We’re ready.
Are you?

Building Data Centre Networks with VXLAN BGP-EVPN

Lukas Krattiger– Principal Technical Marketing Engineer

Cisco live!

Session Objectives

- Focus on Data Centre Networks and Fabrics with Overlays
- Closer Look on Packet Encapsulation (VXLAN)
  - Encapsulation and Forwarding
  - Underlay – the Transport for the Overlay
- Closer Look on Packet Encapsulation (BGP EVPN)
  - Control-Plane – Exchanging Information
  - Optimising the Forwarding

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

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

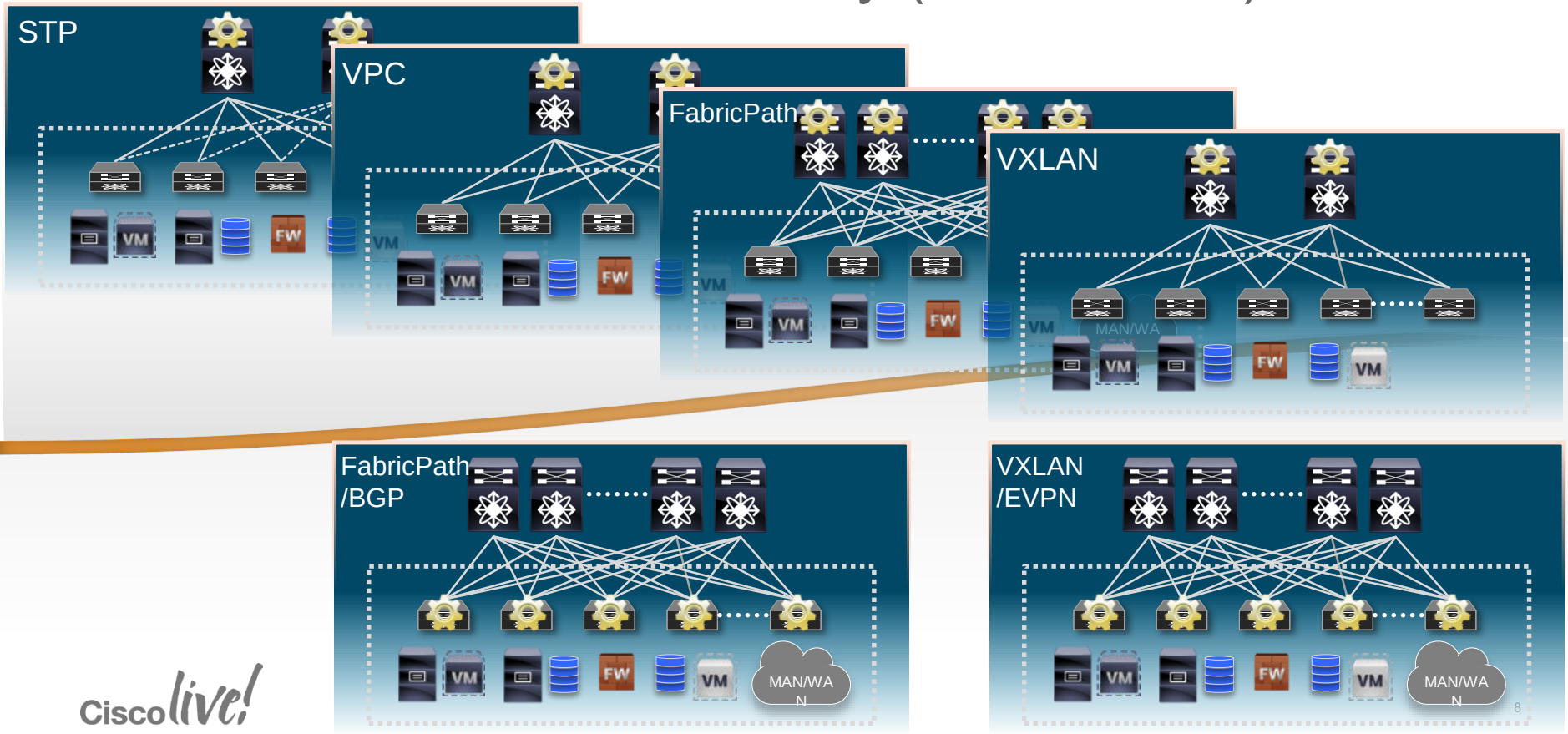
Albert Einstein

Agenda

- Introduction to Data Centre Fabrics
- VXLAN with BGP EVPN
  - Overview
  - Underlay
  - Control & Data Plane
  - Multi-Tenancy
- “Stories” and Use-Cases
- Fabric Management & Automation

Introduction to Data Centre Fabrics

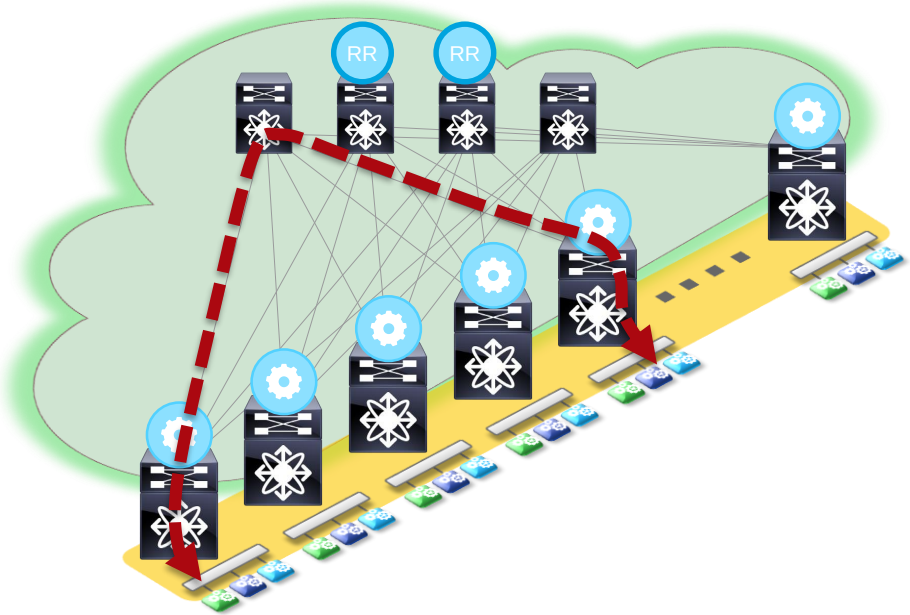
Data Centre “Fabric” Journey (Standalone)



Data Centre Fabric Properties

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

Overlay Based Data Centre Fabrics



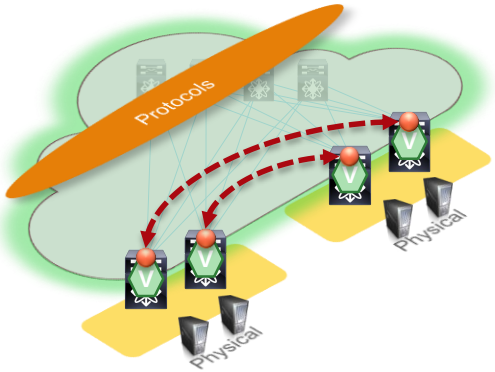
Desirable Attributes:

- Mobility
- Segmentation
- Scale
- Automated & Programmable
- Abstracted consumption models
- Full Cross Sectional Bandwidth
- Layer-2 + Layer-3 Connectivity
- Physical + Virtual

FYI

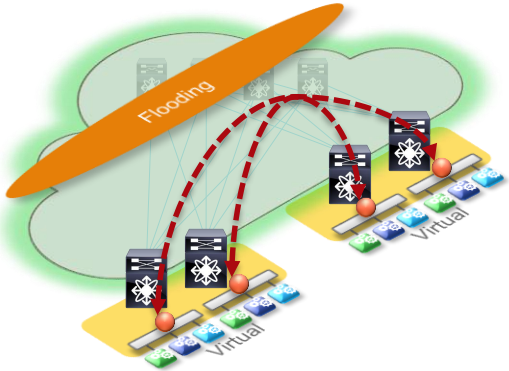
Overlay Based Data Centre: Edge Devices

Network Overlays



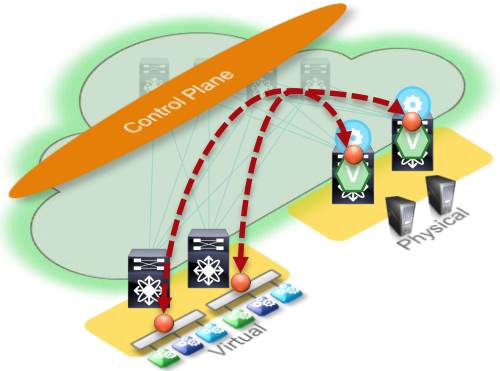
- Router/Switch end-points
- Protocols for Resiliency/Loops
- Traditional VPNs
- VXLAN, OTV, VPLS, LISP, FP

Host Overlays



- Virtual end-points only
- Single admin domain
- VXLAN, NVGRE, STT

Hybrid Overlays

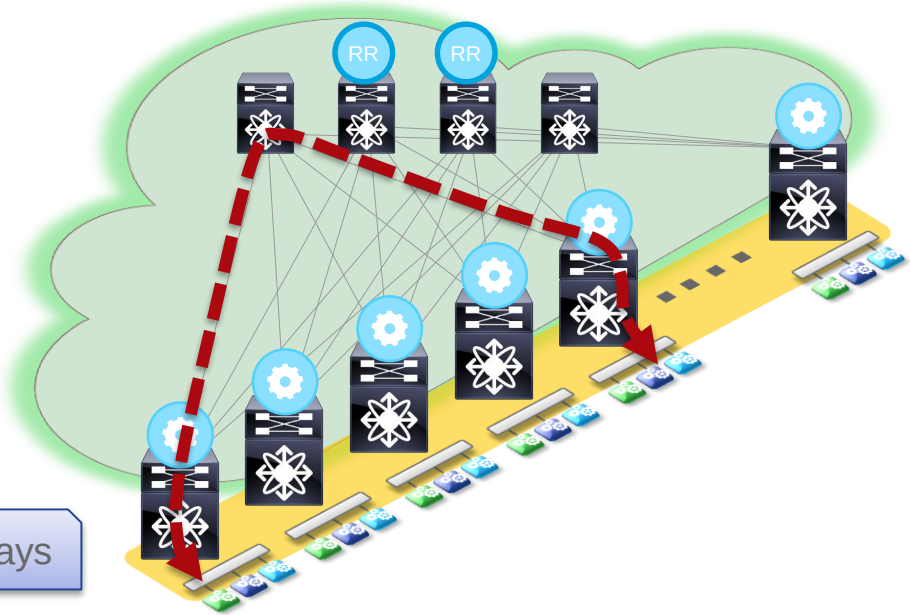


- Physical and Virtual
- Resiliency + Scale
- X-Organisations/Federation
- Open Standards

Cisco live!

Data Centre Fabric Properties

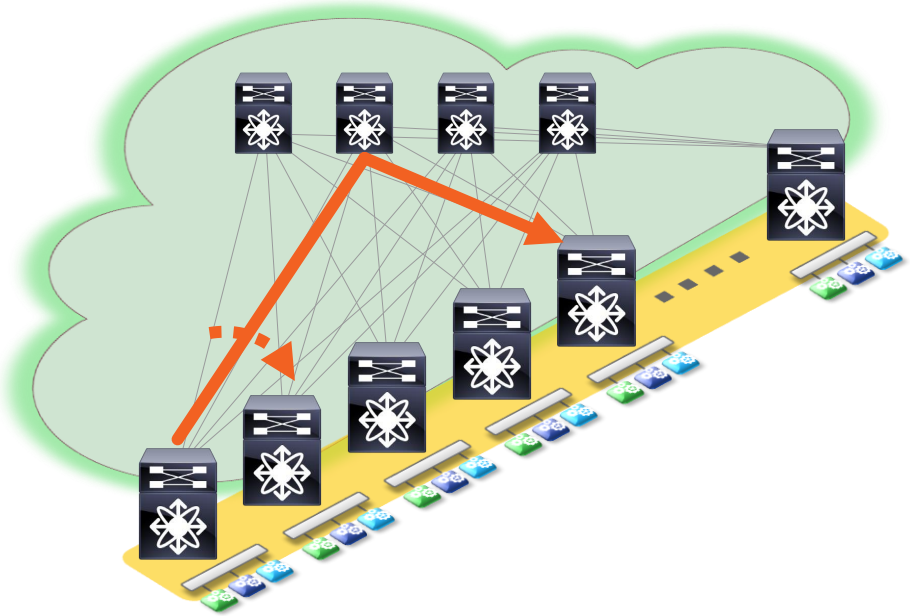
- Any subnet, anywhere, rapidly
- Reduced Failure Domains
- Extensible Scale & Resiliency
- Profile Controlled Configuration



