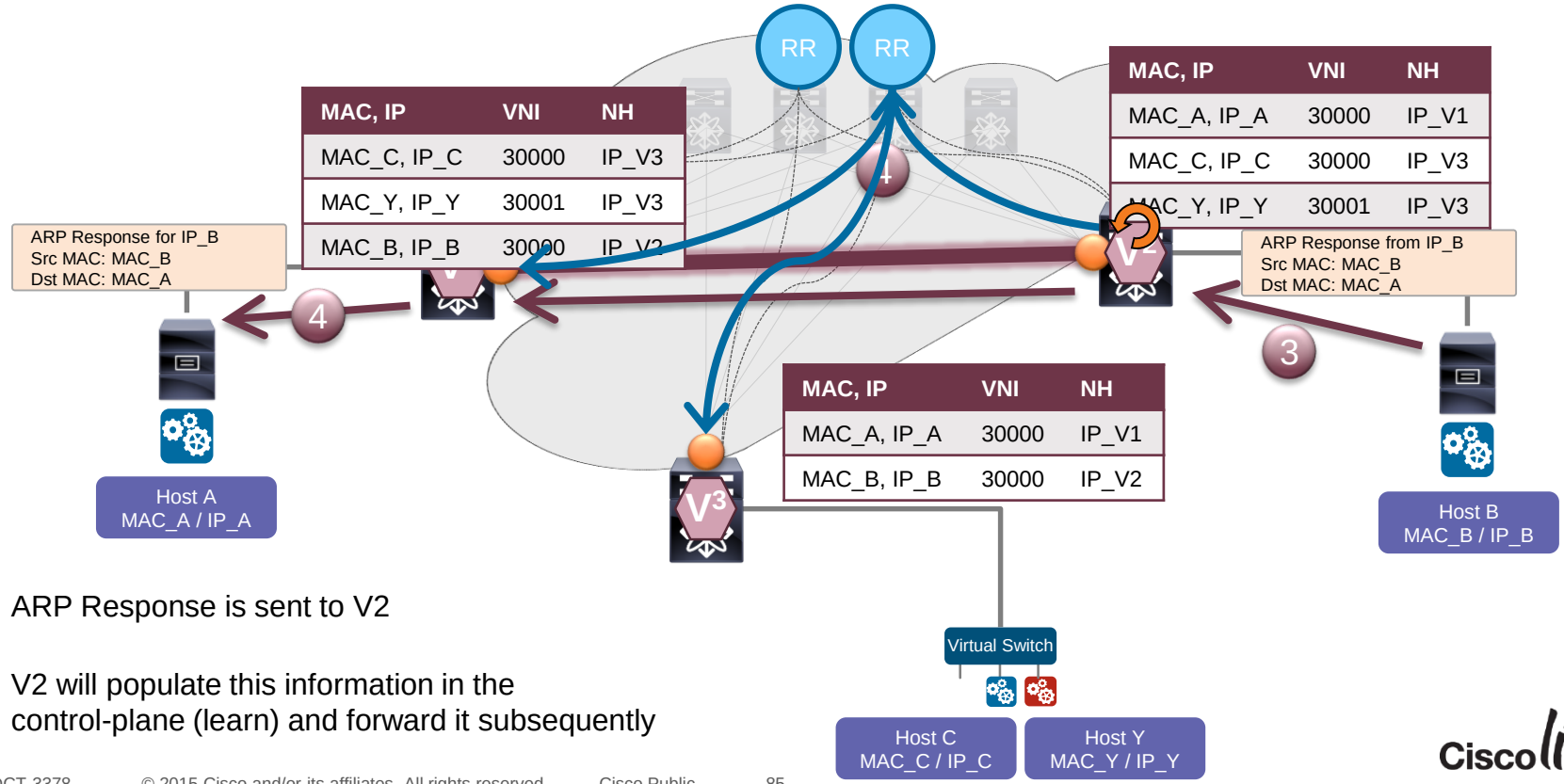
VXLAN/EVPN



1. ARP Request sent for IP_B sent from Host A
2. Miss of IP_B. Forward ARP Request to all Ports except source-port (ARP snooping)

ARP Handling on Lookup “Miss” (2)

VXLAN/EVPN

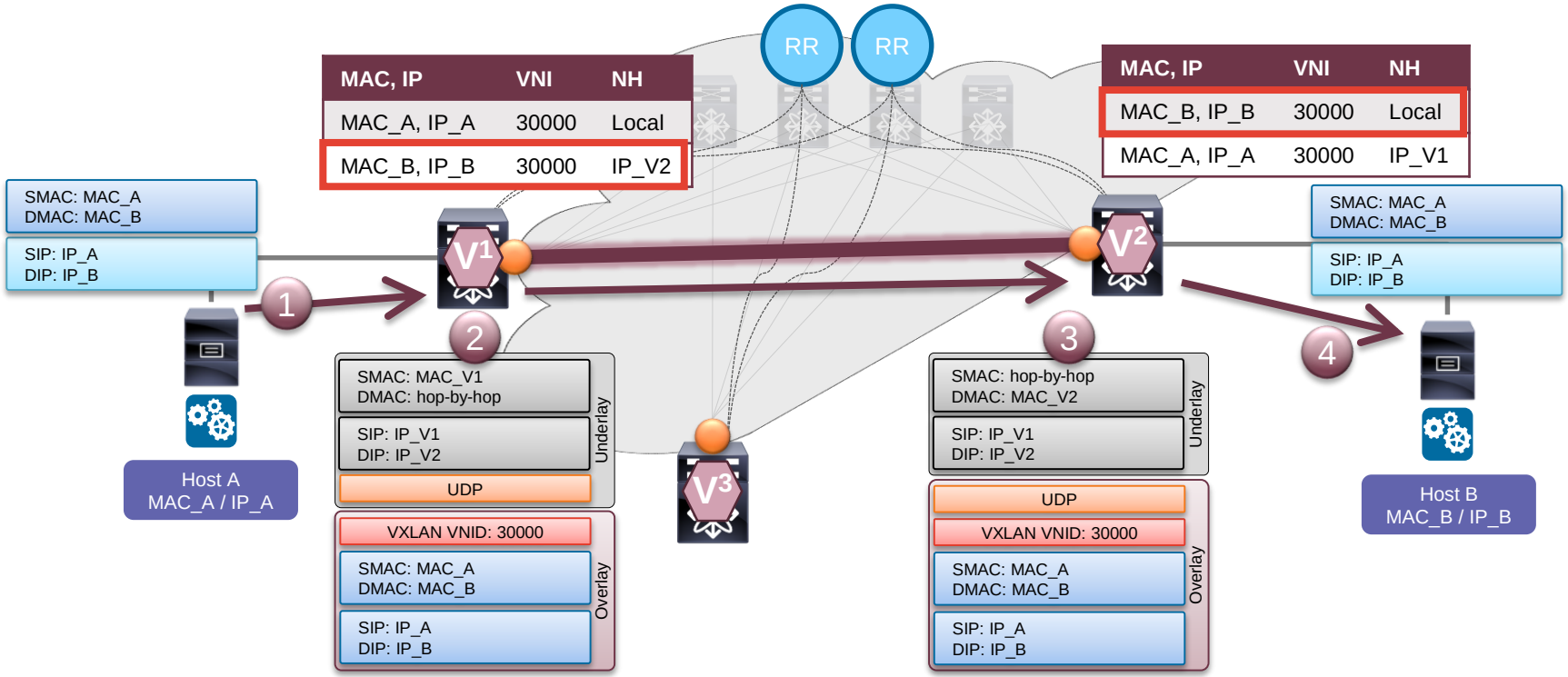


3 ARP Response is sent to V2

4 V2 will populate this information in the control-plane (learn) and forward it subsequently

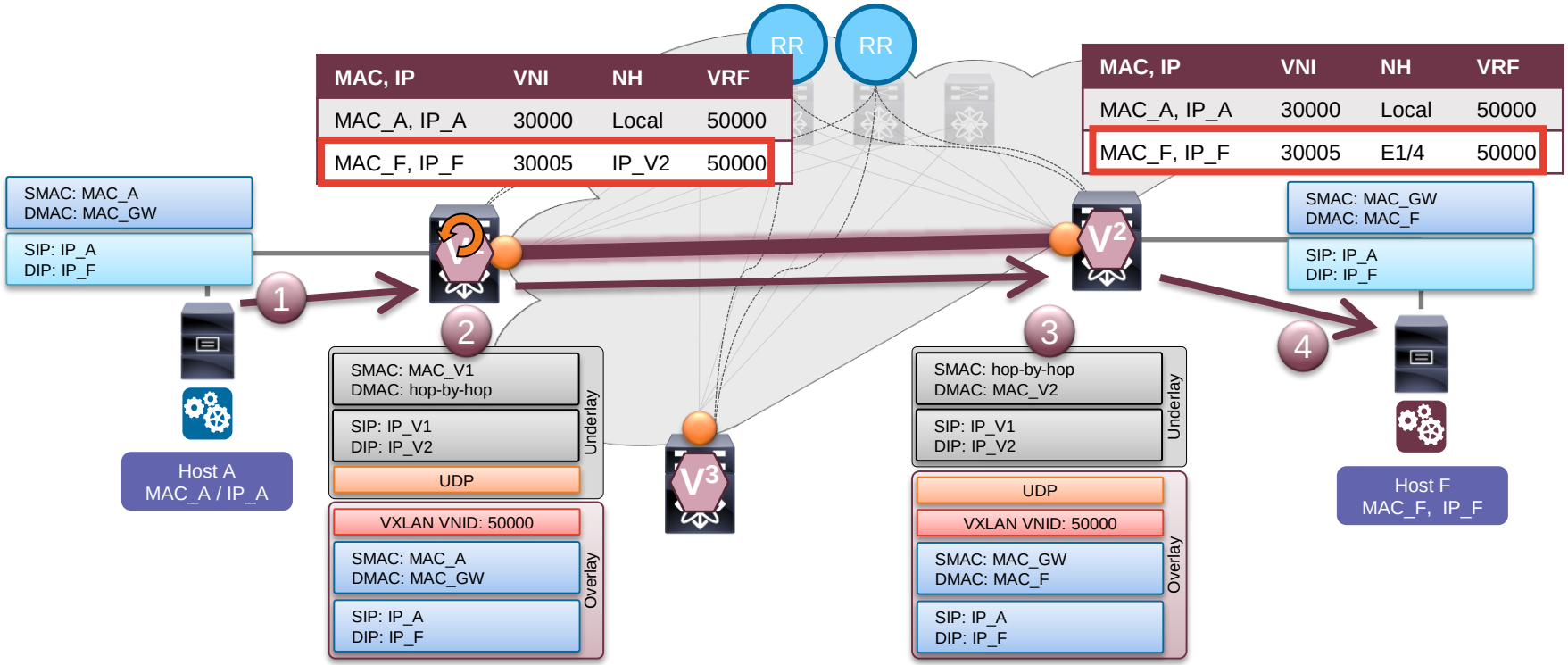
Packet Forwarding (Bridge)

VXLAN/EVPN



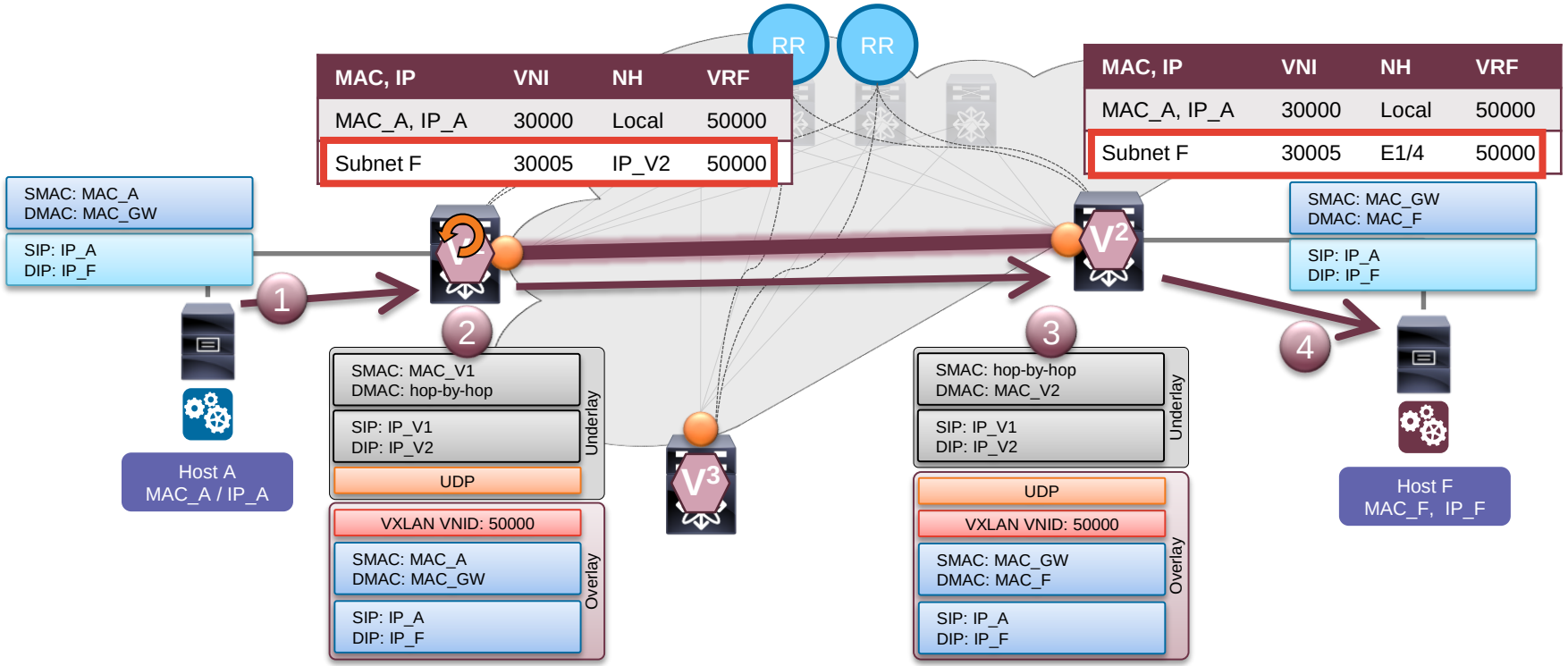
Packet Forwarding (Route)

VXLAN/EVPN



Packet Forwarding (Route) – Silent Host

VXLAN/EVPN



FYI

Protocol Learning

VXLAN/EVPN

Configuration Example for VLAN, VNI, VRF

```
# VLAN to VNI mapping (MT-Lite)
vlan 43
 vn-segment 30000
vlan 55
 vn-segment 30001

# Allocate VLAN for VRF VNI
vlan 500
  vn-segment 50000

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

Configuration Example for MP-BGP EVPN

```
# BGP Configuration for VRF (routing)
router bgp 65535
  address-family ipv4 unicast
  neighbor RR_IP remote-as 65535
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  vrf VRF-A
    address-family ipv4 unicast
      advertise l2vpn evpn

# EVPN Configuration for VNI (bridging)
evpn
  vni 30000 l2
    rd auto
    route-target both auto
 vni 30001 l2
    rd auto
    route-target both auto
```


Cisco live!

Optimized Networks with VXLAN

Data Center Fabric Properties

- ✓Extended Namespace
- ✓Scalable Layer-2 Domains
- ❑Integrated Route and Bridge
- ❑Multi-Tenancy

Cisco live!

VXLAN Evolution

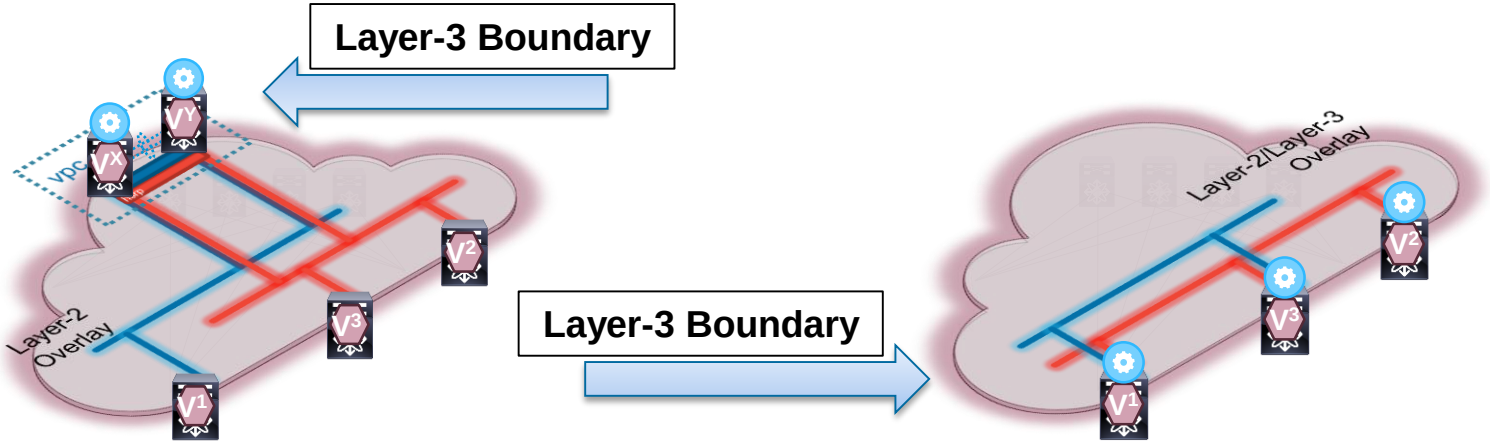
IP Services

- VXLAN Routing
- Distributed Anycast Gateway (requires Overlay Control-Plane)
- Multi-Tenancy

- Forward based on MAC or IP address learnt via Control-Plane (MP-BGP EVPN)
- Make routing decisions at VTEPs
- Scale and Multipathing (ECMP)
- Leverage Layer-3 Gateway capabilities along with Protocol Information
- LISP-ish / LISP-like approach for Host/IP Mobility
  - Location (VTEP), Identifier (MAC, IP of End-Host)

Gateway Functions in VXLAN

VXLAN Routing



Centralized Gateway

- Extra Bridging hop before and after Routing
- Centralized Gateway (Aggregation) for Routing
- Large amounts of state => convergence issues
- Scale problem for large Layer-2 domains
- **Works with VXLAN Flood & Learn or EVPN**

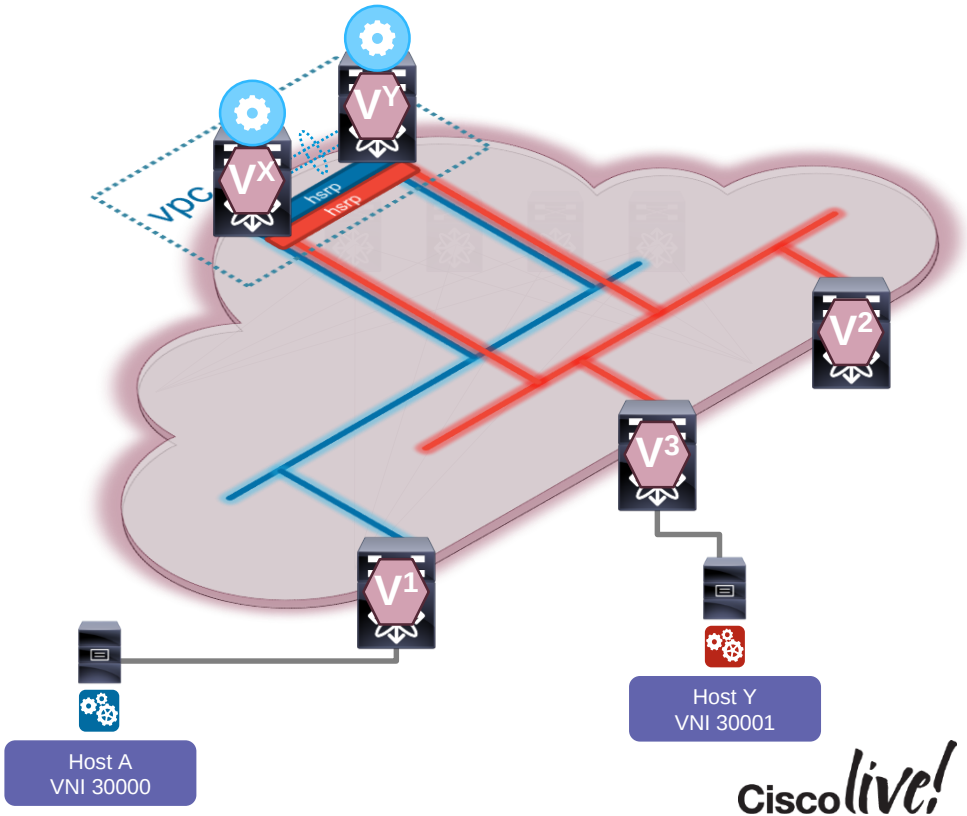
Distributed Gateway

- Route or Bridge at Leaf
- Distributed Gateway (Anycast) for Routing
- Disaggregate state by scale out
- Optimal Scalability
- **Requires VXLAN/EVPN!**

Centralized Gateway (FHRP)

VXLAN Routing

- Centralized Routing in a Layer-2 VXLAN Network
  - Routing between VNI (Different Subnet)
  - Bridging within VNI (Same Subnet)
- Inter-VXLAN Routing at Core/Aggregation Layer
- vPC provides MAC state synchronization and HSRP peering
  - Redundant VTEPs share Anycast VTEP IP address in the Underlay



FYI

Centralized Gateway (FHRP)

VXLAN Routing

Configuration Example for (V^x)

```
# VLAN to VNI mapping (MT-Lite)
vlan 43
 vn-segment 30000
vlan 55
 vn-segment 30001

# Gateway Interface for "BLUE" & "RED"(SVI)
interface vlan 43
 vrf member VRF-A
 ip address 11.11.11.2/24
  hsrp 43
  ip 11.11.11.1

interface vlan 55
 vrf member VRF-A
 ip address 98.98.98.2/24
  hsrp 43
  ip 98.98.98.1
```

Configuration Example for (V^y)

```
# VLAN to VNI mapping
vlan 43
 vn-segment 30000
vlan 55
 vn-segment 30001

# Gateway Interface for "BLUE" & "RED"(SVI)
interface vlan 43
 vrf member VRF-A
 ip address 11.11.11.3/24
  hsrp 43
  ip 11.11.11.1

interface vlan 55
 vrf member VRF-A
 ip address 98.98.98.3/24
  hsrp 43
  ip 98.98.98.1
```

VPC Configuration not shown

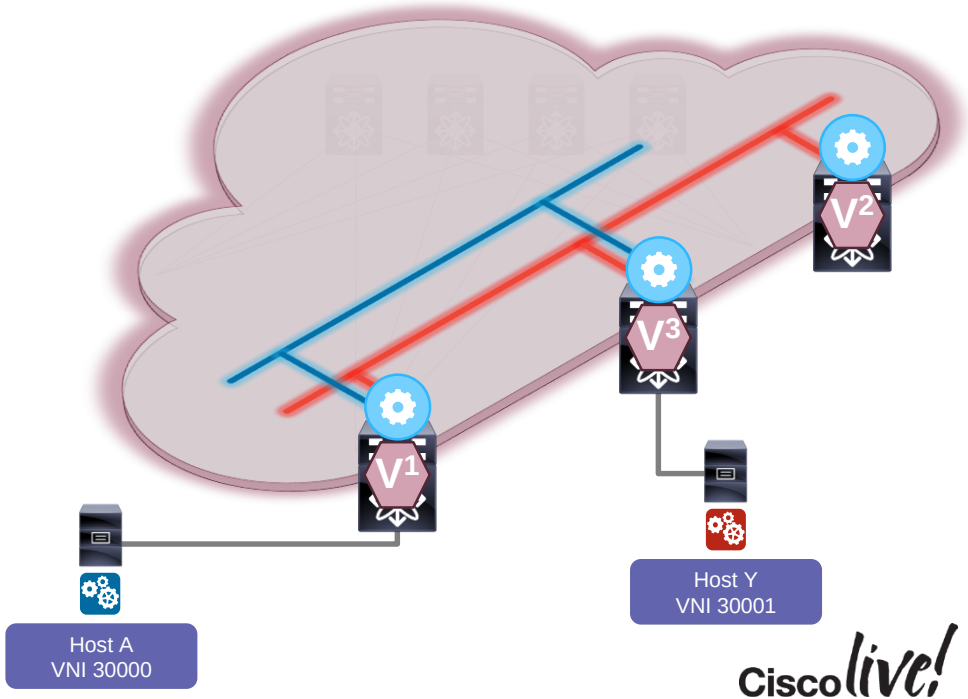
Cisco live!

Distributed IP Anycast Gateway*

VXLAN/EVPN

- Distributed Routing with IP Anycast Gateway (Integrated Route/Bridge IRB)
  - Routing between VNI (Different Subnet)
  - Bridging within VNI (Same Subnet)
- Inter-VXLAN Routing Leaf/Access Layer
  - All Leafs share gateway IP and MAC for a Subnet (No HSRP)
  - A Host will always find its Gateway directly attached anywhere it moves

*Requires EVPN Control-Plane.



FYI

Distributed IP Anycast Gateway*

VXLAN/EVPN

Configuration Example for “BLUE” (V1 & V3)

```
# 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” (V1-3)

```
# 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
```

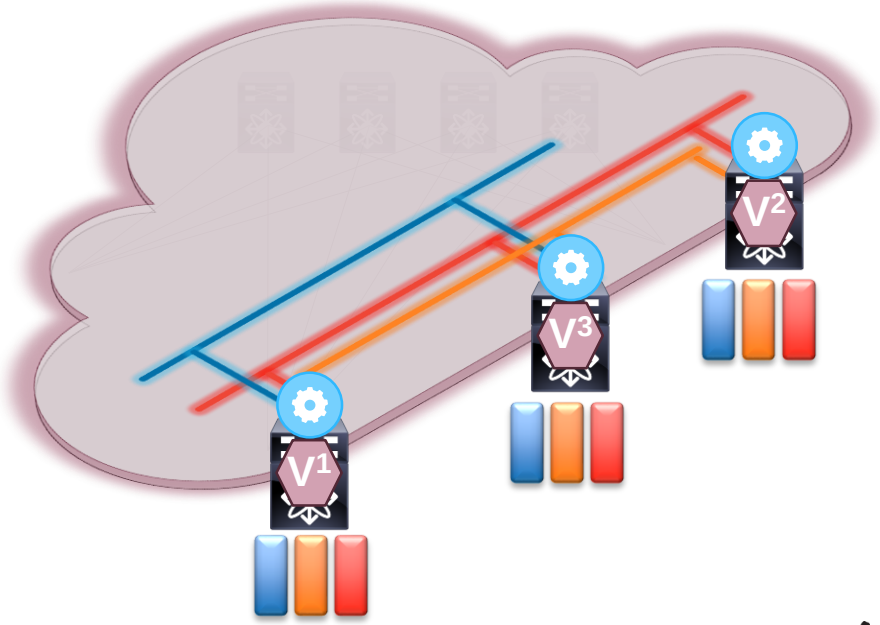
*Requires EVPN Control-Plane. VRF and BGP configuration not shown

Cisco live!