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

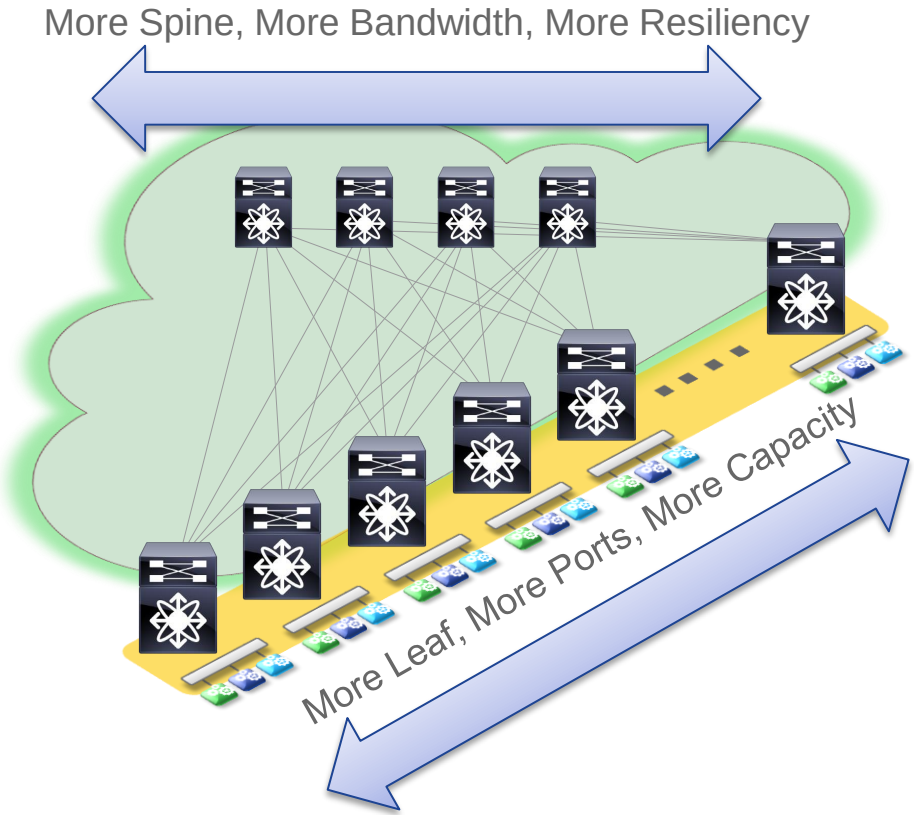
Spine/Leaf Topologies

- 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

- 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



VXLAN with BGP EVPN

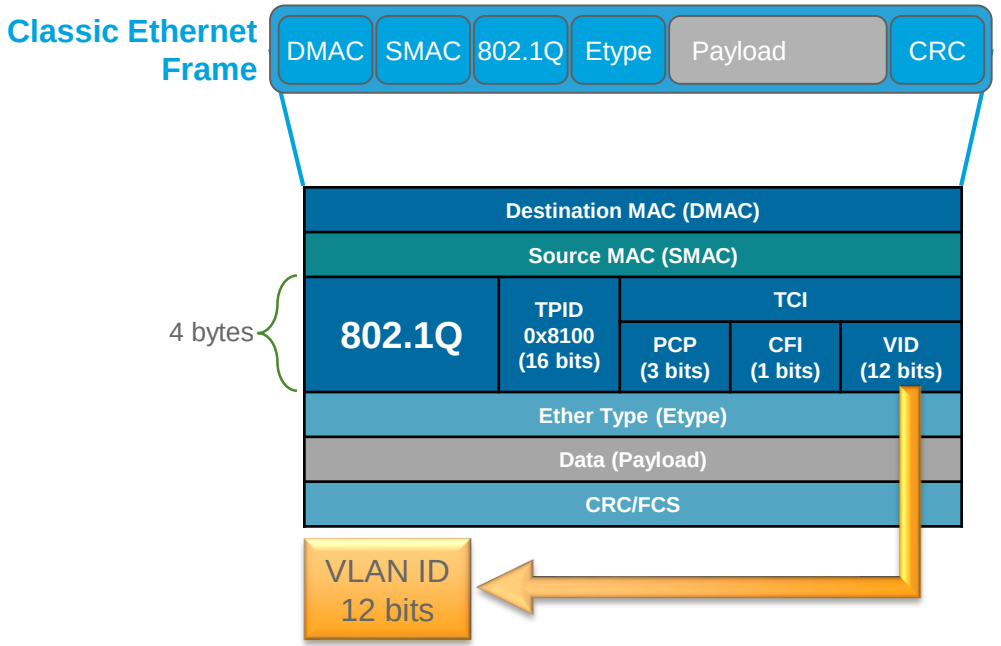
Agenda

- Introduction to Data Centre Fabrics
- VXLAN with BGP EVPN
  - **Overview**
  - Underlay
  - Control & Data Plane
  - Multi-Tenancy
- “Stories” and Use-Cases
- Fabric Management & Automation

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 Centre to 4096 VLANs

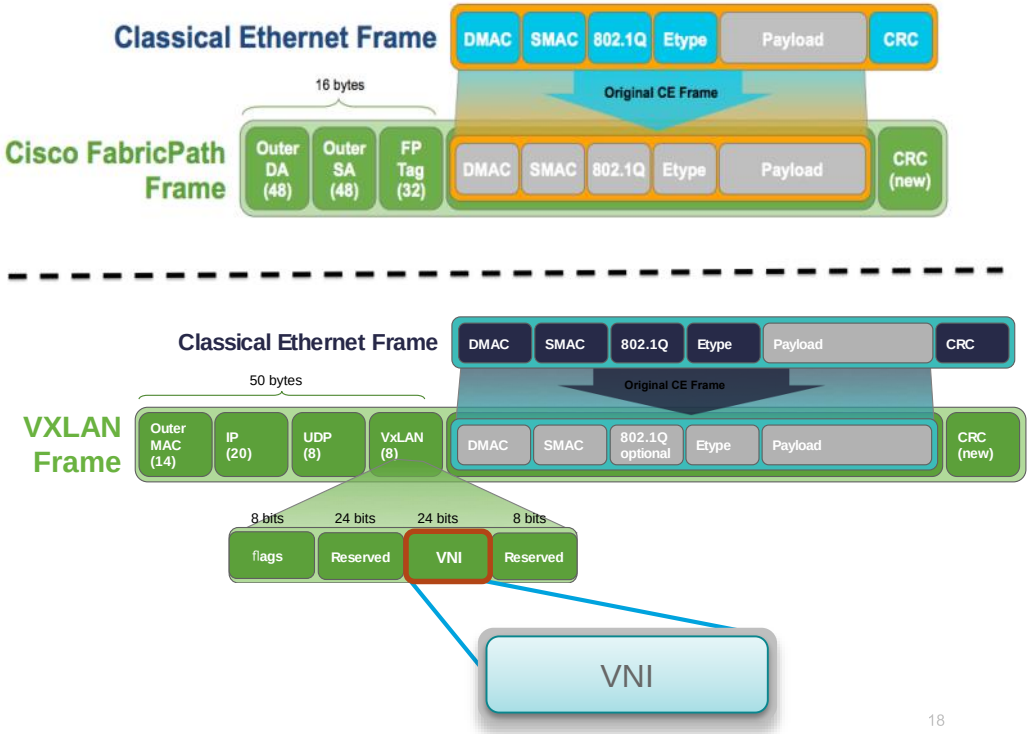


TPID = Tag Protocol Identifier, TCI = Tag Control Information, PCP = Priority Code Point, CFI = Canonical Format Indicator, VID = VLAN Identifier

Overview

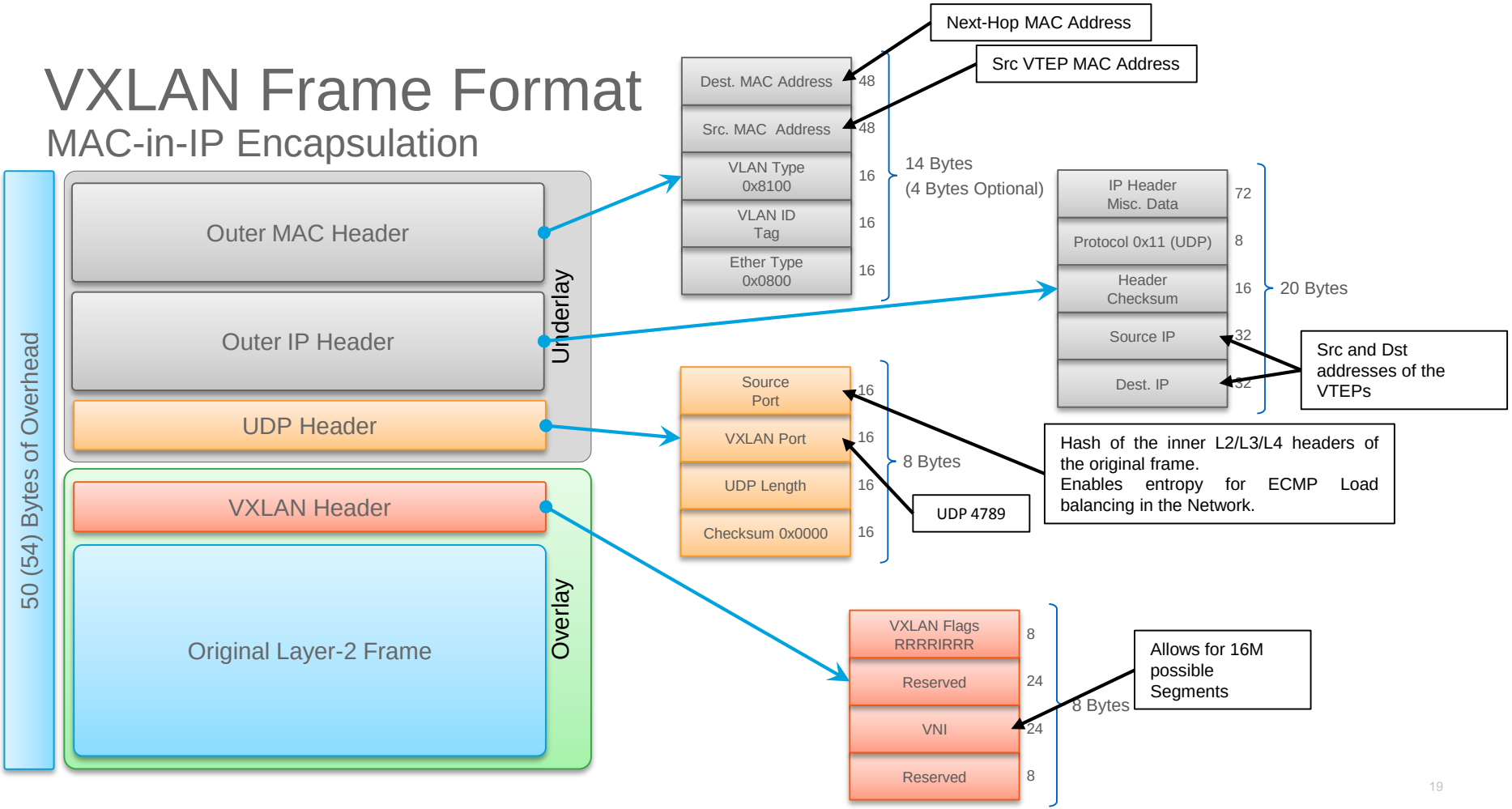
Introducing VXLAN

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



Data Centre Fabric Properties

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

Understanding Overlay Technologies

Overlay Services

- Layer 2
- Layer 3
- Layer 2 and Layer 3

Tunnel Encapsulation

Underlay Transport Network

Control Plane

- Peer Discovery mechanism
- Route Learning and Distribution
  - Local Learning
  - Remote Learning

Data Plane

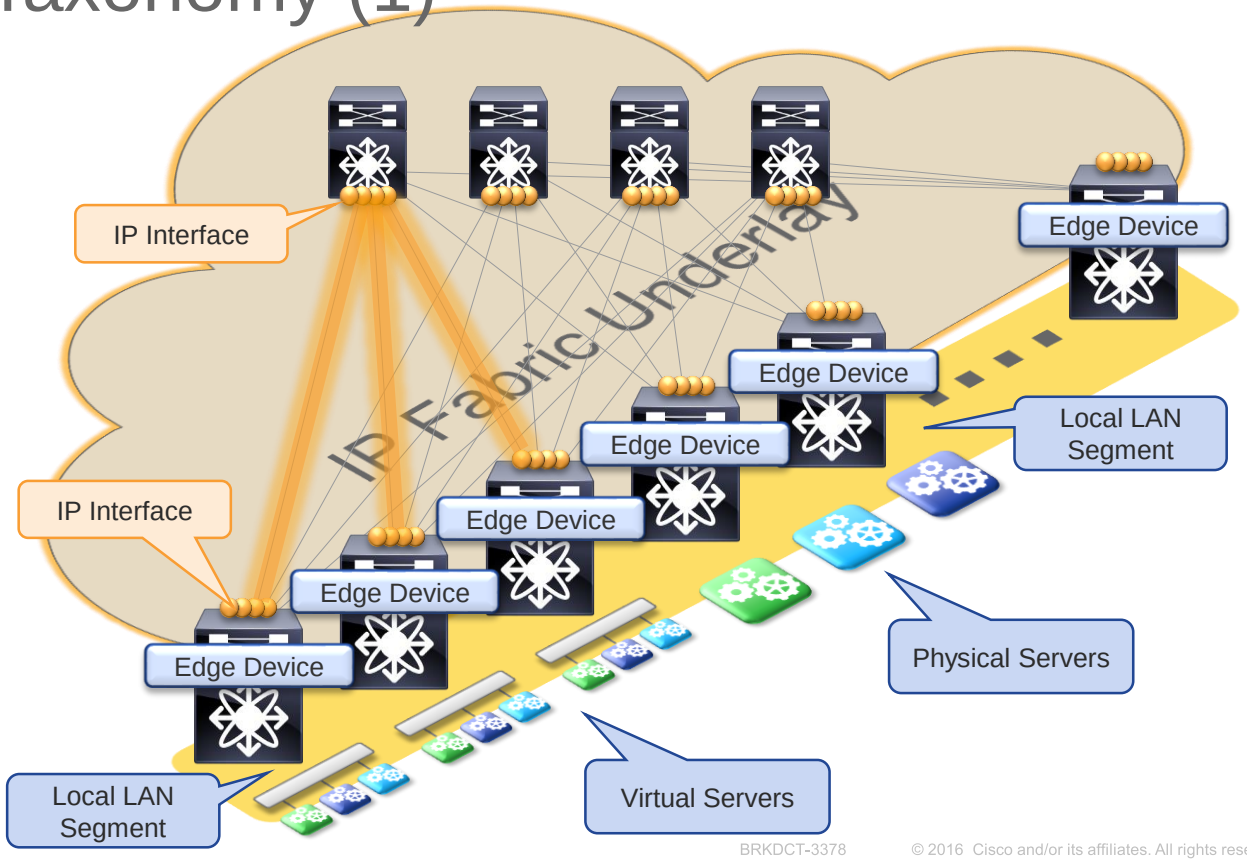
- Overlay Layer 2/Layer 3 Unicast traffic
- Overlay Broadcast, Unknown Unicast, Multicast traffic (BUM traffic) forwarding
  - Ingress Replication
  - Multicast

Why VXLAN?