Consistent Configuration with Distributed Gateway

VXLAN/EVPN

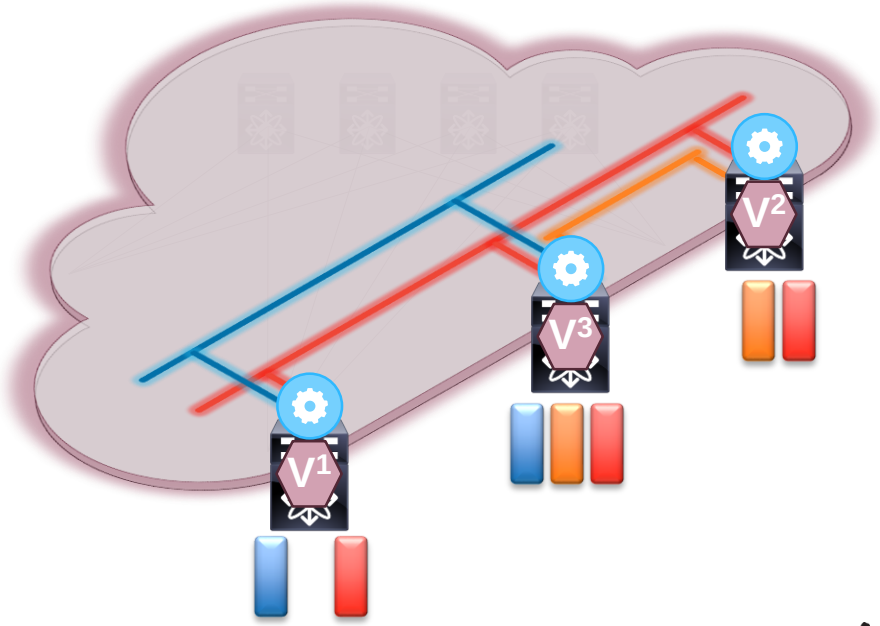
- Logical Configuration instantiated on ALL Leafs (consistent/brute force)
- ARP & MAC state of all hosted VLAN/VNI and SVI across the Network
- Flooding to ALL Leafs (all Leafs have all VLAN/VNI instantiated)



Scoping Configuration with Distributed Gateway

VXLAN/EVPN

- Logical Configuration only instantiated at respective Leaf (scoped)
- ARP & MAC state only for local hosted VLAN/VNI and SVI
- Flooding only to respective Leaf (where VLAN/VNI is instantiated)
- Host demands provisioning; two models available
  - top-down Orchestration, push to Leaf
  - bottom-up Orchestration, pull by Leaf

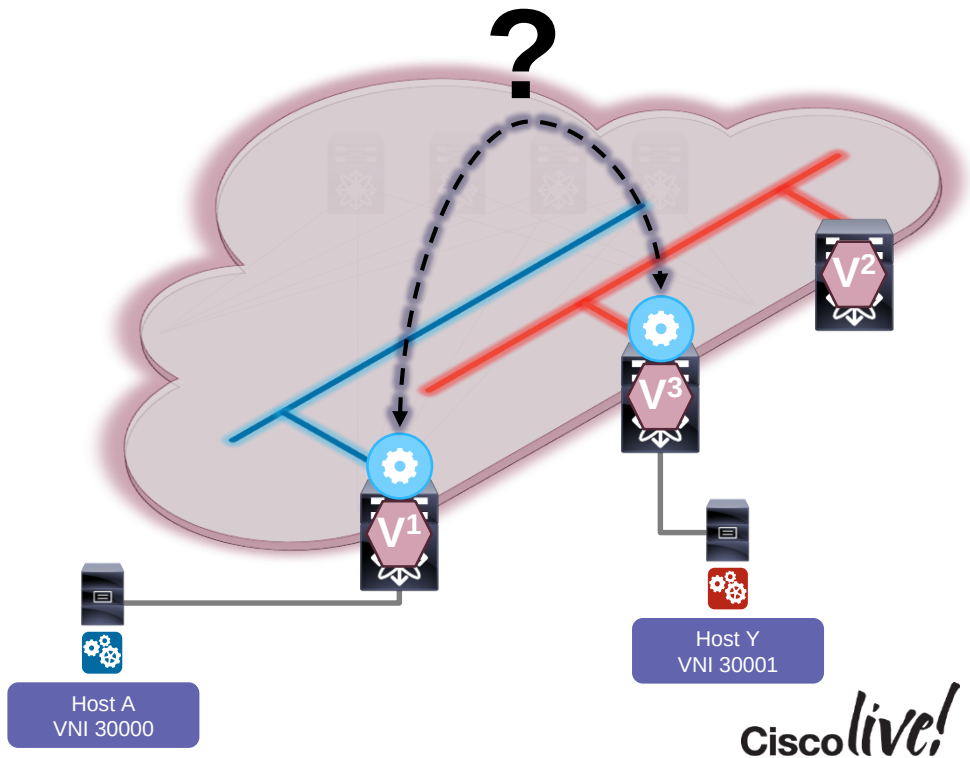


FYI

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

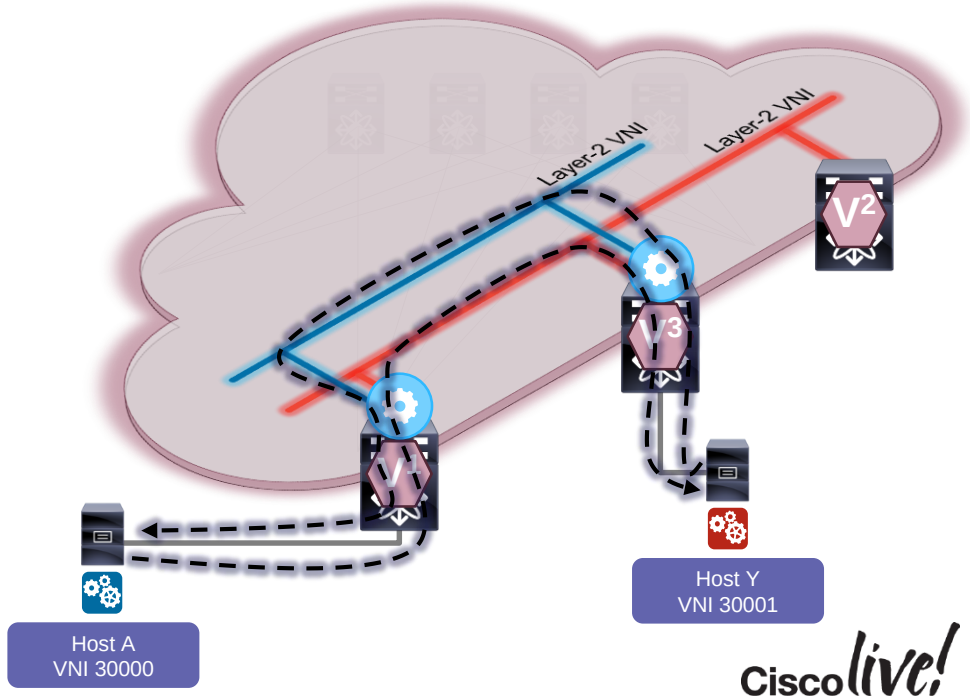


FYI

Asymmetric IRB

VXLAN Routing

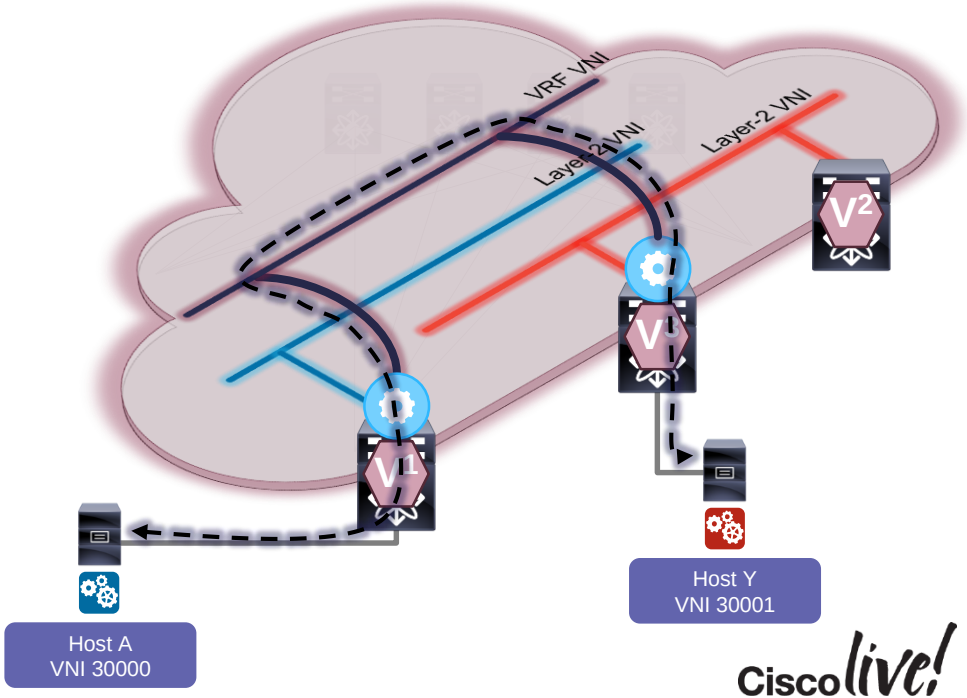
- Asymmetric
  - Similar to Inter-VLAN routing
  - Source and Destination VNI has to exist on Switch where routing happens
  - Post Routing traffic shares destination VNI with Bridged traffic
  - Not very suitable for distributed Routing
    - From Host A via VLAN/VNI “blue” routed at V^1 to VNI “red” reaching destination VLAN “red”
    - From Host Y via VLAN/VNI “red” routed at V^3 to VNI “blue reaching destination VLAN “blue”



Symmetric IRB

VXLAN Routing

- Symmetric
  - Similar to creating a Transit Segment
  - Regardless of where Source or Destination VNI exists
  - Post Routing traffic uses different VNI than Bridged traffic
  - Additional VNI for Routing traffic (per VRF)
    - From Host A via VLAN “blue” routed at V^1 to VNI “purple” reaching destination VLAN “red”
    - From Host Y via VLAN “red” routed at V^3 to VNI “purple” reaching destination VLAN “blue”
  - Used in Cisco VXLAN/EVPN

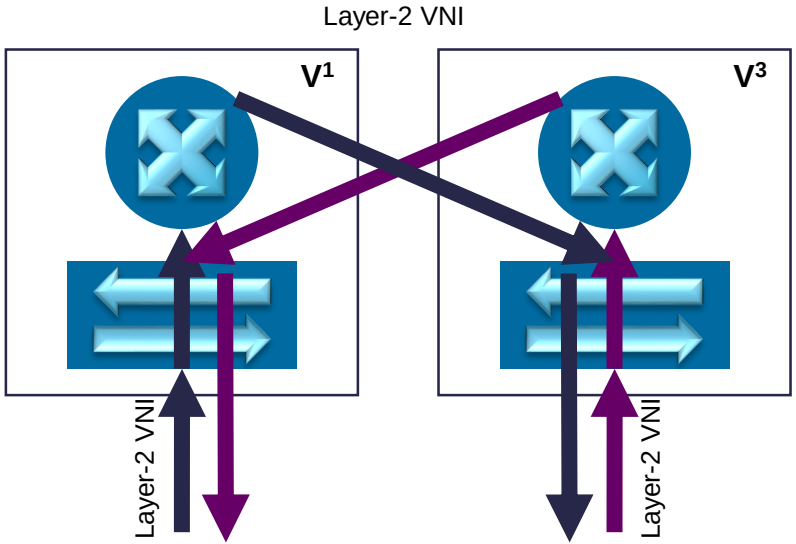


FYI

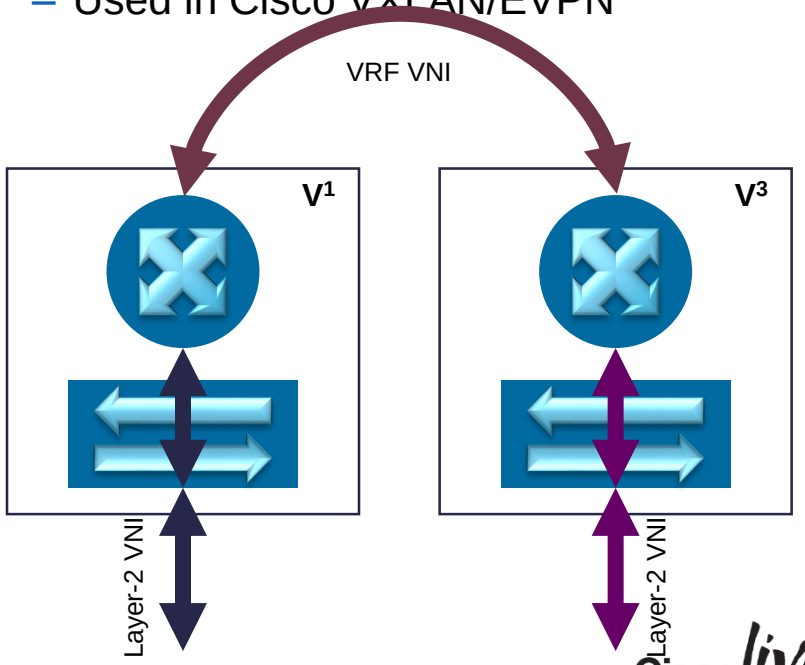
Asymmetric vs. Symmetric IRB

VXLAN Routing - Block Diagram

- Asymmetric IRB



- Symmetric IRB
  - Used in Cisco VXLAN/EVPN



Optimized Networks with VXLAN

Data Center Fabric Properties

- ✓Extended Namespace
- ✓Scalable Layer-2 Domains
- ✓Integrated Route and Bridge
- ❑Multi-Tenancy

Cisco live!

Simplified Underlay

- No IP addressing on Point-2-Point connections*
- No ARP flooding on Underlay
- Optimized Multi-Destination Topology for Scale and Convergence

Optimized Overlay

- VXLAN with distributed Default-Gateway
- End Host Discovery and Distribution (aka Control-Plane)
- Minimized Flood & Learn across Overlay

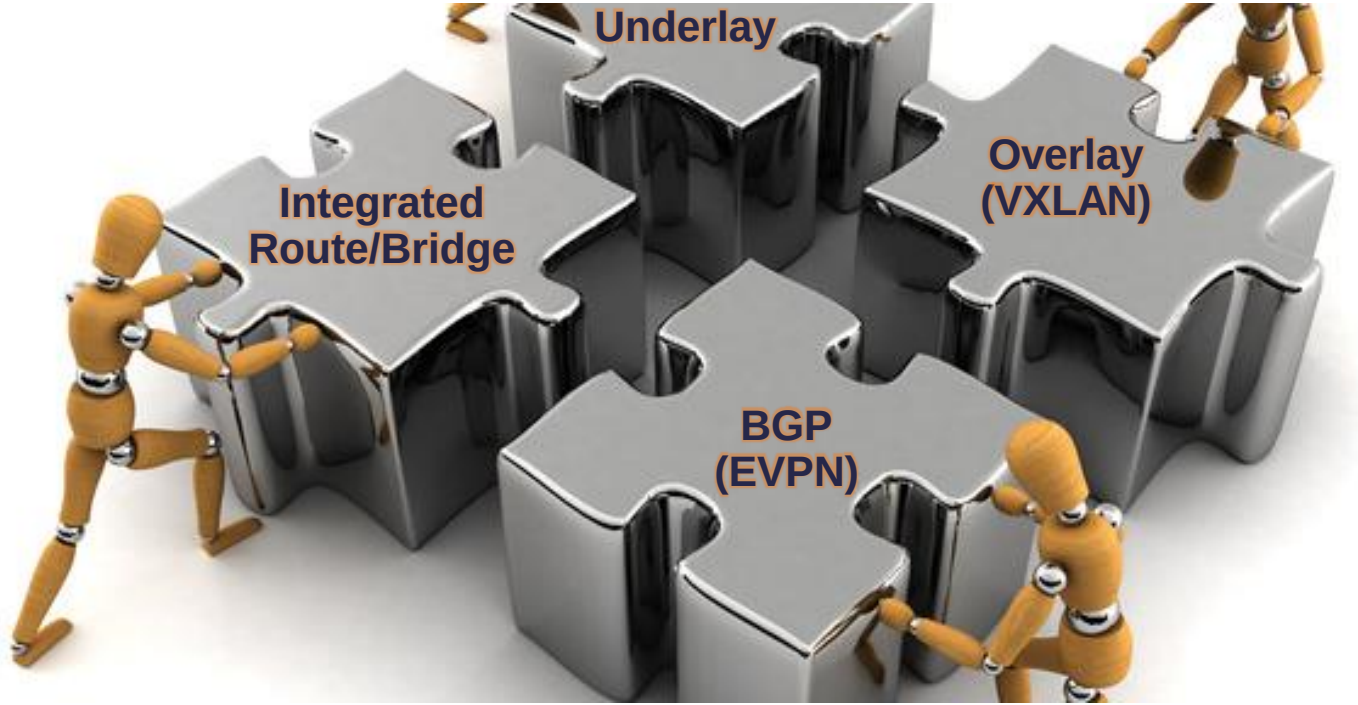
Automated Configuration

*Related to VXLAN Layer-3 Underlay (IP Un-Numbered)

Cisco live!

Getting the Puzzle Together!

Optimized Networks with VXLAN



VXLAN EVPN, Flood&Learn and Cisco FabricPath

Compared

	VXLAN/EVPN	VXLAN	FabricPath
Encapsulation	**Packet Encapsulation (PE)**	Packet Encapsulation (PE)	Frame Encapsulation (FE)
Transport Medium Requirement	**Layer-3**	Layer-3	Layer-1 (mandatory)
End-Host Reachability and Distribution	**MP-BGP EVPN**	Flood & Learn	Flood & Learn (+Conversational Learning)
End-Host Detection	**Localized Flood & Learn w/ ARP suppression**	Flood & Learn	Flood & Learn
Multi-Destination Traffic (BUM*) forwarding	**Multicast (PIM) / Ingress-Replication**	Multicast (PIM)	FabricPath IS-IS
Underlay Control-Plane	**Any Unicast Routing Protocol (static, OSPF, IS-IS, BGP)**	Any Unicast Routing Protocol (static, OSPF, IS-IS, eBGP)	FabricPath IS-IS
Unique Node Identifier	**VTEP IP**	VTEP IP	SwitchID
Standard Reference	**RFC 7348 + draft-sd-l2vnp-evpn-overlay**	RFC 7348	TRILL based (Cisco Proprietary)

*BUM: Broadcast, Unknown Unicast, Multicast

Cisco live!

VXLAN applicability evolves as the Control Plane evolves!

- Yesterday: VXLAN, yet another Overlay
  - Data-Plane only (Multicast based Flood & Learn)
- Today: VXLAN for the creation of scalable DC Fabrics – Intra-DC
  - Control-Plane, active VTEP discovery, Multicast and Unicast (Head-End Replication)
- Future: VXLAN for DCI – Inter-DC
  - DCI Enhancements (ARP caching/suppress, Multi-Homing, Failure Domain isolation, Loop Protection etc.)

What is the Elephant in the Room?

Cisco live!

Note sure if it is a Elephant

VXLAN for Interconnecting Networks

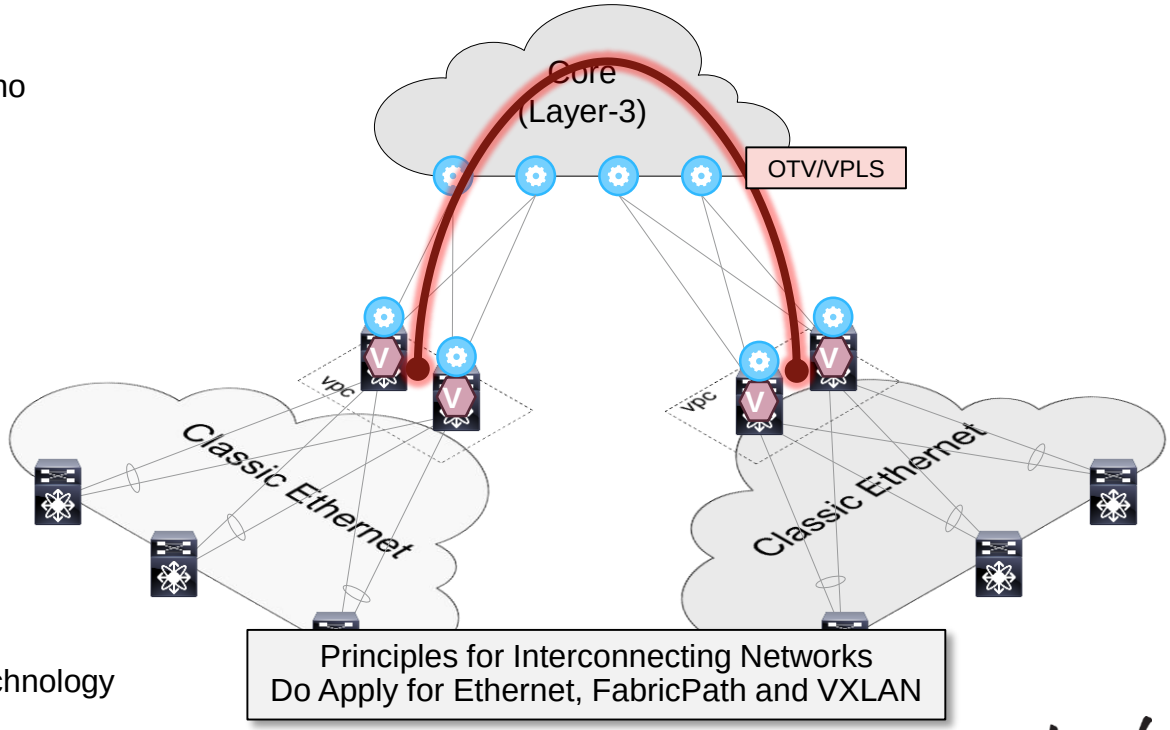
Cisco live!

FYI

Principles of Interconnecting Networks at Layer-2 (1)

Inter-Pod Connectivity

- Control-Plane
  - Learn and Distribute MAC information (no Flood&Learn)
- Multi-Homing
  - Automated Multi-Homing for Resiliency
- Loop Prevention
  - Using redundant Path
  - Providing Loop protection
- Fault Containment
  - Separate Control-Plane information
  - Limit Flood (ARP caching)
- Transport Agnostic
  - Can leverage literally any Transport Technology

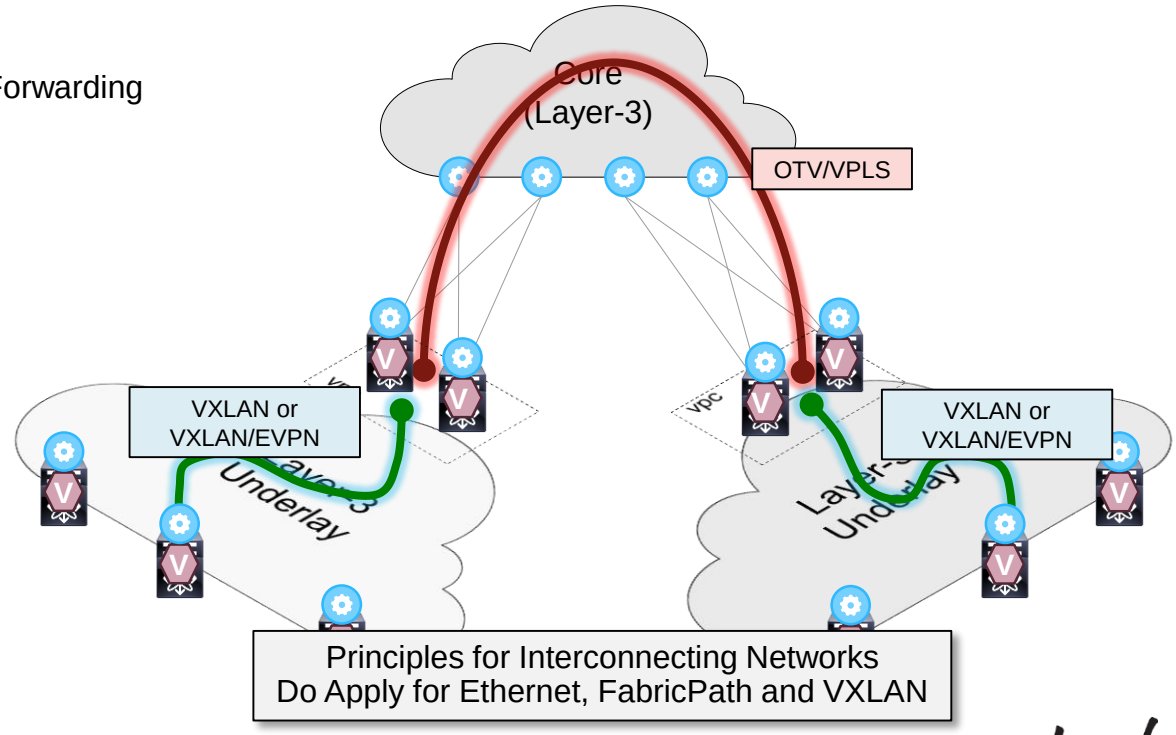


FYI

Principles of Interconnecting Networks at Layer-2 (2)