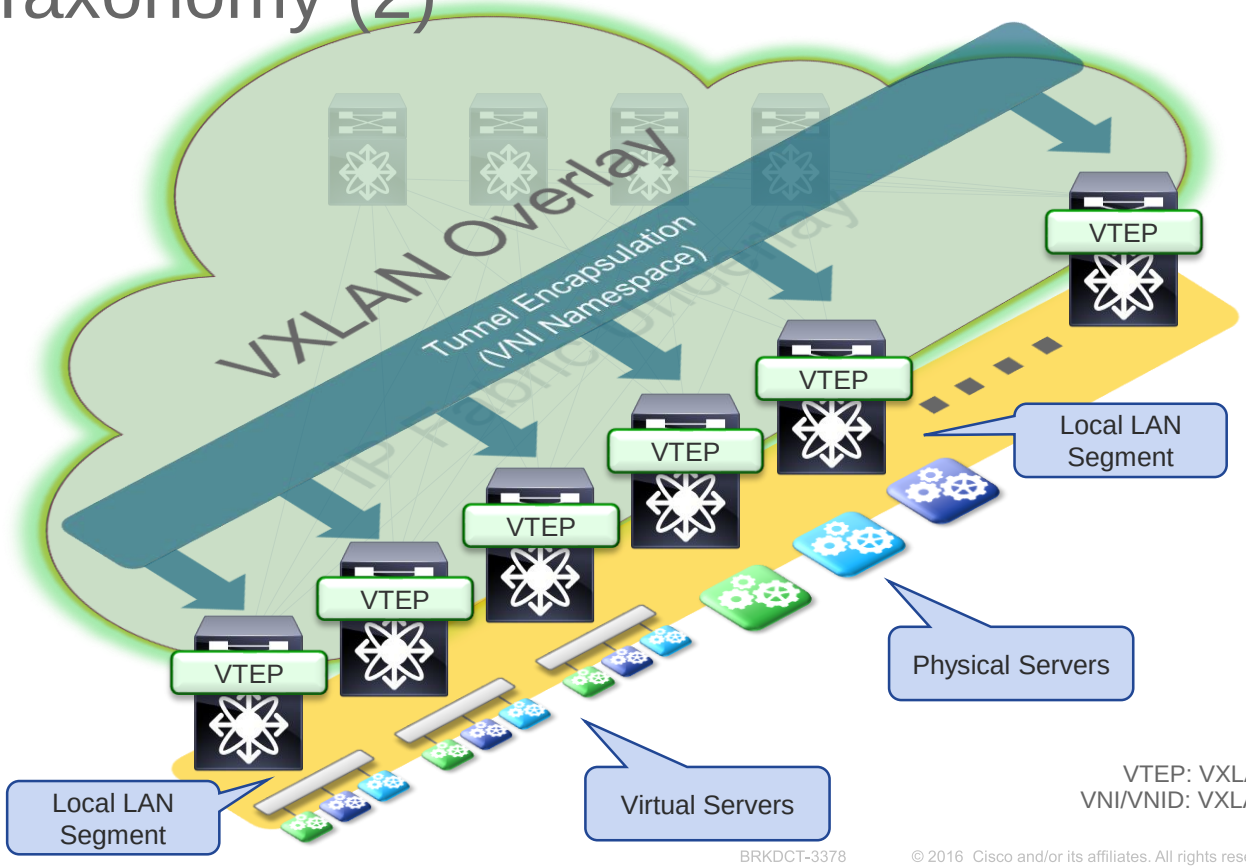
VXLAN provides a Network with Segmentation, IP Mobility, and Scale

- “Standards” based Overlay (RFC 7348)
- 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

Getting the Puzzle Together!

Driving Standards based Overlay-Evolution with

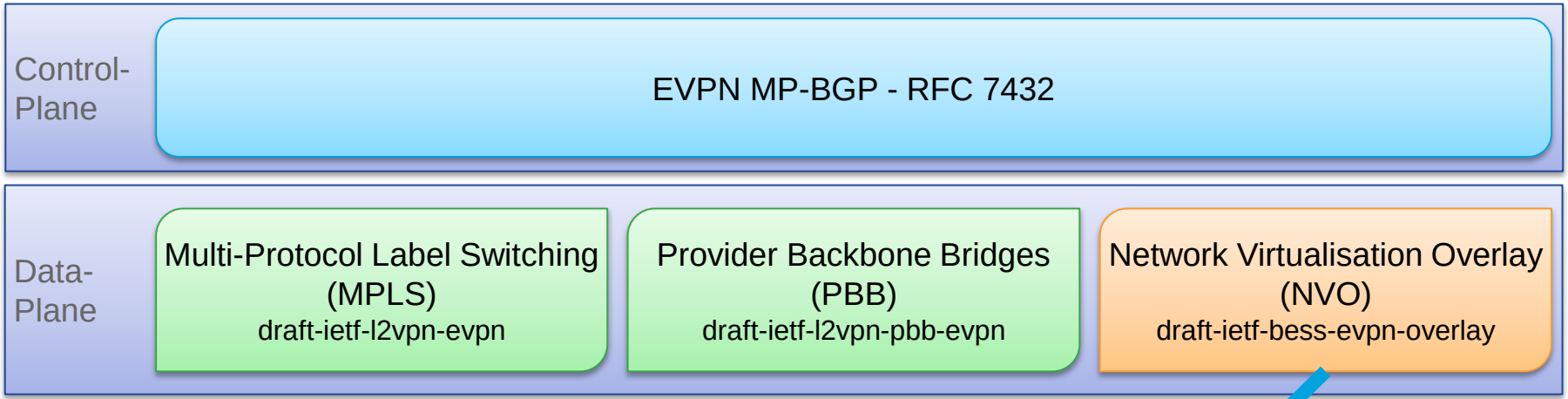
VXLAN BGP EVPN

What is VXLAN with BGP EVPN?

- Standards based Overlay (VXLAN) with Standards based Control-Plane (BGP)
- Layer-2 MAC and Layer-3 IP information distribution by Control-Plane (BGP)
- Forwarding decision based on Control-Plane (minimises flooding)
- Integrated Routing/Bridging (IRB) for Optimised Forwarding in the Overlay
- Multi-Tenancy At Scale

FYI

EVPN – Ethernet VPN



- EVPN over NVO Tunnels (ie VXLAN) for Data Centre Fabric encapsulations
- Provides Layer-2 and Layer-3 Overlays over simple IP Networks

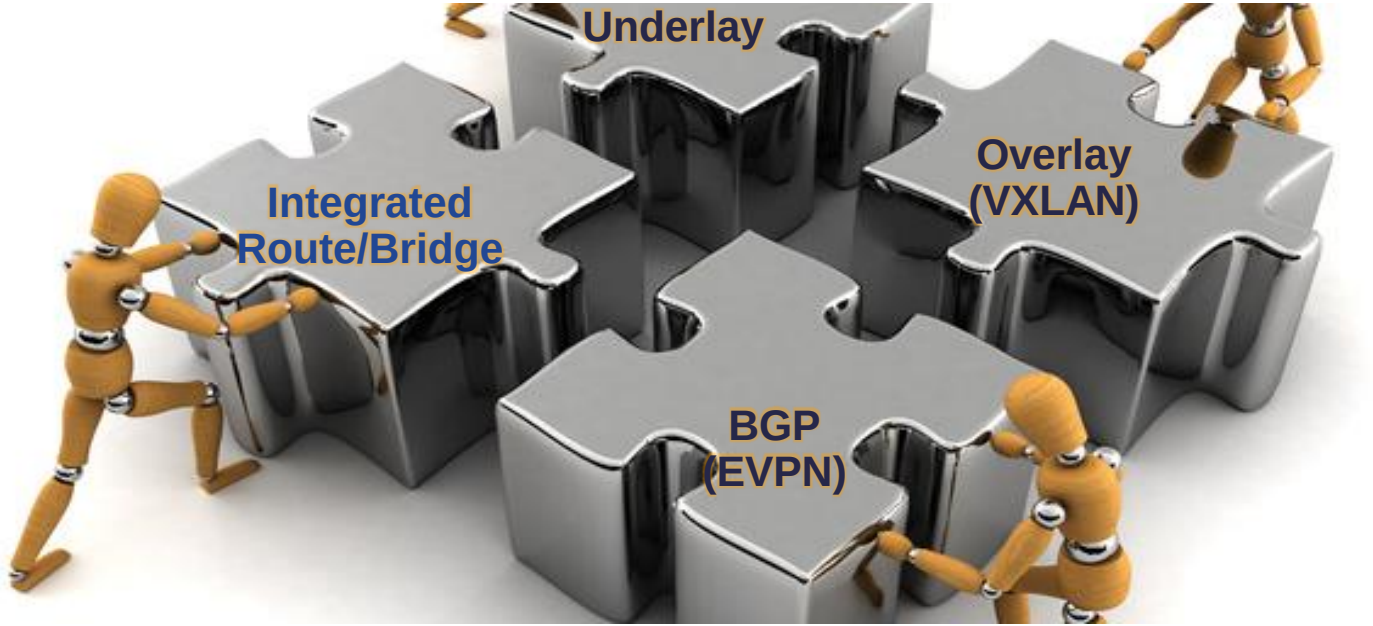
FYI

Cisco's VXLAN related IETF RFCs & Drafts

ID	Title	Category
RFC 7348	Virtual eXtensible Local Area Network	Data Plane
RFC 7432	BGP MPLS based Ethernet VPNs	Control Plane
draft-ietf-bess-evpn-overlay	A Network Virtualisation Overlay Solution using EVPN	Control Plane
draft-ietf-bess-evpn-inter-subnet-forwarding	Integrated Routing and Bridging in EVPN	Control Plane
draft-ietf-bess-l2vpn-evpn-prefix-advertisement	IP Prefix Advertisement in E-VPN	Control Plane
draft-tissa-nvo3-oam-fm	NVO3 Fault Management / OAM	Management Plane

Getting the Puzzle Together!

Optimised Networks with VXLAN

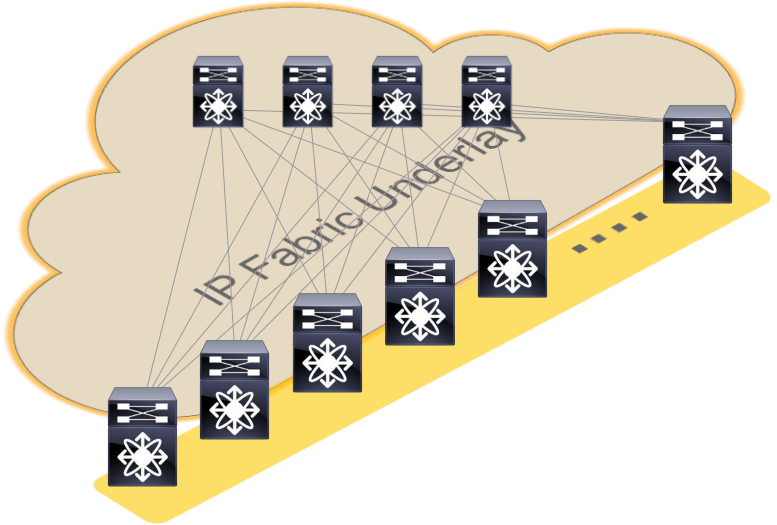


Agenda

- Introduction to Data Centre Fabrics
- VXLAN with BGP EVPN
  - Overview
  - **Underlay**
  - Control & Data Plane
  - Multi-Tenancy
- “Stories” and Use-Cases
- Fabric Management & Automation

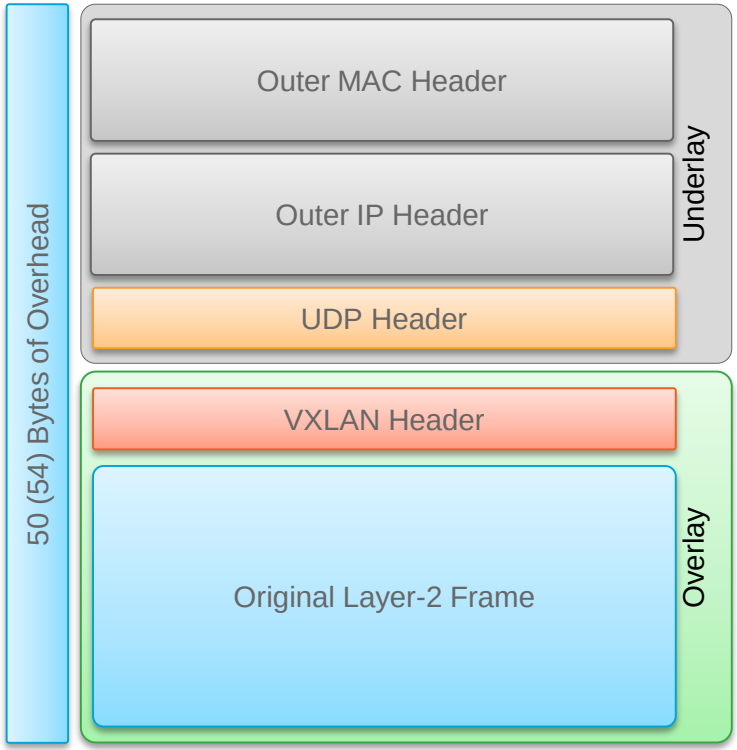
Deployment Considerations

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



*BUM: Broadcast, Unknown Unicast & Multicast

MTU and VXLAN



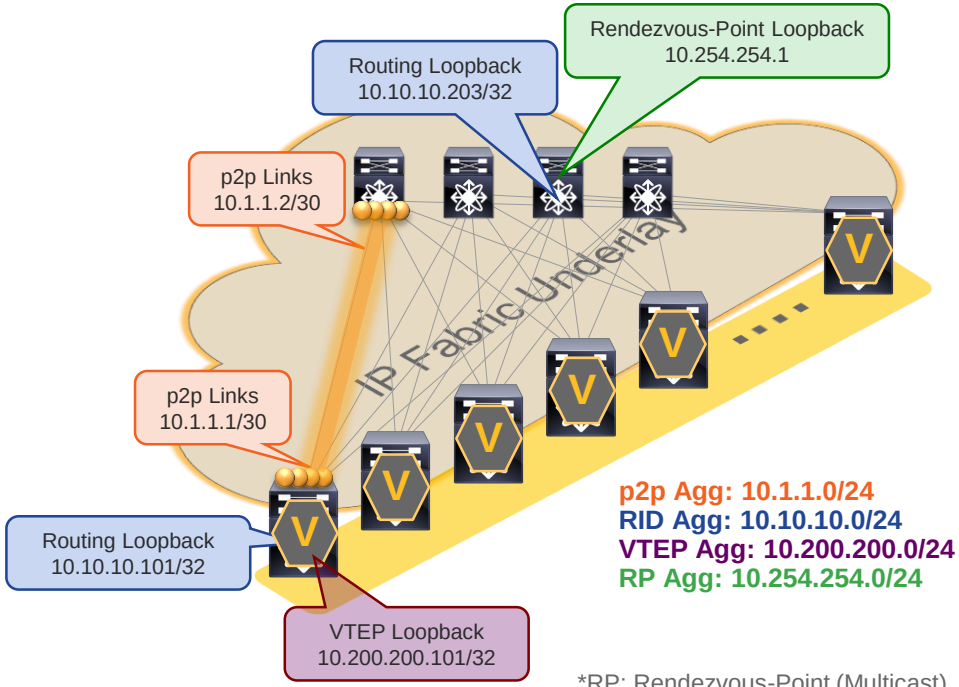
No Fragmentation Needed

- VXLAN adds 50 Bytes (or 54 Bytes) to the Original Ethernet Frame
- Avoid Fragmentation by adjusting the IP Networks MTU
- Data Centres 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 (1)