Inter-Pod Connectivity

- Simplified Transport Requirement
  - Multicast dependent and independent Forwarding of BUM* Traffic (no hairpin)
- Multicast Optimization
  - Offers optimized Multicast Forwarding
- Path Diversity
  - Flow based Entropy
- Multi-Site
  - Provides Site to Multi-Site connectivity



FYI

Principles of Interconnecting Networks at Layer-2

Inter-Pod Connectivity

		Control-Plane	Multi-Homing	Loop Prevention	Fault Containment	Transport Agnostic	Multicast Optimization	Path Diversity	Multi-Site
Good	FabricPath	✗	✓[1]	✓✓	✗	✗	✗	✓	✗
	VXLAN (Flood&Learn)	✗	✓[1]	✓✓	✗	✓	✗	✓✓	✗
Better	VXLAN-EVPN	✓✓	✓[1]	✓✓	✗	✓	✓	✓✓	✗
	VPLS	✗	✓[1]	✓✓	✗	✗	✗	✓	✓
Best	OTV	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓

BRKDCT-2049 - Overlay Transport Virtualization
Brian Farnham, Technical Marketing Engineer
Thursday, 2:30p

1) Only with Multi-Chassis Link Aggregation (MC-LAG / VPC)
2) Limited Overlay Loop Prevention

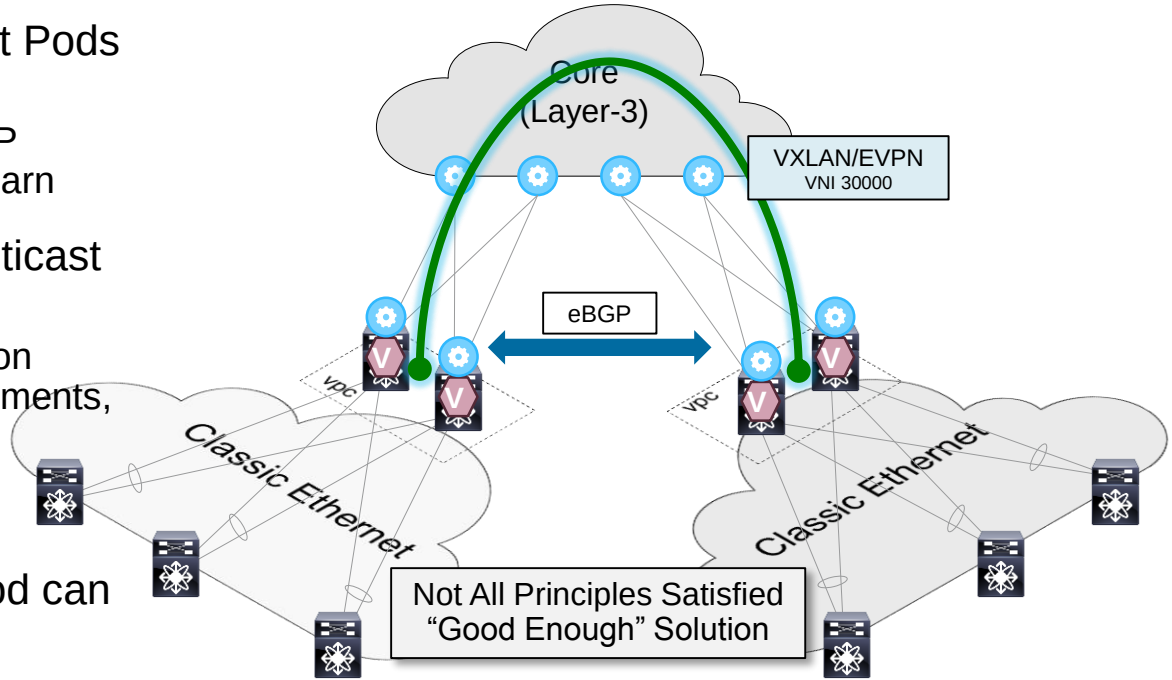
Cisco live!

FYI

Interconnecting Classic Ethernet Networks (Layer-2)

Inter-Pod Connectivity

- Interconnecting Classic Ethernet Pods with VXLAN/EVPN is possible
  - EVPN Control-Plane provides ARP suppression and avoids Flood&Learn
- Ingress Replication to avoid Multicast requirement in Underlay
  - Depends on various communication requirement factors (traffic requirements, amount of Sites, etc)
- Multi-Homing requires VPC(!)
- Loop in one Classic Ethernet Pod can influence the other Pod(s)

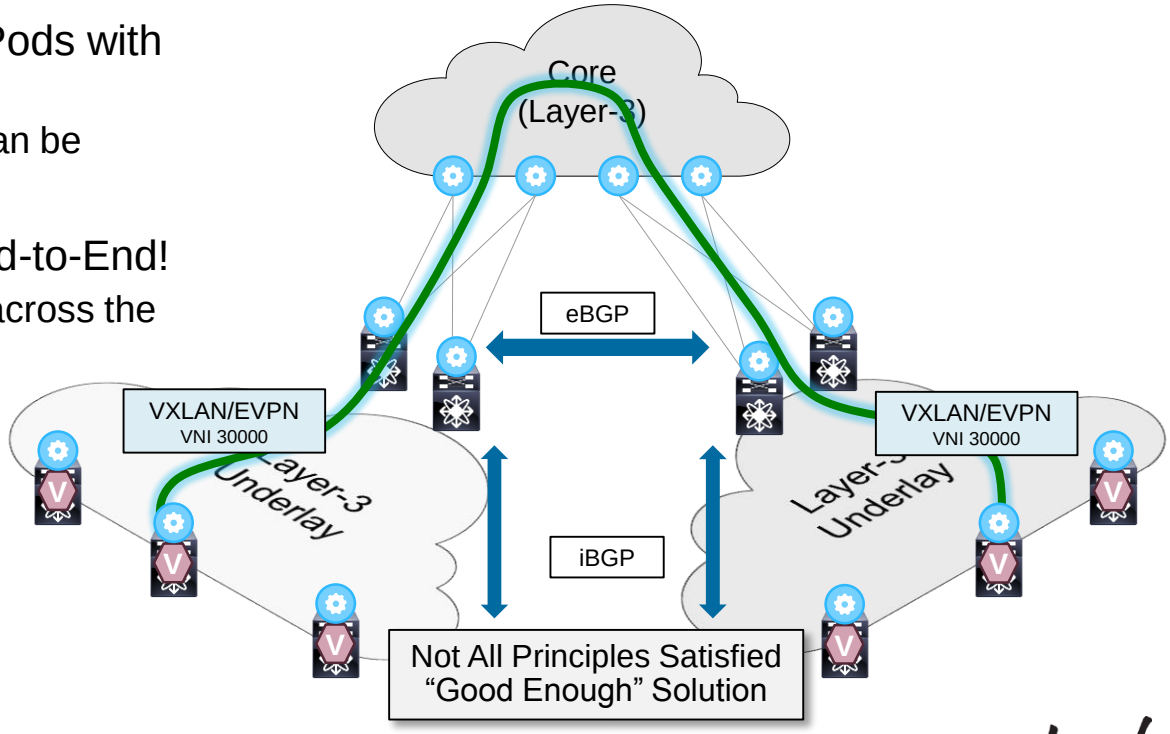


FYI

Interconnecting VXLAN Networks (Layer-2)

Inter-Pod Connectivity

- Interconnecting VXLAN/EVPN Pods with VXLAN/EVPN is possible
  - Control-Plane Domains (EVPN) can be separated (iBGP/eBGP)
- Data-Plane Encapsulation is End-to-End!
  - Leaf/TOR knows about all VTEP across the two Data Centers
  - BUM Traffic is across Pods
- Decision on Ingress Replication (Unicast) or Multicast is across all Pods being interconnected

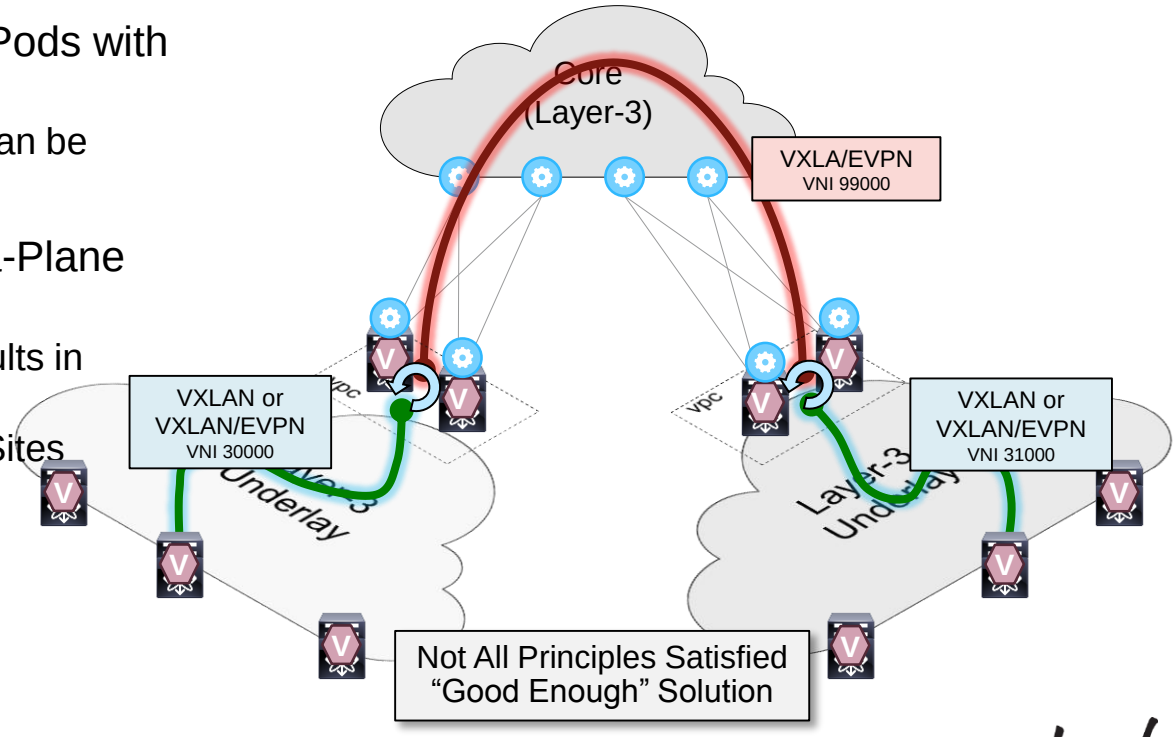


FYI

Interconnecting VXLAN Networks (Layer-3)

Inter-Pod Connectivity

- Interconnecting VXLAN/EVPN Pods with VXLAN/EVPN is possible
  - Control-Plane Domains (EVPN) can be separated (iBGP/eBGP)
- With Layer-3 interconnect, Data-Plane Encapsulation is separated
  - Routing decision at DC-Edge results in Decapsulation
  - Requires a Transit VNI between Sites
- No Layer-2 Interconnect!



Optimized Networks with VXLAN

Data Center Fabric Properties

- ✓Extended Namespace
- ✓Scalable Layer-2 Domains
- ✓Integrated Route and Bridge
- ✓Inter-Pod connectivity
- ❑Multi-Tenancy

Cisco live!

Agenda

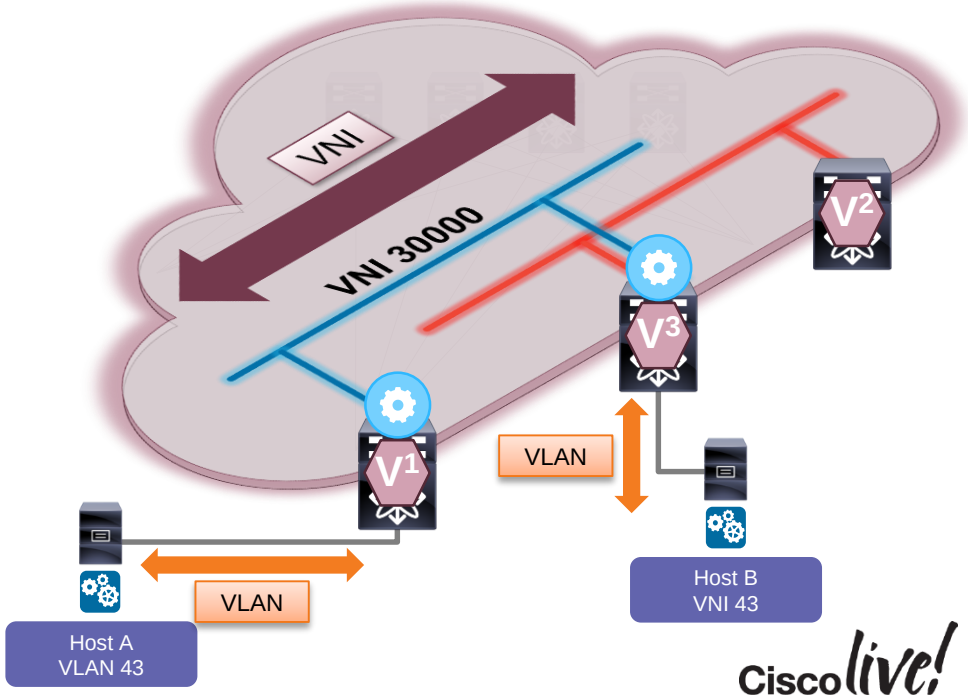
- Data Center Fabric Properties
- Optimized Networks with VXLAN
  - Overview
  - Underlay
  - Control- & Data-Plane
  - **Multi-Tenancy**
- Optimized Networks with FabricPath
- Fabric Management & Automation

Cisco live!

Layer-2 Multi-Tenancy

Multi-Tenancy

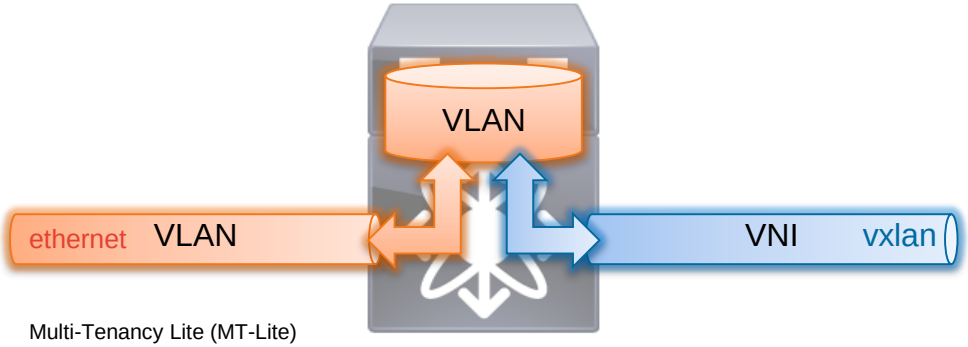
- VNI are utilized for providing isolation at Layer-2 and Layer-3 across VXLAN
  - Received frames must be mapped to specific VNI for VXLAN transport
  - The VLAN-to-VNI mapping can be performed
    - per-Switch (MT-Lite)
    - per-Port (MT-Full) level
- VLANs become locally significant on the Leaf/Port and 1:1 mapped to a VNI (global)



Layer-2 Multi-Tenancy

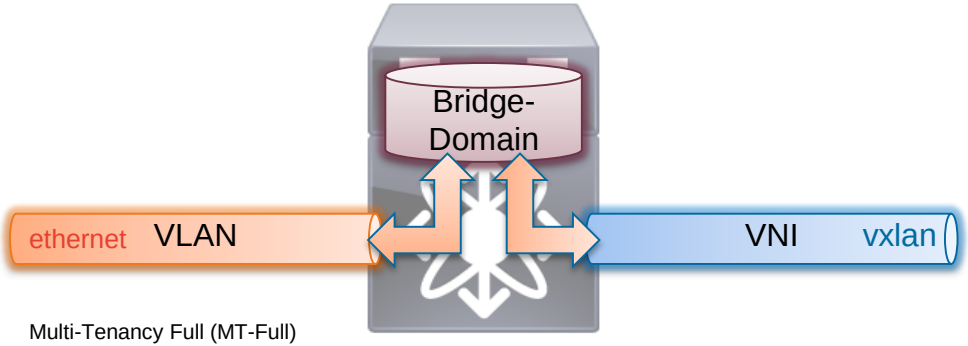
Multi-Tenancy

- Two different Interface “Mode of Operation”
- MT-Lite (Multi-Tenancy Lite)
  - VLAN to Segment ID mapping



Multi-Tenancy Lite (MT-Lite)

- MT-Full (Multi-Tenancy Full)
  - Leverages Ethernet Flow Point (EFP) or Virtual Services Instance (VSI) approach
  - Brings flexibility for Multi-Encap Gateway (VLAN to VXLAN, MPLS etc)

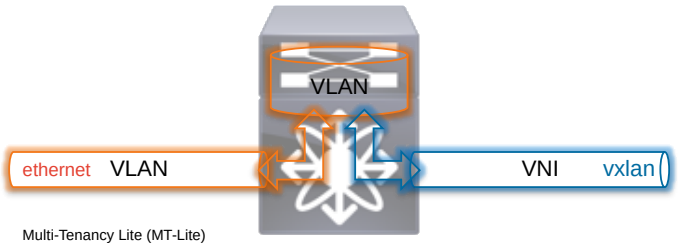


Multi-Tenancy Full (MT-Full)

Layer-2 Interface “Mode of Operation” (1)

Multi-Tenancy

- MT-Lite configuration is simplified Method of 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!



Multi-Tenancy Lite (MT-Lite)

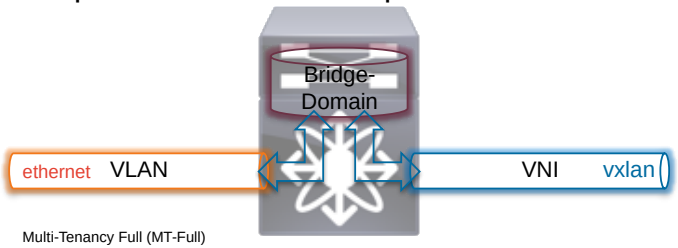
Configuration Example MT-Lite

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

# Layer-2 Interface Configuration; Trunk (MT-Lite)
interface Ethernet 1/8
 switchport mode trunk

# Layer-2 Interface Configuration; Access (MT-Lite)
interface Ethernet 1/12
 switchport mode access
 switchport access vlan 43

# Layer-3 Instance (MT-Lite)
interface Vlan 43
 ip address x.x.x.1/24
```

Layer-2 Interface “Mode of Operation” (2)

Multi-Tenancy

- MT-Full configuration is extended to allow Per-Port VLAN to Bridge-Domain mapping. Bridge-Domain will be mapped to VXLAN VNI
  - VLAN to VNI configuration on a per-Port based
  - VLAN becomes “Port Local Identifier”
  - Bridge-Domain becomes “Switch Local Identifier”
  - VNI becomes “Network Global Identifier”
  - 4k VLAN limitation resides only on a per-Dot1Q Trunk
  - Independent from VLAN Space!



Multi-Tenancy Full (MT-Full)

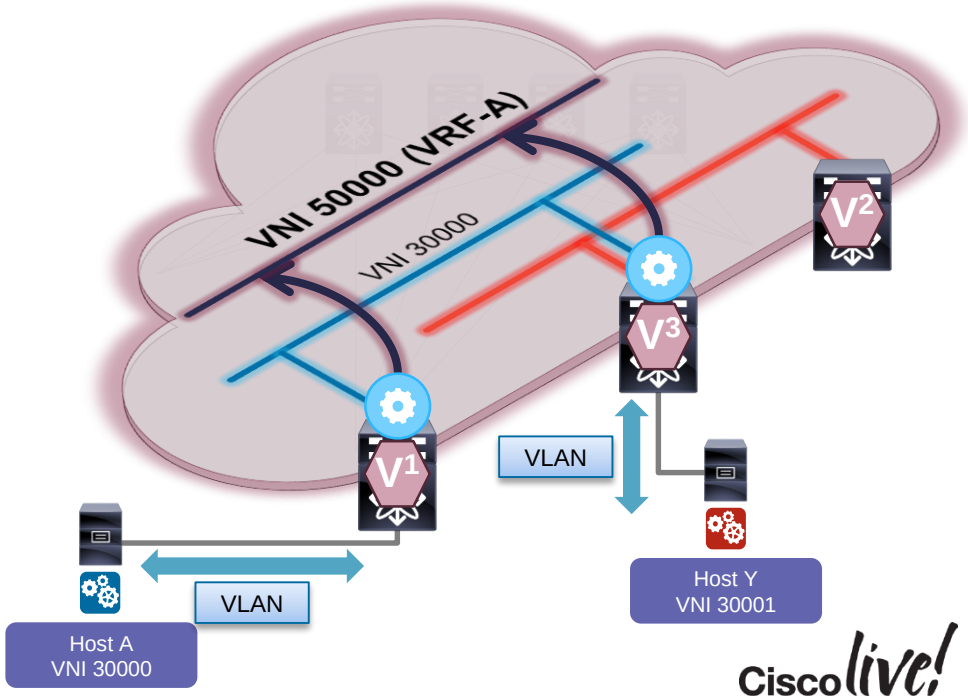
Configuration Example MT-Full

```
# VLAN to VNI mapping (MT-Full)
vni 30000

bridge-domain 100
 member vni 30000

encapsulation profile vni MyProfile
 dot1q 43 vni 30000

# Interface Configuration (MT-Full)
interface Ethernet 1/12
 no switchport
 service instance 1 vni
  encapsulation profile MyProfile default

# Layer-3 Instance (MT-Full)
interface Bdi 100
 ip address x.x.x.1/24
```

Layer-3 Multi-Tenancy

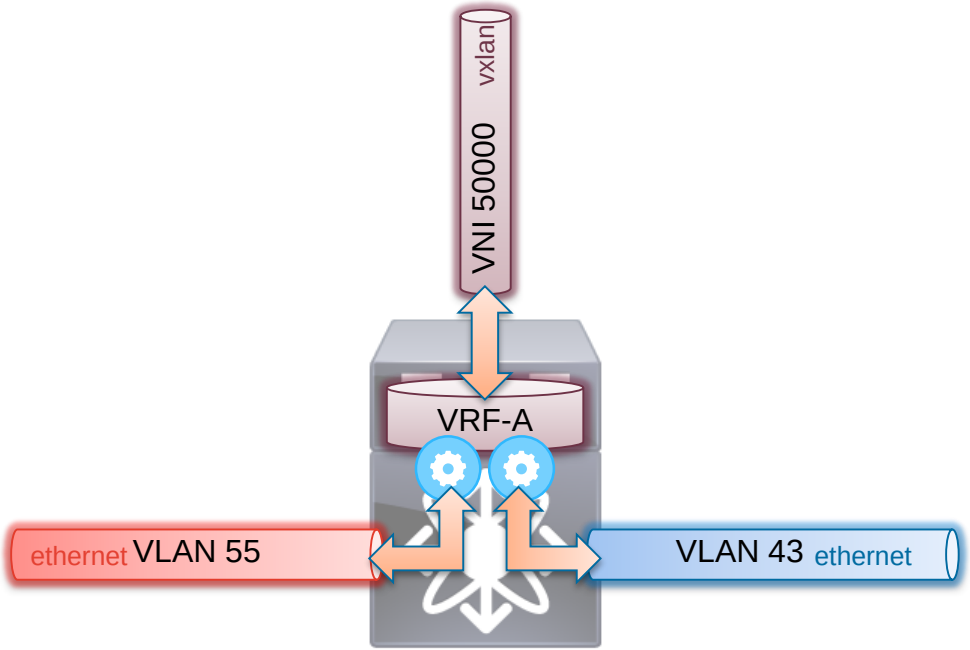
Multi-Tenancy

- VNI utilized for providing isolation at Layer-2 and Layer-3 across VXLAN
  - Received frames must be mapped to specific VNI for VXLAN transport
  - The VLAN-to-VNI mapping is performed on Routing
- All Routed Traffic uses the VNI assigned to the VRF



Layer-3 VRF “Mode of Operation”

Multi-Tenancy



Configuration Example VRF-A with 2 Subnet

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

# Layer-3 Instance (MT-Lite)
interface vlan 43
 vrf member VRF-A
 ip address 11.11.11.1/24
 fabric forwarding mode anycast-gateway

interface vlan 55
 vrf member VRF-A
 ip address 98.98.98.1/24
 fabric forwarding mode anycast-gateway
```

Optimized Networks with VXLAN

Data Center Fabric Properties

- ✓Extended Namespace
- ✓Scalable Layer-2 Domains
- ✓Integrated Route and Bridge
- ✓Multi-Tenancy

Cisco live!