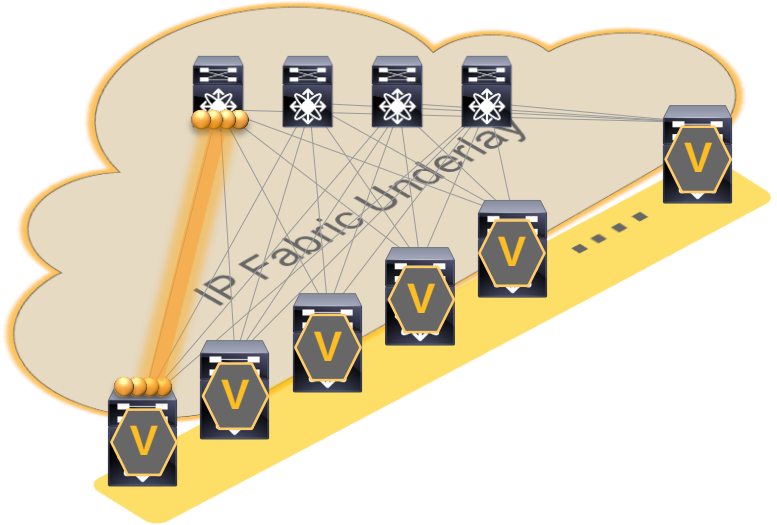
- Know your IP addressing and IP scale requirements
- Separate VTEP from Routing Protocol from RP* Loopback
- Best to use individual Aggregates for the Underlay
  - Unicast Routing p2p** Links
  - Unicast Routing Loopbacks
  - VTEP (NVE) Loopback
  - Multicast Routing Loopback (RP)
- IPv4 only (today)



*RP: Rendezvous-Point (Multicast)
**p2p: Point-to-Point

Building your IP Network – Interface Principles (2)

- Routed Ports/Interfaces
  - Layer-3 Interfaces between Spine and Leaf (no switchport)
  - For each Point-2-Point (P2P) connection, minimum /31 required
  - Alternative, use IP Unnumbered (/32)
- Use Loopback as Source-Interface for VTEP (NVE*)



*NVE: Network Virtualisation Edge
VTEP: VXLAN Tunnel End-Point

Building your IP Network – Some Math

Example from depicted topology:
4 Spine * 6 Leaf = 24 Point-2-Point (P2P) Links
24 Links * 2 (/31) + 10 RID* + 6 VTEP + 4 Spine

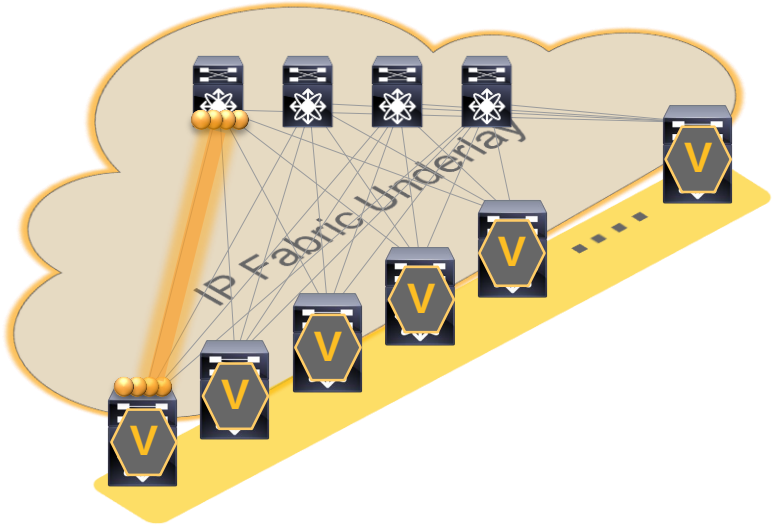
= 48 IP Addresses for P2P Links
= 20 IP Addresses for Loopback Interfaces

68 IP Addresses required == /25 Prefix

A More Realistic Scenario:
4 Spine * 40 Leaf = 160 Point-2-Point (P2P) Link
160 Links * 4 (/30) + 44 RID* + 80 VTEP + 4 Spine

= 640 IP Addresses for P2P Links
= 128 IP Addresses for Loopback Interface

768 IP Addresses required == /22 Prefix



*RID: Router ID; Unicast Routing Loopback

IP Unnumbered– Simplifying the Math

Example from depicted topology:
4 Spine + 6 Leaf = 10 Individual Devices

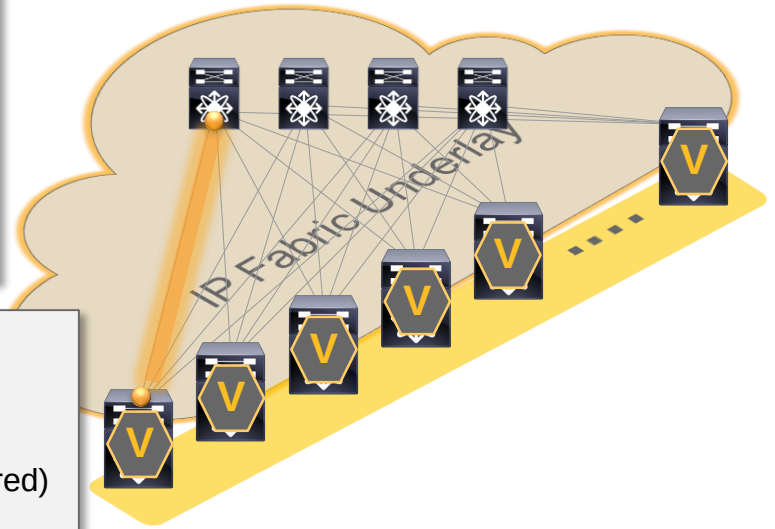
= 6 IP Addresses for Loopback Interface (Used for VTEP)
= 10 IP Address Loopback Interface (RID* & IP Unnumbered)

16 IP Addresses required == /28 Prefix

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

= 40 IP Addresses for Loopback Interface (Used for VTEP)
= 44 IP Addresses for Loopback Interface (RID* & IP Unnumbered)

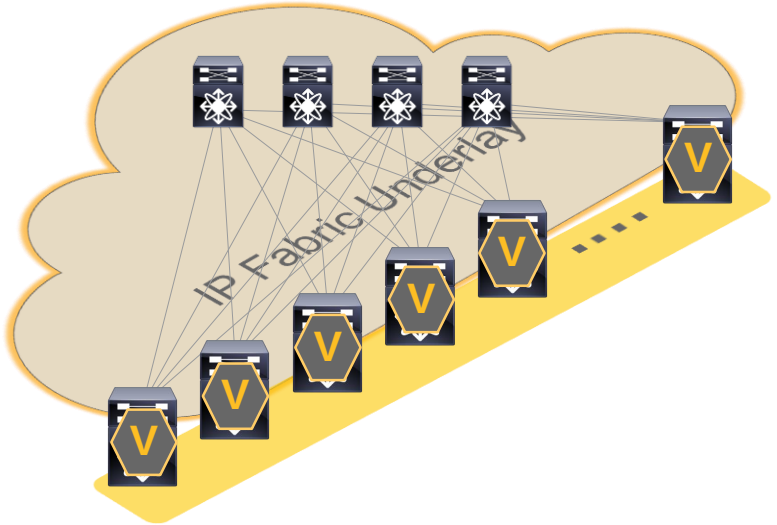
84 IP Addresses required == /25 Prefix



*RID: Router ID; Unicast Routing Loopback

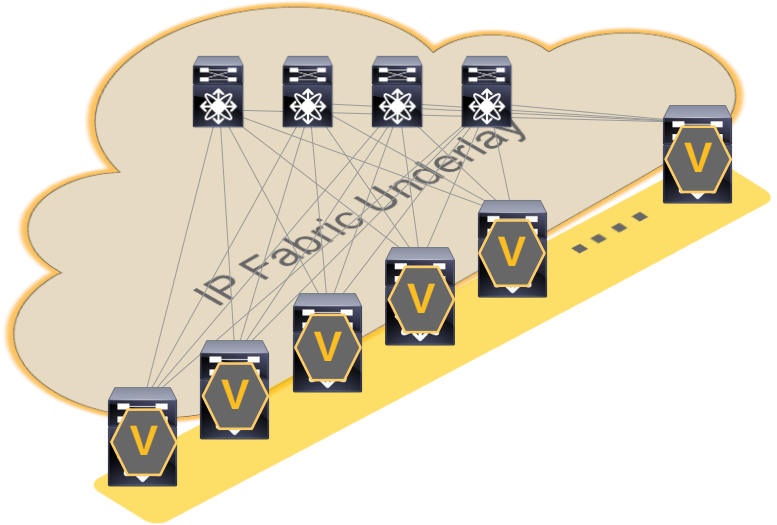
Building your IP Network – Routing Protocols; OSPF

- 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

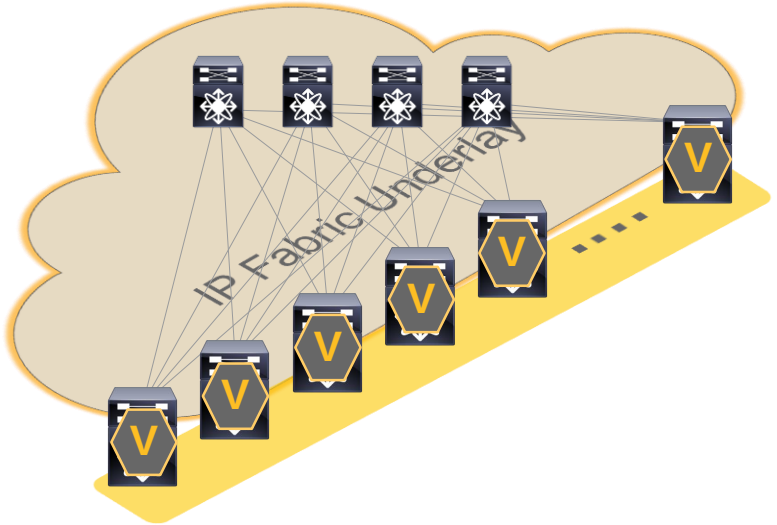
- IS-IS – what was this CLNS?
  - Independent of IP (CLNS)
  - Well suited for routed interfaces/ports
  - No SPF calculation on Link change; only if Topology changes
  - Fast Re-convergence
  - Not everyone is familiar with it



*CLNS: Connection-Less Network Service

Building your IP Network – Routing Protocols; eBGP

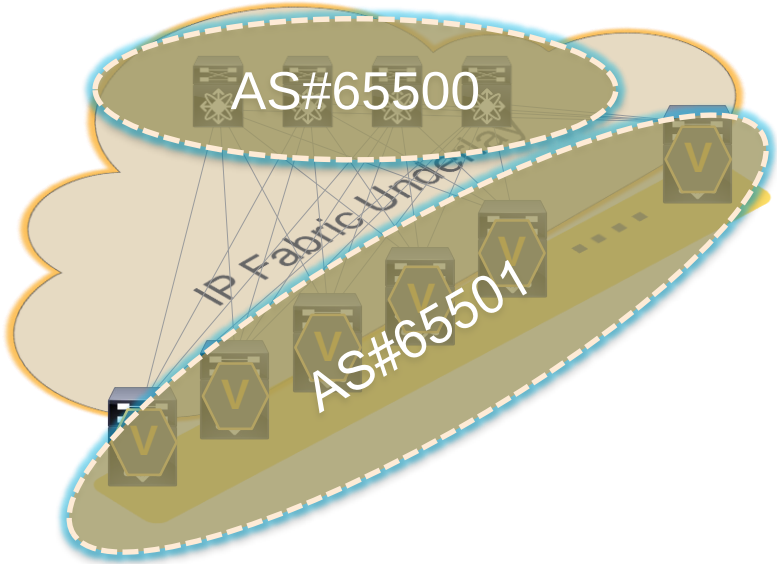
- eBGP – Service Provider style
  - Two Different Models
    - Two-AS
    - Multi-AS
  - BGP is a Distance Vector
    - AS* are used to calculate the Path (AS_Path)
  - If Underlay is eBGP, your Overlay becomes eBGP