Agenda

- Data Center Fabric Properties
- Optimized Networks with VXLAN
  - Overview
  - Underlay
  - Control- & Data-Plane
  - Multi-Tenancy
- **Optimized Networks with FabricPath**
- Fabric Management & Automation

Fabric Control Plane

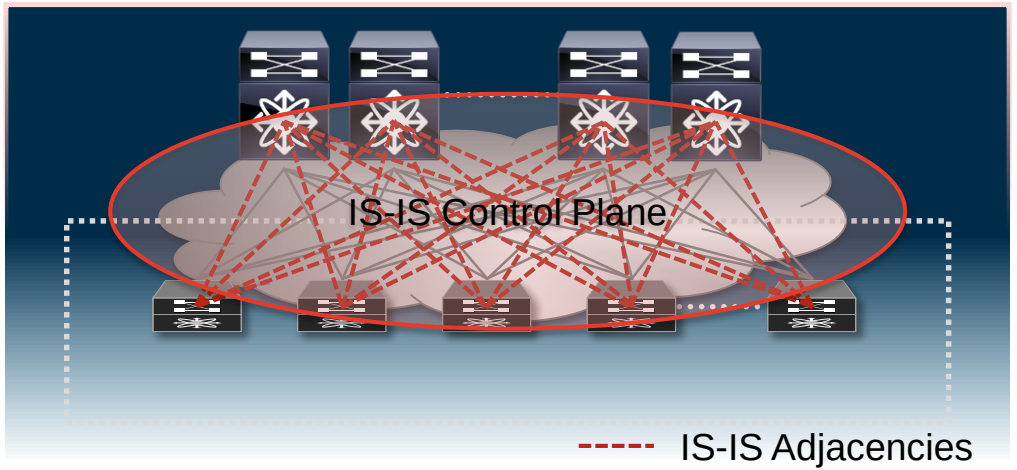
Optimized Networks with FabricPath

IS-IS for fabric link state distribution

- Fabric node reachability for overlay encapsulation
- Building multi-destination trees for multicast and broadcast traffic
- Quick reaction to fabric link/node failure (Layer-2 BFD)
- Enhanced for mesh topologies

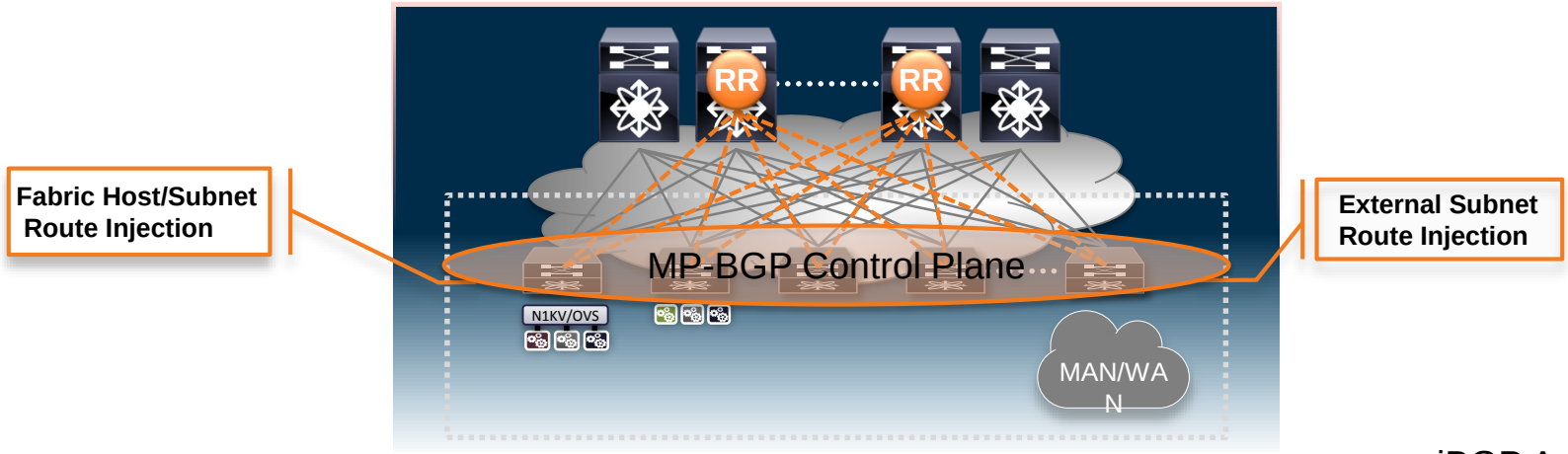
Fabric Control Protocol doesn't distribute

- Host Routes
- Host originated control traffic
- Server subnet information



Host and Subnet Route Distribution

Optimized Networks with FabricPath

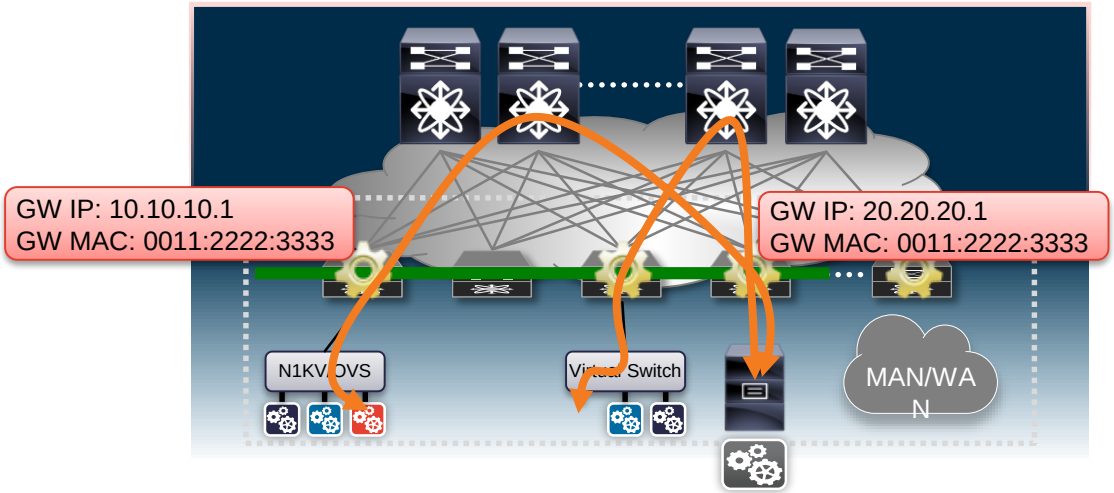


- Host Route Distribution decoupled from the Fabric link state protocol
- Use MP-BGP on the leaf nodes to distribute internal host/subnet routes and external reachability information
- MP-BGP enhancements to carry up to 1.2 Million routes and reduce convergence time

Note: Route-Reflectors deployed for scaling purposes

Distributed Gateway at the Leaf

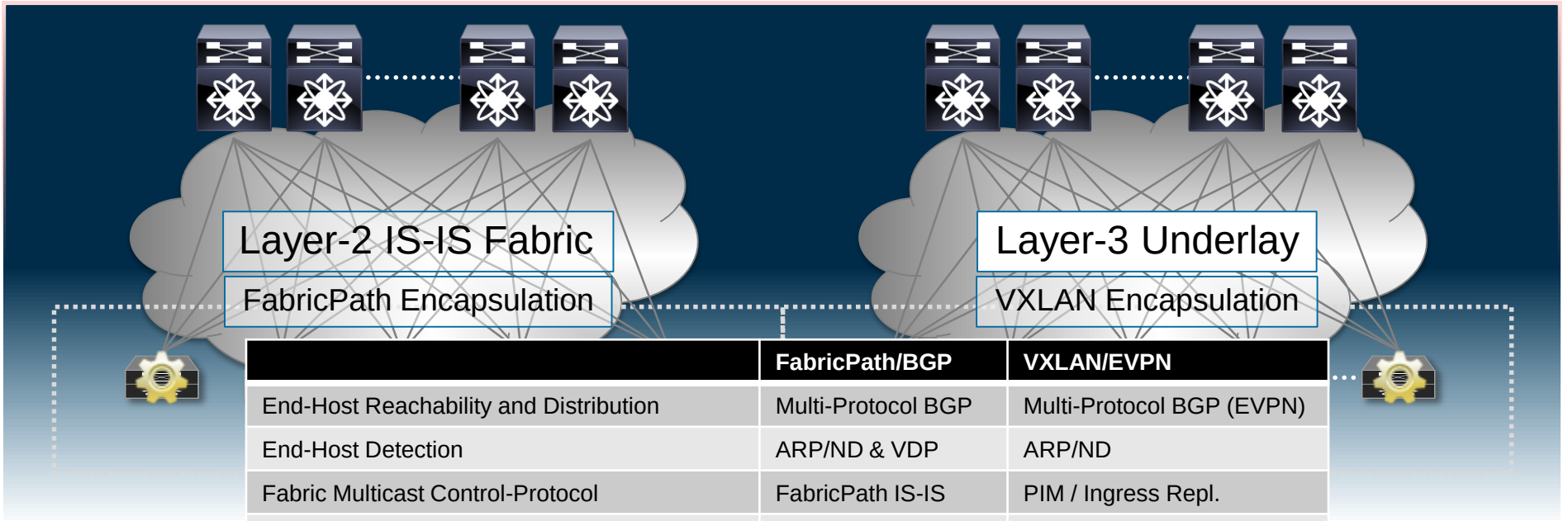
Optimized Networks with FabricPath



- Any Subnet anywhere => Any Leaf can instantiate ANY Subnet
  - All Leafs share gateway IP and MAC for a Subnet **(No HSRP)**
  - ARPs are terminated on Leafs, No Flooding beyond Leaf
- Facilitates VM Mobility, workload distribution, arbitrary clustering
- Seamless Layer-2 or Layer-3 communication between physical hosts and virtual machines

Forwarding Modes Comparison

Optimized Networks with FabricPath



	FabricPath/BGP	VXLAN/EVPN
End-Host Reachability and Distribution	Multi-Protocol BGP	Multi-Protocol BGP (EVPN)
End-Host Detection	ARP/ND & VDP	ARP/ND
Fabric Multicast Control-Protocol	FabricPath IS-IS	PIM / Ingress Repl.
Fabric Topology Control-Plane	FabricPath IS-IS	Layer-3 OSPF, IS-IS, BGP
Leaf IP/MAC binding distribution within Fabric	FabricPath IS-IS	Multi-Protocol BGP (EVPN)

VXLAN/EVPN & Cisco FabricPath/BGP

Compared

	VXLAN/EVPN		FabricPath/BGP
Encapsulation	Packet Encapsulation (PE)		Frame Encapsulation (FE)
Transport Medium Requirement	Layer-3		Layer-1 (mandatory)
End-Host Reachability and Distribution	MP-BGP EVPN		MP-BGP L3VPN
End-Host Detection	Localized Flood & Learn w/ ARP suppression		ARP/ND & VDP
Multi-Destination Traffic (BUM*) forwarding	Multicast (PIM) / Ingress-Replication		FabricPath IS-IS
Underlay Control-Plane	Any Unicast Routing Protocol (static, OSPF, IS-IS, BGP)		FabricPath IS-IS
Unique Node Identifier	VTEP IP		SwitchID
Standard Reference	RFC 7348 + draft-sd-l2vnp-evpn-overlay		TRILL based (Cisco Proprietary)

*BUM: Broadcast, Unknown Unicast, Multicast

Forwarding Modes Comparison

	Proxy-Gateway (FabricPath/BGP)	Anycast-Gateway (FabricPath/BGP)	Anycast-Gateway + ARP Suppression (VXLAN/EVPN)
VLAN/Subnets stretched between Leafs	✓	✓	✓
Common Anycast GW IP across Leafs	✓	✓	✓
Common Anycast GW MAC across Leafs	✓	✓	✓
Use local Proxy-ARP/ND	✓ (respond to ARP/ND only if the destination is available in the RIB)	✗	✗ (impersonated ARP response from Host)
ARP Flooding in Layer-2 Domain	✗	✓ (floods across Network)	✓ (ARP suppression)
Intra-Subnet forwarding	Always routed (TTL decrement)	Bridged + Flood&Learn	Bridged + Protocol Learning
Silent Host Discovery	✗	✓	✓
Non-IP Forwarding	✓	✓	✓

Agenda

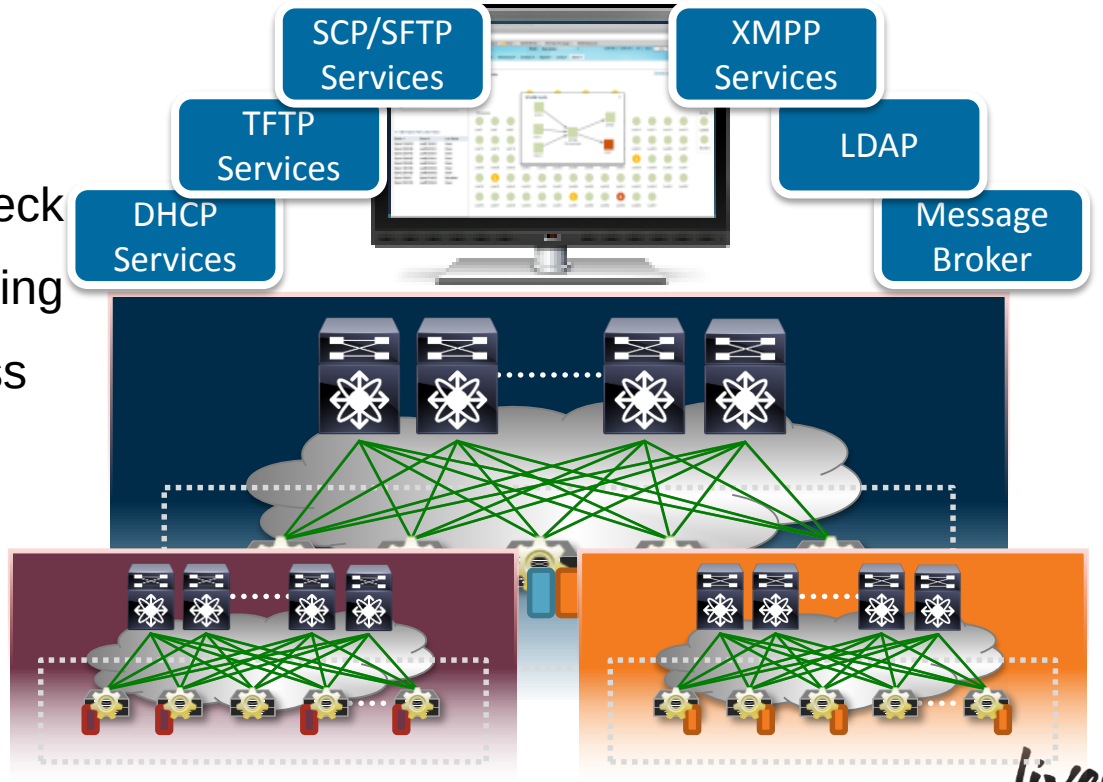
- Data Center Fabric Properties
- Optimized Networks with VXLAN
  - Overview
  - Underlay
  - Control- & Data-Plane
  - Multi-Tenancy
- Optimized Networks with FabricPath
- **Fabric Management & Automation**

Cisco live!

Simplifying Fabric Management & Optimizing Fabric Visibility

Advantages

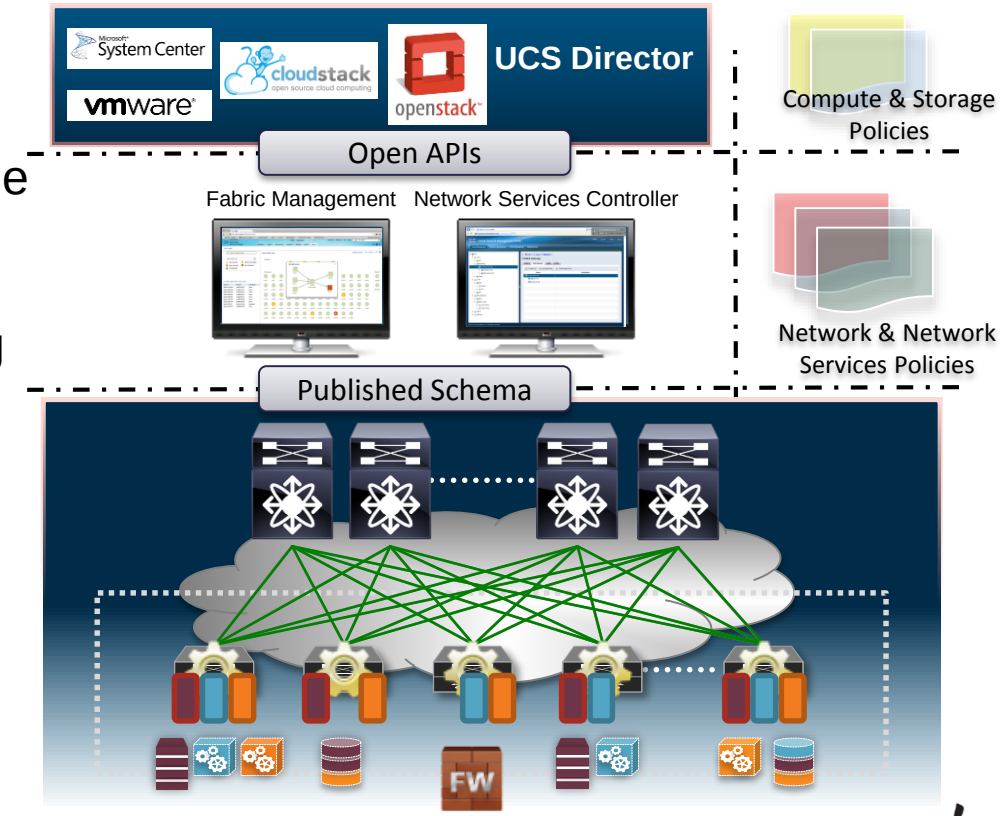
- Device Auto-Configuration
- Cabling Plan Consistency Check
- Automated Network Provisioning
- Common point of fabric access
- Tenant, Virtual Fabric & Host Visibility



Workload Automation & Open Environment

Advantages

- Any workload, Anywhere, Anytime
- Open Integration: Orchestration
- Automated Scalable Provisioning
- Workload aware fabric



How to achieve Data Center Automation

- Simplify
  - Do not start with the most difficult task (low hanging Fruits)
- Standardize
  - Find common Denominators and create Templates
- Automate repetitive Tasks
  - Use Templates for Simple Tasks and use Automation (e.g. create VLAN, SVI, VRF)
- Abstract
  - Take a step back and look at the WHOLE
  - Cisco ACI

Simplified Underlay

- No IP addressing on Point-2-Point connections*
- No ARP flooding on Underlay
- Optimized Multi-Destination Topology for Scale and Convergence

Optimized Overlay

- VXLAN with distributed Default-Gateway
- End Host Discovery and Distribution (aka Control-Plane)
- Minimized Flood & Learn across Overlay

Automated Configuration

- Automated Configuration of Tenant and Network
- Fabric Management and Troubleshooting

*Related to VXLAN Layer-3 Underlay (IP Un-Numbered)

Cisco live!

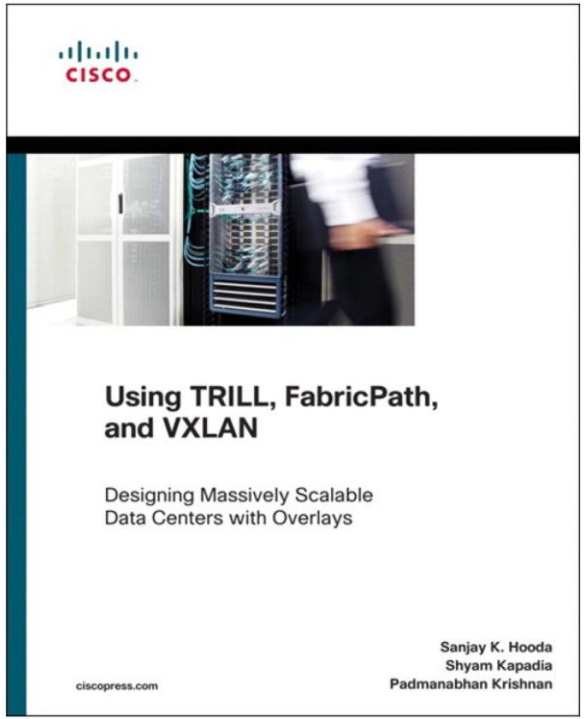
Session Marketing ☺

Session ID	Session Title	Speaker	Day / Time
LTRDCT-1224	Implementing VXLAN in Datacenter	Lilian Quan Technical Marketing Engineer	Tuesday, 9:30 Wednesday, 9:00
BRKAPP-9004	Data Center Mobility, VXLAN & ACI Fabric Architecture	Bradley Wong Principal Engineer	Tuesday, 11:15
PNLDCT-2001	Panel: Overlays in the Data Center - A Customer Perspective	Panel	Tuesday, 4:45p
BRKDCT-2328	Evolution of Network Overlays in Data Center Clouds	Victor Moreno Distinguished TME	Wednesday, 11:30
BRKDCT-2456	Building Multi-tenant DC using Cisco Nexus Switching	Elyor Khakimov Technical Marketing Engineer	Wednesday, 2:30p
BRKDCT-2404	VXLAN deployment models - A practical perspective	Victor Moreno Distinguished TME	Thursday, 9:00
BRKACI-2001	Integration and Interoperation of existing Nexus networks into an ACI architecture	Mike Herbert Principal Engineer	Thursday, 11:30
BRKDCT-2049	Overlay Transport Virtualization	Brian Farnham Technical Marketing Engineer	Thursday, 2:30p

Call to Action

- Visit the World of Solutions for
  - Cisco Campus – Overlay Demos (VXLAN, OTV, LISP) / Programmability Demo
  - Walk in Labs – LABDCT-2227 (Wednesday & Thursday @ 2:30p
  - Technical Solution Clinics
- Meet the Engineer – Wednesday & Thursday @ 10:00 – 11:00
- Lunch time Table Topics – Always Happy to have Geek Talk during Lunch
- DevNet zone related labs and sessions
  - DevNet1052 (Tuesday, 2:00p)
  - DevNet-1009 (Thursday, 1:30p)
- Recommended Reading: for reading material and further resources for this session, please visit www.pearson-books.com/CLMilan2015

Recommended Reading



Using TRILL, FabricPath, and VXLAN: Designing Massively Scalable Data Centers (MSDC) with Overlays

- Sanjay K. Hooda
- Shyam Kapadia
- Padmanabhan Krishnan

ISBN-10: 1-58714-393-3

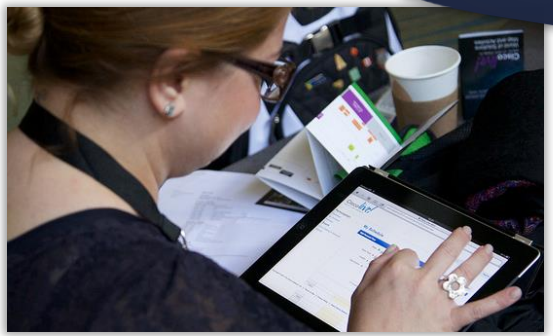
ISBN-13: 978-1-58714-393-9

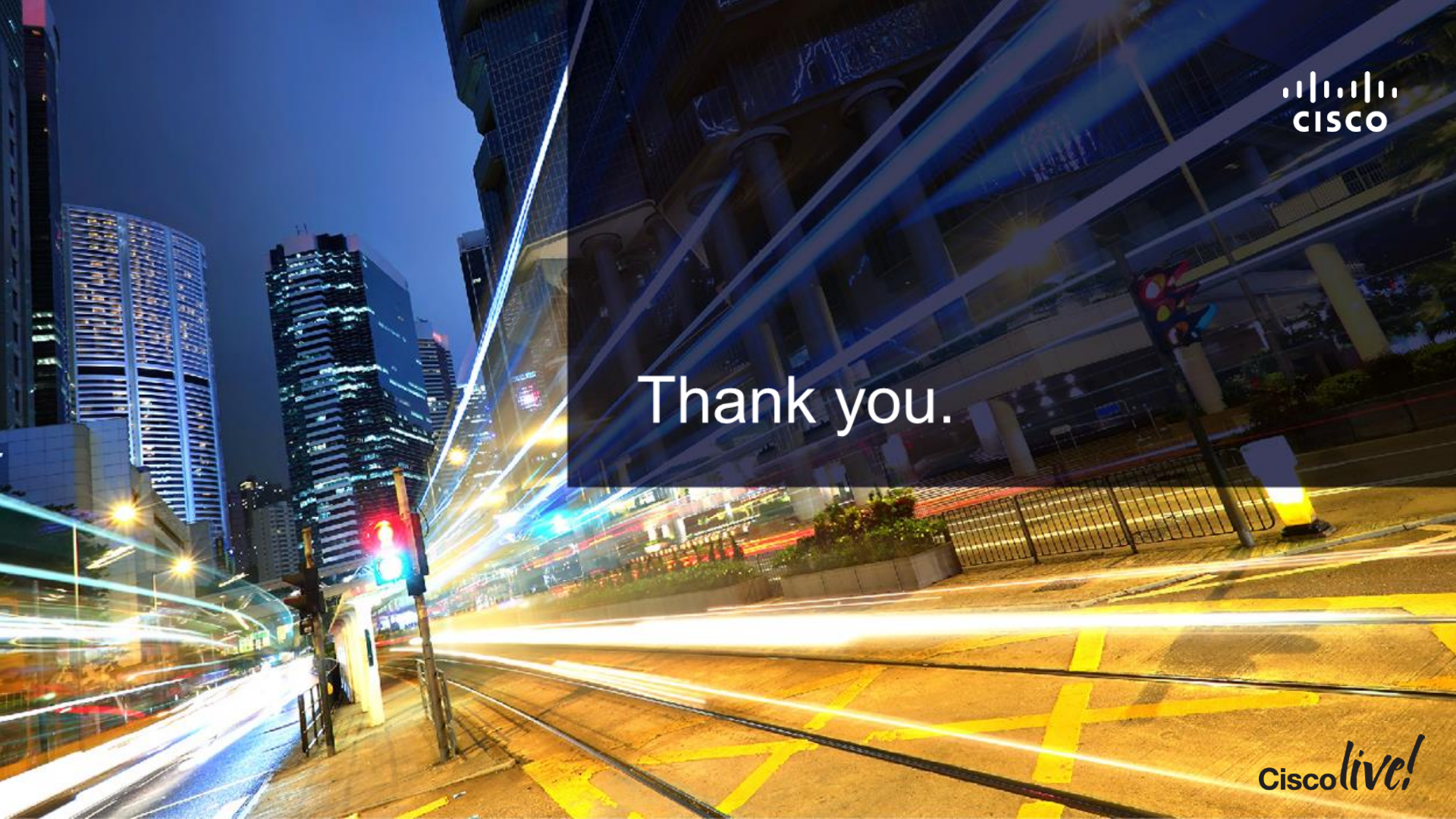
Cisco live!

Complete Your **Online Session Evaluation**

- Please complete your online session evaluations after each session. Complete 4 session evaluations & the Overall Conference Evaluation (available from Thursday) to receive your Cisco Live T-shirt.

- All surveys can be completed via the Cisco Live Mobile App or the Communication Stations




Thank you.
CISCO
Cisco live!


CISCO