*AS: Autonomous System

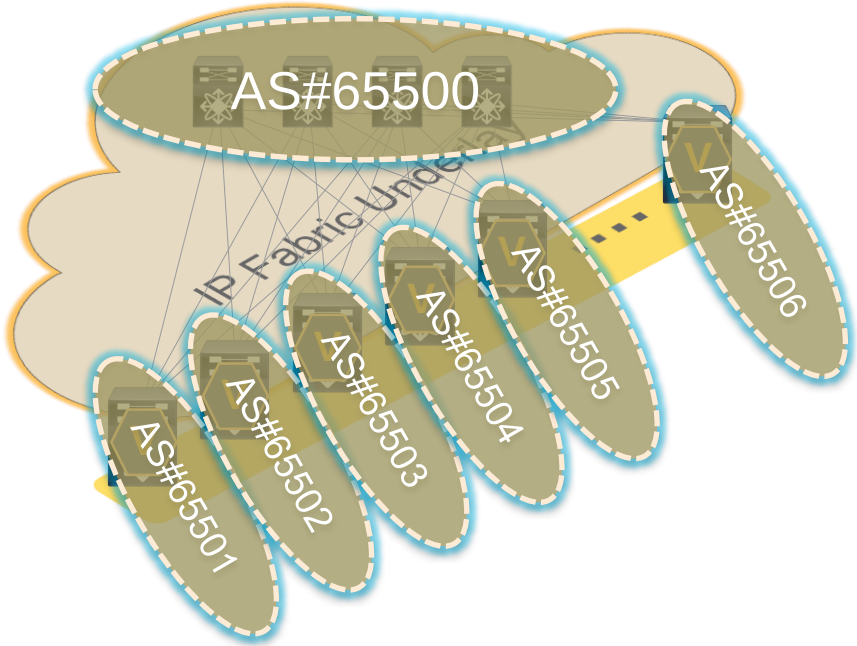
Building your IP Network – Routing Protocols; eBGP

- eBGP – TWO-AS, yes it works!
  - Total of 8 eBGP Peering (with 4 Spine)
    - eBGP peering for Underlay-Routing based on physical interface
      - 4 Spines = 4 BGP Peering per Leaf
      - Advertise all Infrastructure Loopbacks
    - eBGP peering for Overlay-Routing (EVPN)
      - Loopback to Loopback Peering
      - 4 Spines = 4 BGP Peering
  - Requires some BGP config knobs
    - Disable BGP AS-Path check
    - Next-Hop needs to be Unchanged
    - Retain all Routes on Spine (not a RR)



Building your IP Network – Routing Protocols; eBGP

- eBGP – Multi-AS
  - Total of 8 eBGP Peering (with 4 Spine)
    - eBGP peering for Underlay-Routing based on physical interface
      - 4 Spines = 4 BGP Peering per Leaf
      - Advertise all Infrastructure Loopbacks
    - eBGP peering for Overlay-Routing (EVPN)
      - Loopback to Loopback Peering
      - 4 Spines = 4 BGP Peering
  - Requires some BGP config knobs
    - Next-Hop needs to be Unchanged
    - Retain all Routes on Spine (not a RR)



FYI

Multicast Enabled Underlay

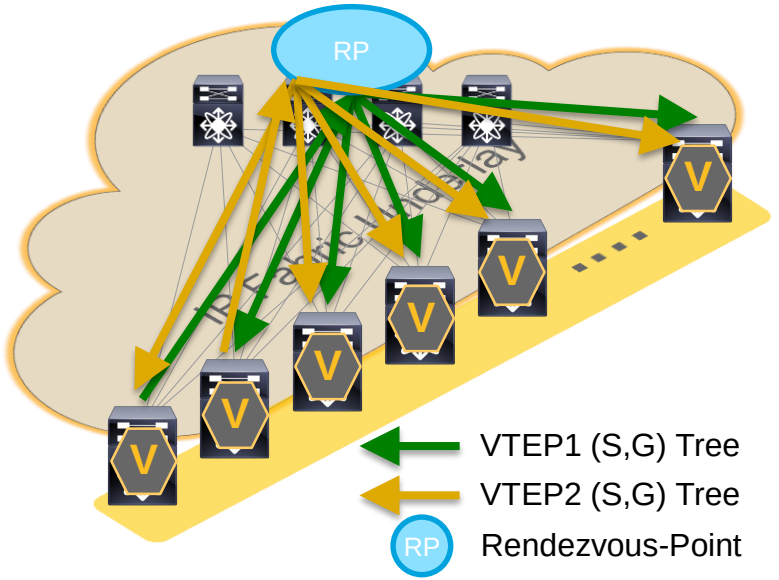
May use PIM-ASM or PIM-BiDir (Different hardware has different capabilities)

	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 optimise 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*

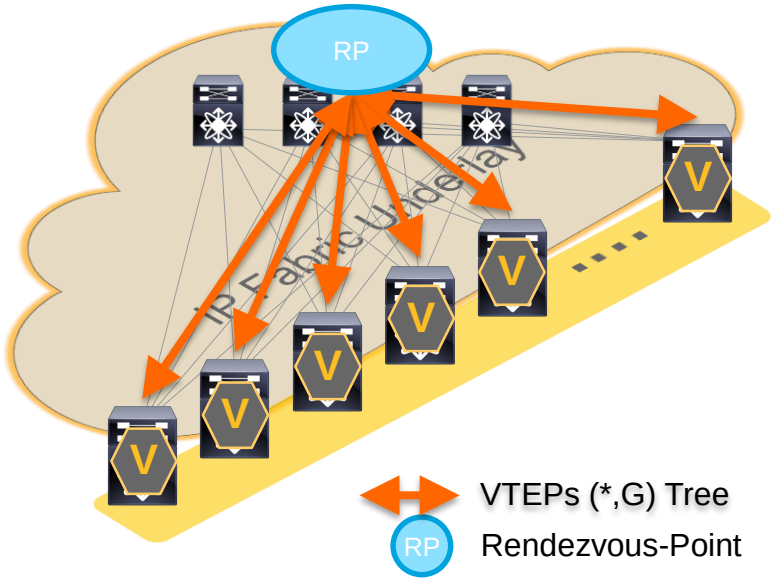
- PIM Sparse-Mode (ASM)
- Redundant Rendezvous-Point using PIM Anycast-RP or MSDP
- Source-Tree or Unidirectional Shared-Tree (Source-Tree shown)
  - Shared-Tree will always use RP for forwarding
- 1 Source-Tree per Multicast-Group per VTEP (each VTEP is Source & Receiver)



*ASM: Any-Source Multicast

Multicast Enabled Underlay – BiDir-PIM*

- 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



*BiDir-PIM: Bidirectional PIM

FYI

To Remember - Multicast Enabled Underlay

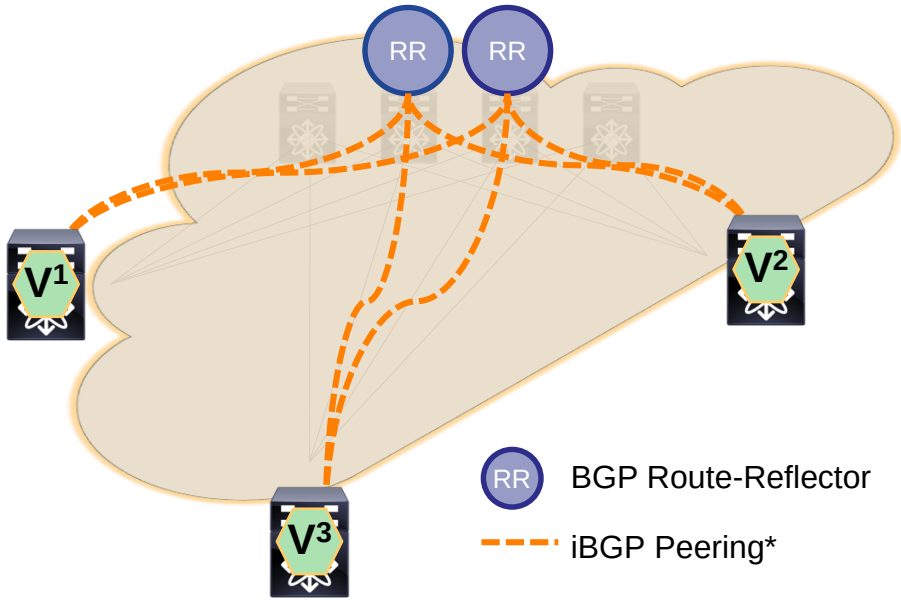
- Multi-Destination Traffic (Broadcast, Unknown Unicast, etc.) needs to be replicated to ALL VTEPs serving a given VNI
  - Each VTEP is Multicast Source & Receiver
- For a given VNI, all VTEPs act as a Sender and a Receiver
- Head-End Replication will depend on hardware scale/capability
- Resilient, 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)

Agenda

- Introduction to Data Centre Fabrics
- VXLAN with BGP EVPN
  - Overview
  - Underlay
  - **Control & Data Plane**
  - Multi-Tenancy
- “Stories” and Use-Cases
- Fabric Management & Automation

Multiprotocol BGP (MP-BGP) Primer

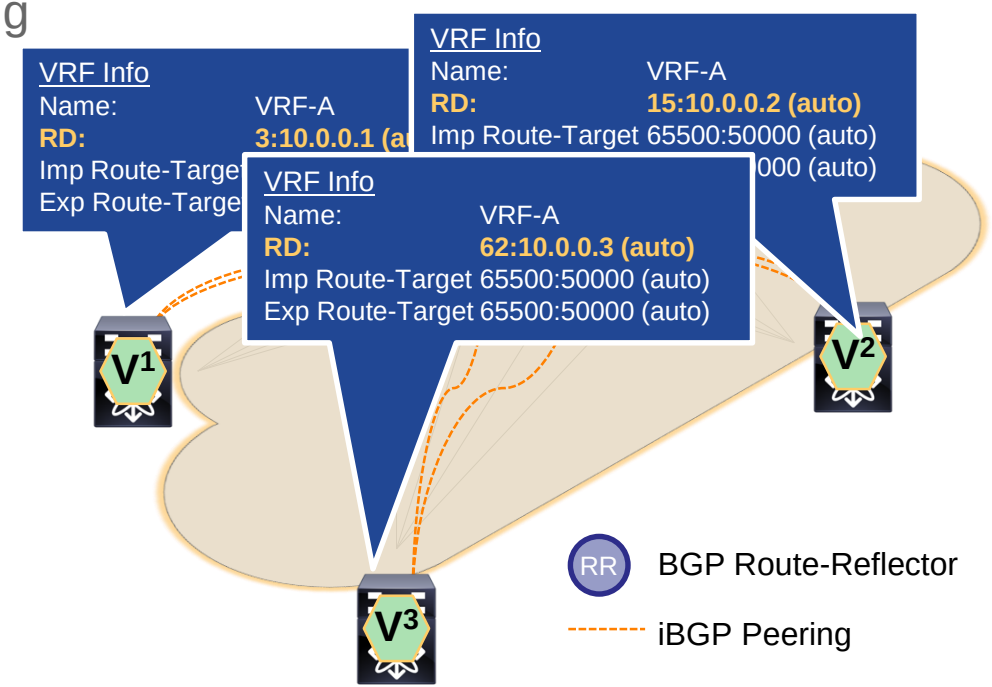
- Multiprotocol BGP (MP-BGP)
- Extension to Border Gateway Protocol (BGP) - RFC 4760
- VPN Address-Family:
  - Allows different types of address families (e.g. VPNv4, VPNv6, L2VPN EVPN, MVPN)
  - Information transported across single BGP peering



*eBGP supported without BGP Route-Reflector

Multiprotocol BGP (MP-BGP) Primer

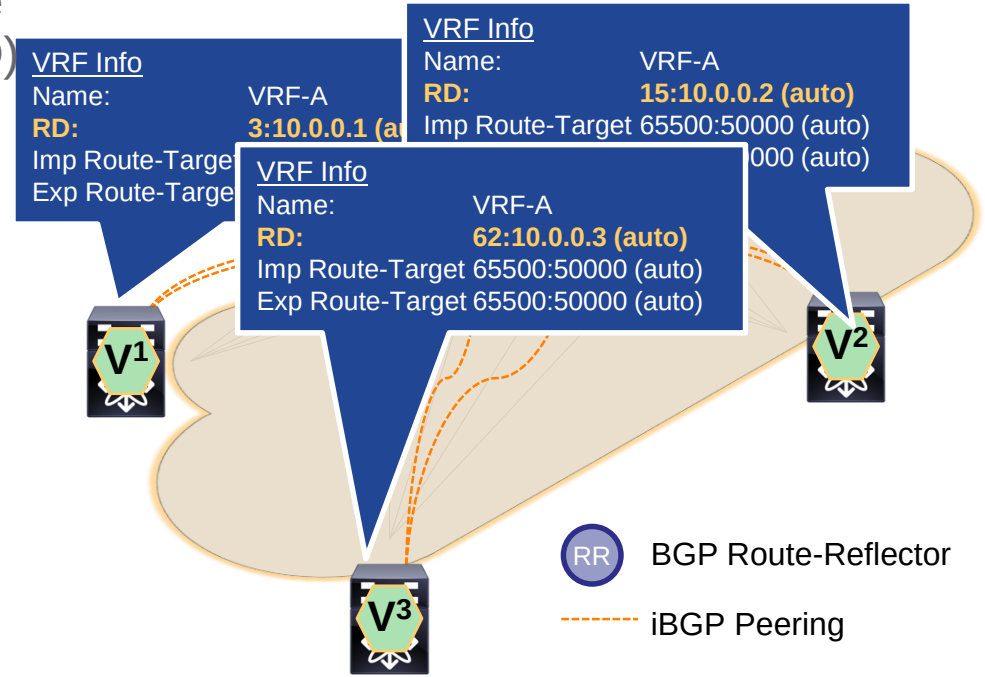
- VPN segmentation for tenant routing (Multi-Tenancy)
  - Route Distinguisher (RD)
  - 8-byte field of VRF parameters
  - value to make VPN prefix unique:
    - RD + VPN prefix



Multiprotocol BGP (MP-BGP) Primer

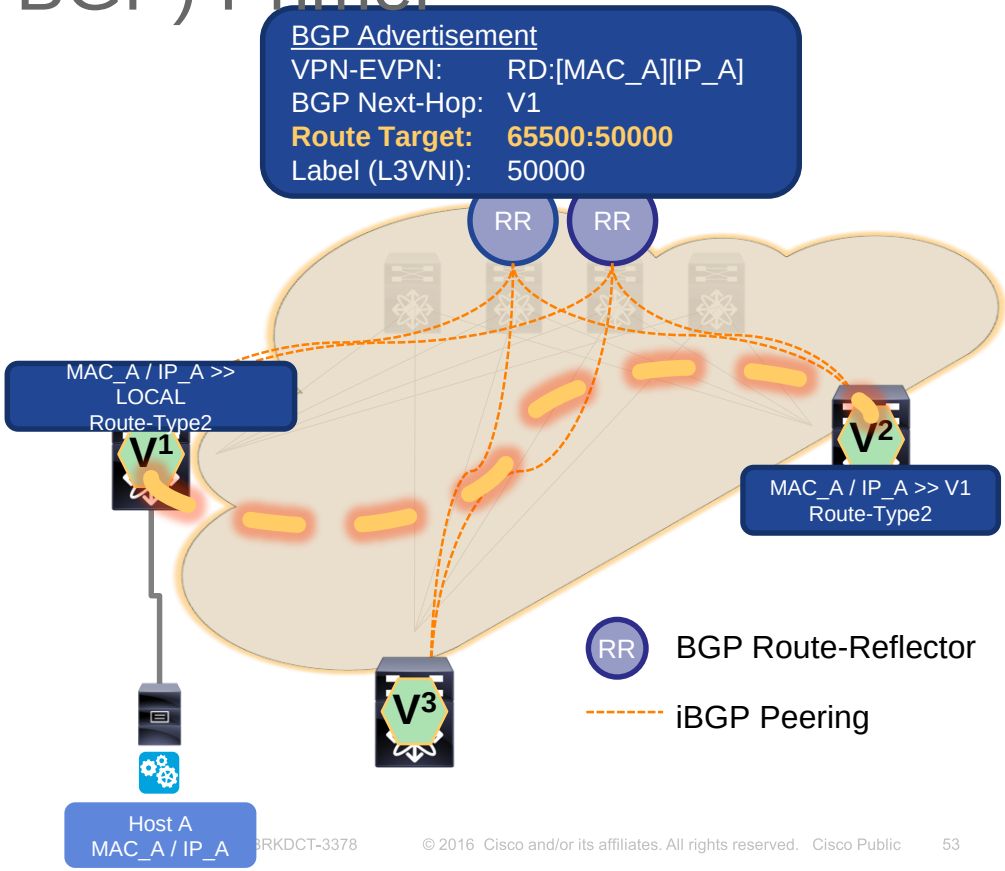
- Cisco's VXLAN/EVPN does provide automated Route Distinguisher (RD)
  - Automatic uses Type 1 format
  - 4-byte IP Address (Router ID)
  - 4-byte Value (VRF ID)

```
vrf context VRF-A
  vni 50000
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
  address-family ipv6 unicast
    route-target both auto
    route-target both auto evpn
```



Multiprotocol BGP (MP-BGP) Primer

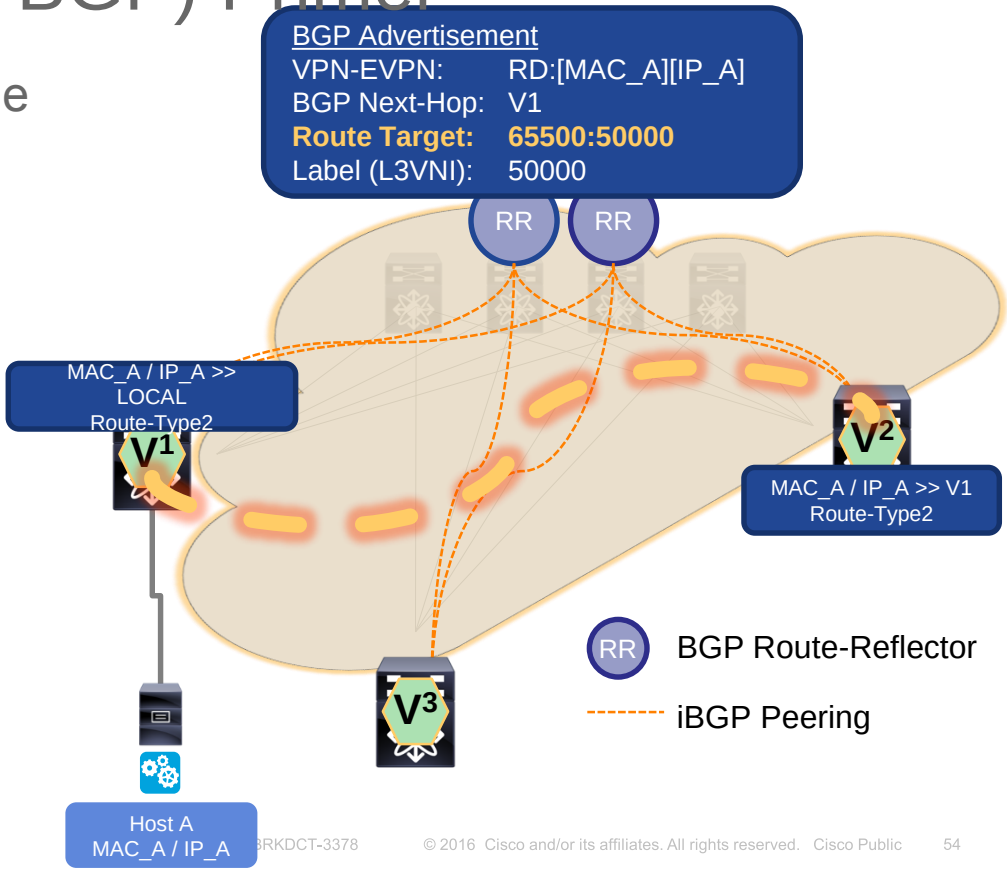
- VPN Segmentation for tenant routing (Multi-Tenancy)
- Selective distribute VPN routes - Route Target (RT)
  - 8-byte field of VRF parameter
  - unique value to define the import/export rules for VPN prefix



Multiprotocol BGP (MP-BGP) Primer

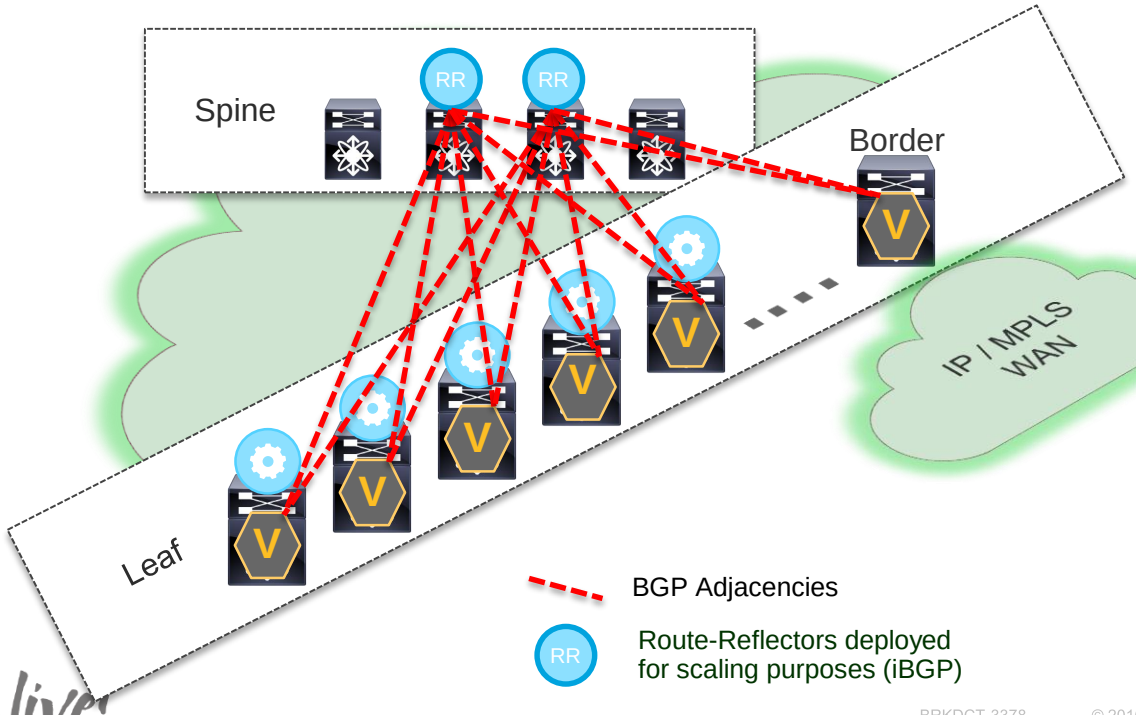
- Cisco's VXLAN/EVPN does provide automated Route Target (RT)
  - 8-byte Route Target (2 x 4-byte)
  - ASN : VNI

```
vrf context VRF-A
  vni 50000
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
  address-family ipv6 unicast
    route-target both auto
    route-target both auto evpn
```



Overlay with Optimised Routing

EVPN Control Plane -- Host and Subnet Route Distribution



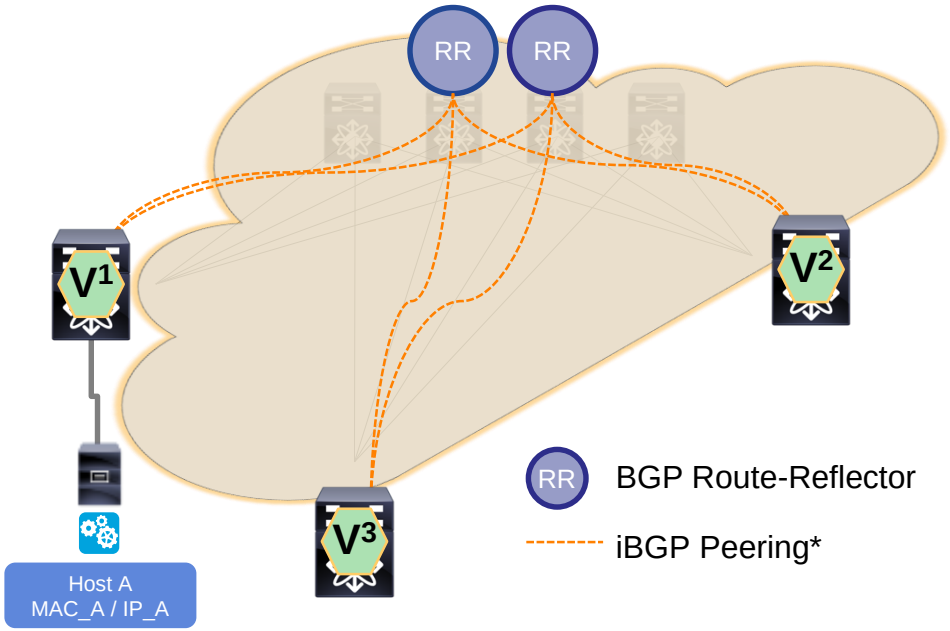
BGP Update

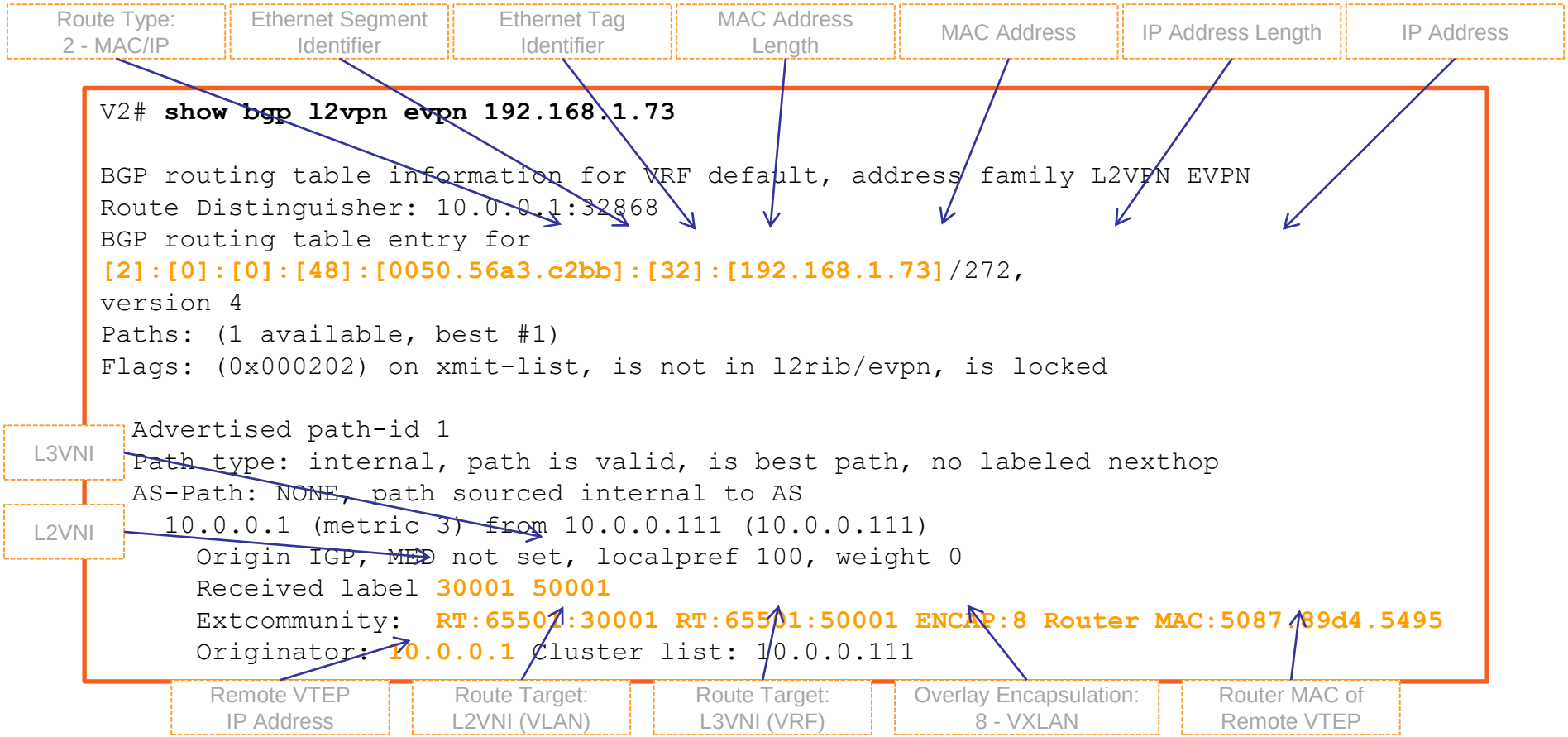
- Host-MAC
- Host-IP
- Internal IP Subnet
- External Prefixes

Host Advertisement

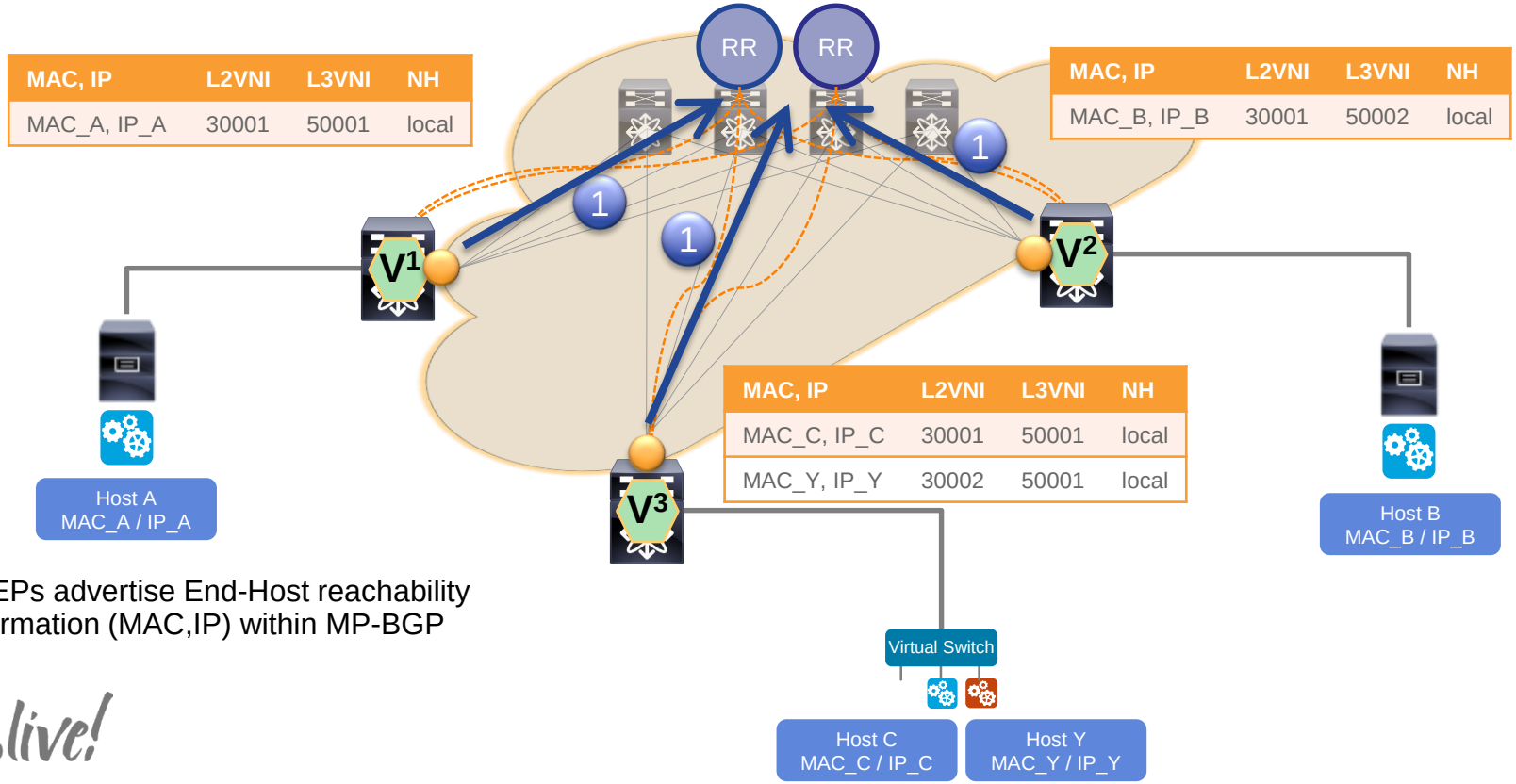
- Host Attaches
  - Host “A” attaches to Edge Device (VTEP)
- VTEP V1 advertises Host “A” reachability information
  - MAC and L2VNI [mandatory]
  - IP and L3VNI [optional]
  - depending on ARP
- Additional route attributes advertised
  - MPLS Label1 (L2VNI)
  - MPLS Label2 (L3VNI)
  - Extended Communities

Route Type	MAC, IP	L2VNI (“VLAN”)	L3VNI (“VRF”)	NH	Encap	Seq
2	MAC_A, IP_A	30001	50001	IP_V1	8:VXLAN	0



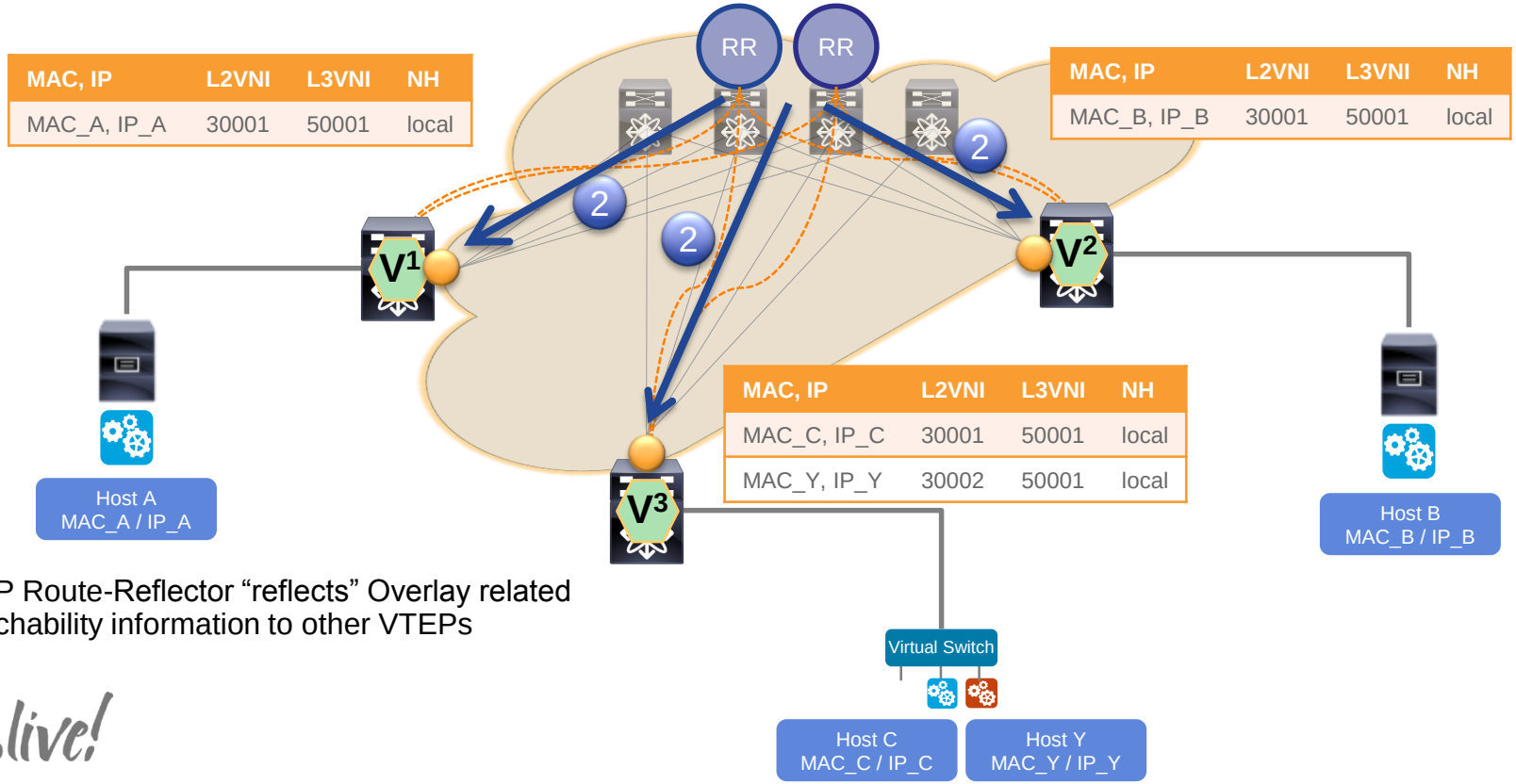


Protocol Learning & Distribution



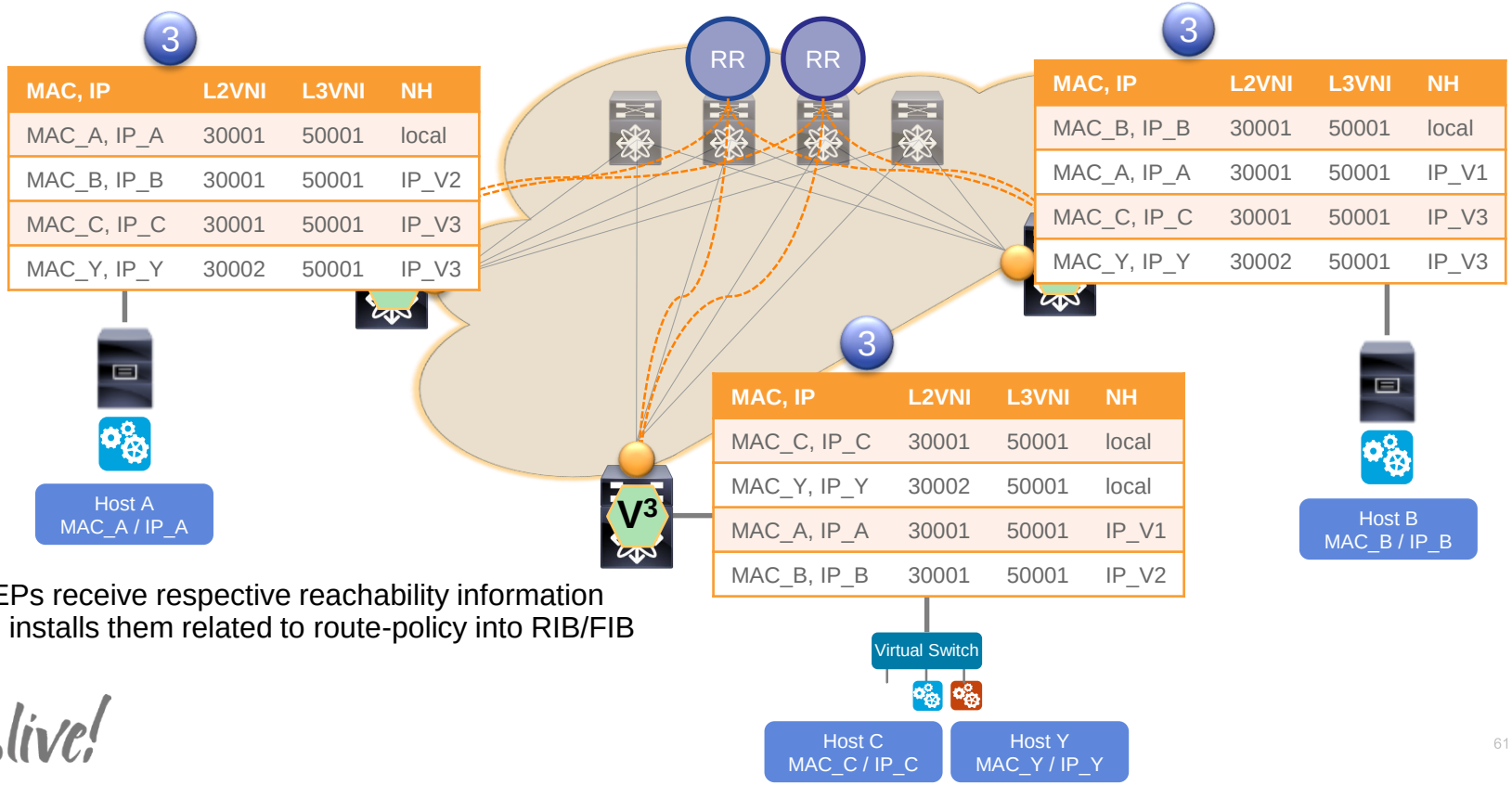
1 VTEPs advertise End-Host reachability information (MAC,IP) within MP-BGP

Protocol Learning & Distribution



2 BGP Route-Reflector “reflects” Overlay related reachability information to other VTEPs

Protocol Learning & Distribution

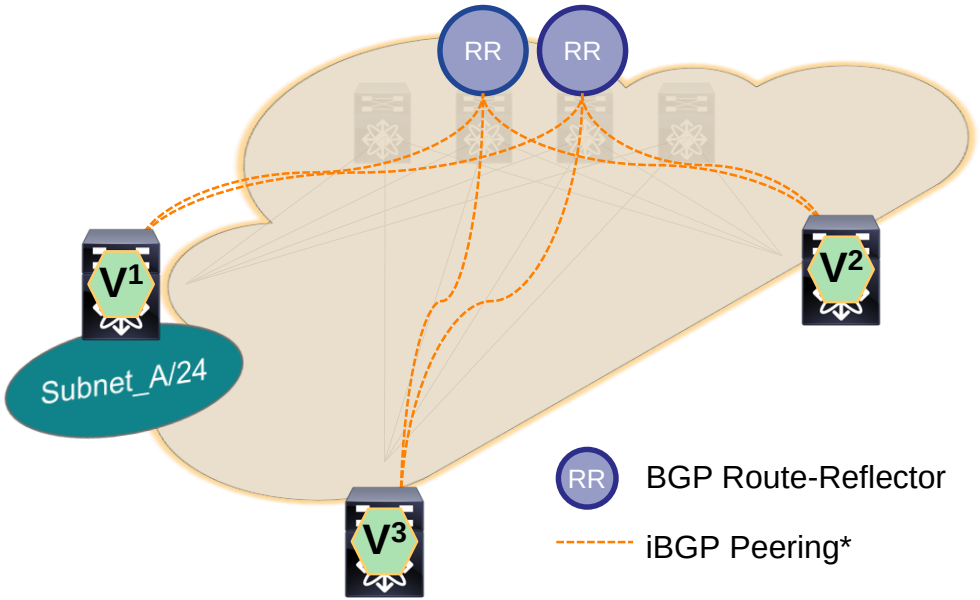


3 VTEPs receive respective reachability information and installs them related to route-policy into RIB/FIB

Subnet Route Advertisement

- IP Prefix Redistribution
  - From “Direct” (connected), “Static” or dynamically learned Routes
- VTEP V1 advertises local Subnet through redistribution of “Direct” (connected) routes
  - IP Prefix, IP Prefix Length, and L3VNI
- Additional route attributes advertised
  - MPLS Label (L3VNI)
  - Extended Communities

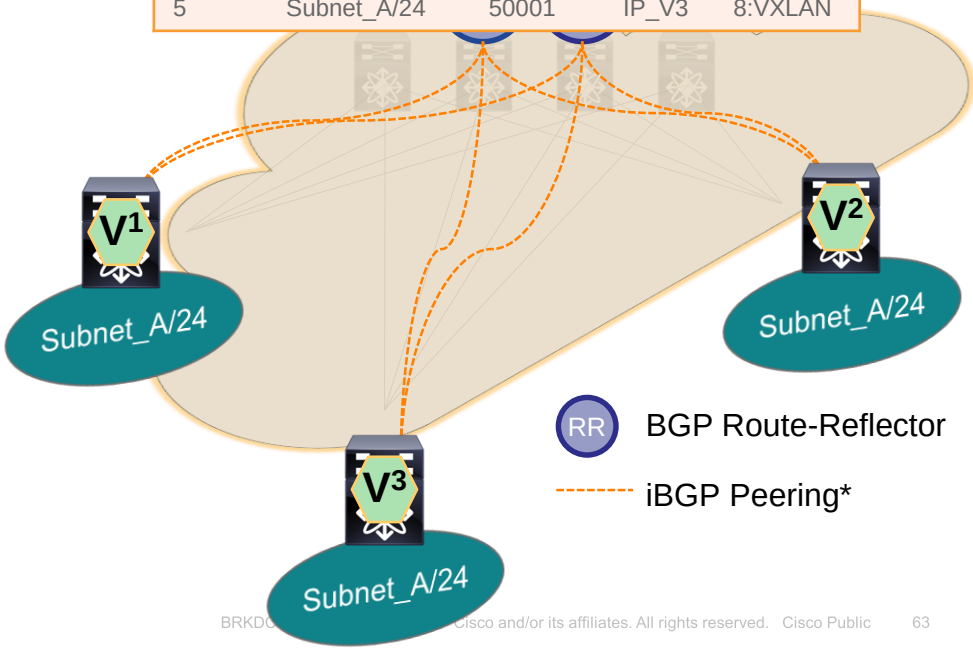
Route Type	MAC, IP	L3VNI (“VRF”)	NH	Encap
5	Subnet_A/24	50001	IP_V1	8:VXLAN



Subnet Route Advertisement

- If multiple VTEP announce same IP Prefix, Equal Cost Multipath (ECMP) will apply
- VTEP V1 advertises local Subnet through redistribution of “Direct” (connected) routes
  - IP Prefix, IP Prefix Length, and L3VNI
- Additional route attributes advertised
  - MPLS Label (L3VNI)
  - Extended Communities

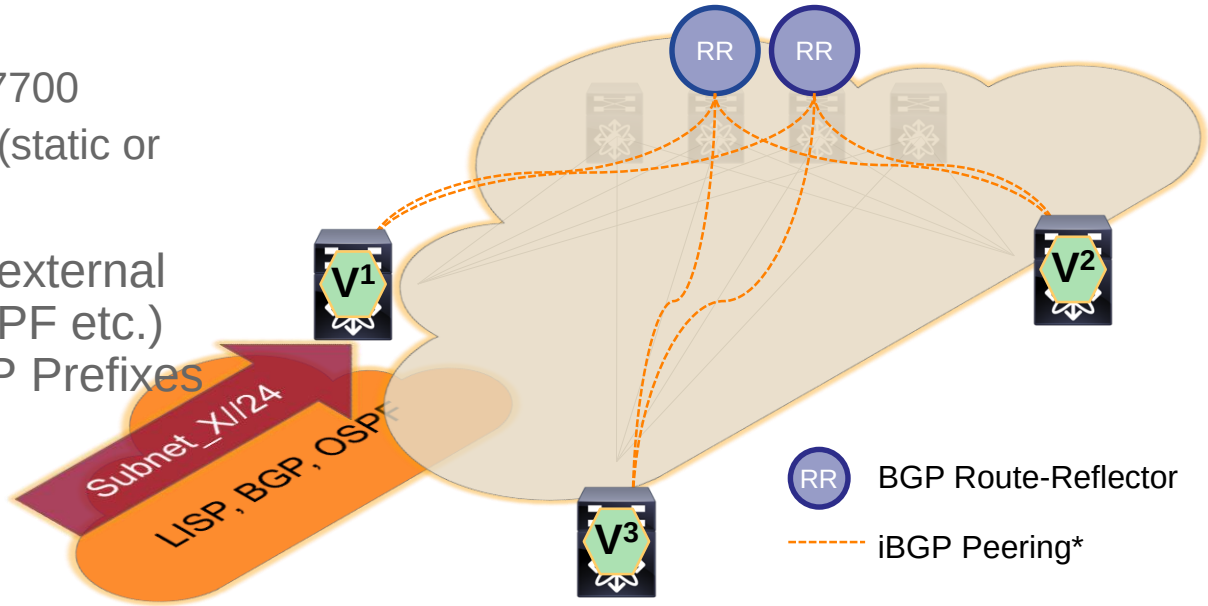
Route Type	MAC, IP	L3VNI (“VRF”)	NH	Encap
5	Subnet_A/24	50001	IP_V1	8:VXLAN
5	Subnet_A/24	50001	IP_V2	8:VXLAN
5	Subnet_A/24	50001	IP_V3	8:VXLAN

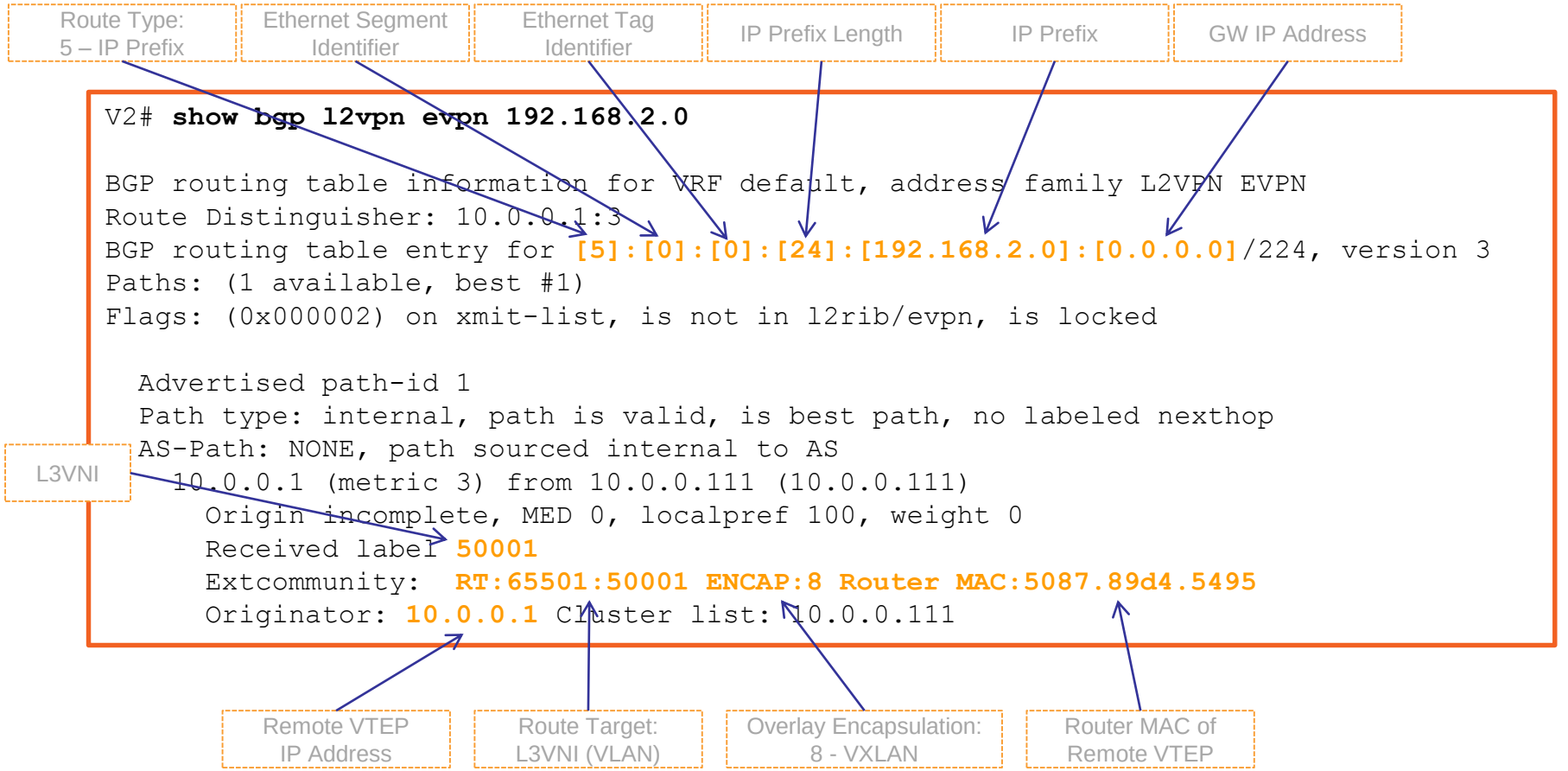


Subnet Route Advertisement

- IP Prefix Learning
  - via BGP with VRF-Lite (Inter-AS Option A)
  - via LISP on Nexus 7000/7700
  - via other routing protocol (static or dynamic)
- VTEP V1 participated in external Peering (LISP, BGP, OSPF etc.) and advertises learned IP Prefixes into the Fabric
  - IP Prefix
  - IP Prefix Length
  - L3VNI

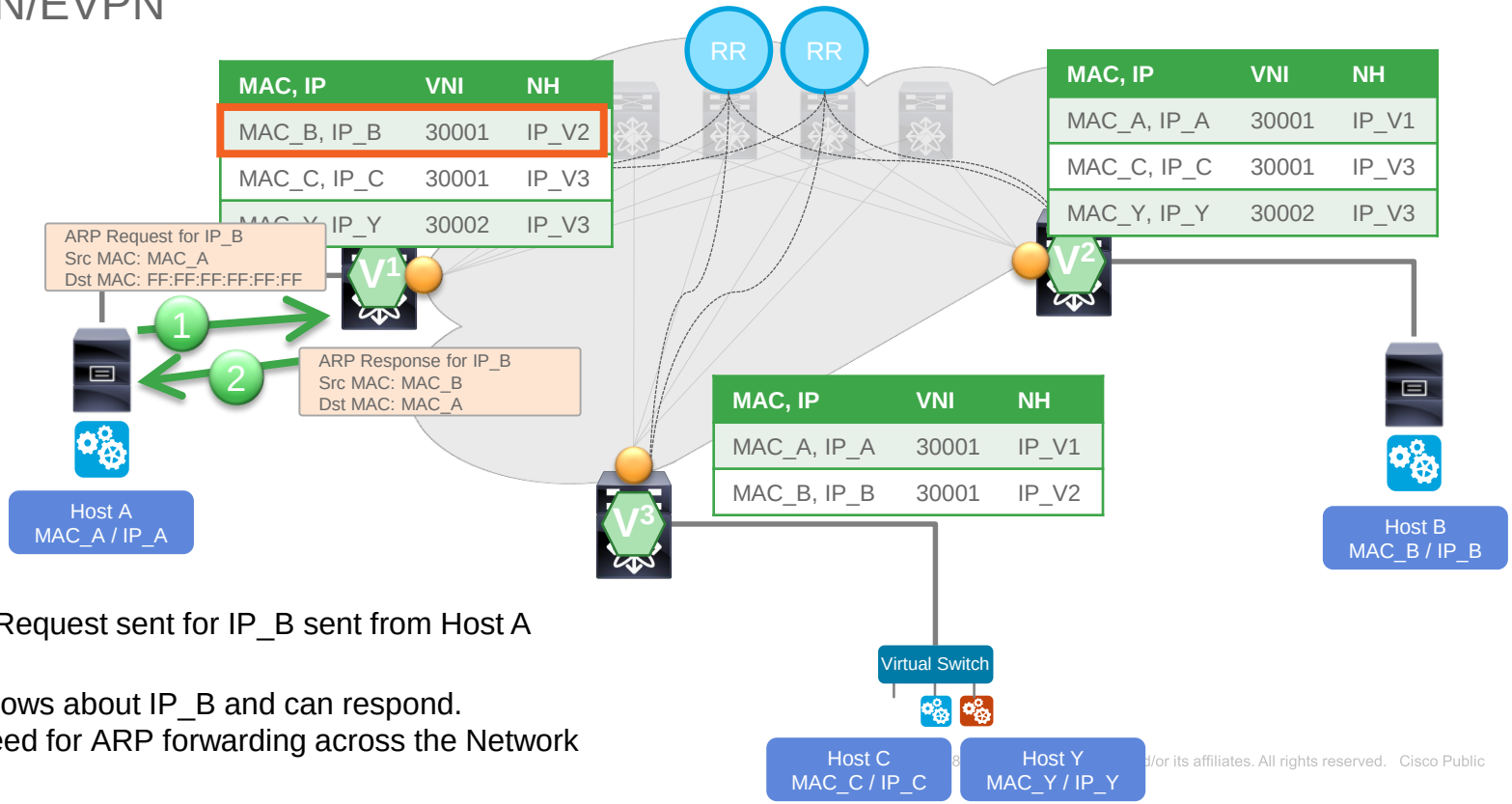
Route Type	MAC, IP	L3VNI ("VRF")	NH	Encap
5	Subnet_X/24	50001	IP_V1	8:VXLAN





ARP Suppression

VXLAN/EVPN

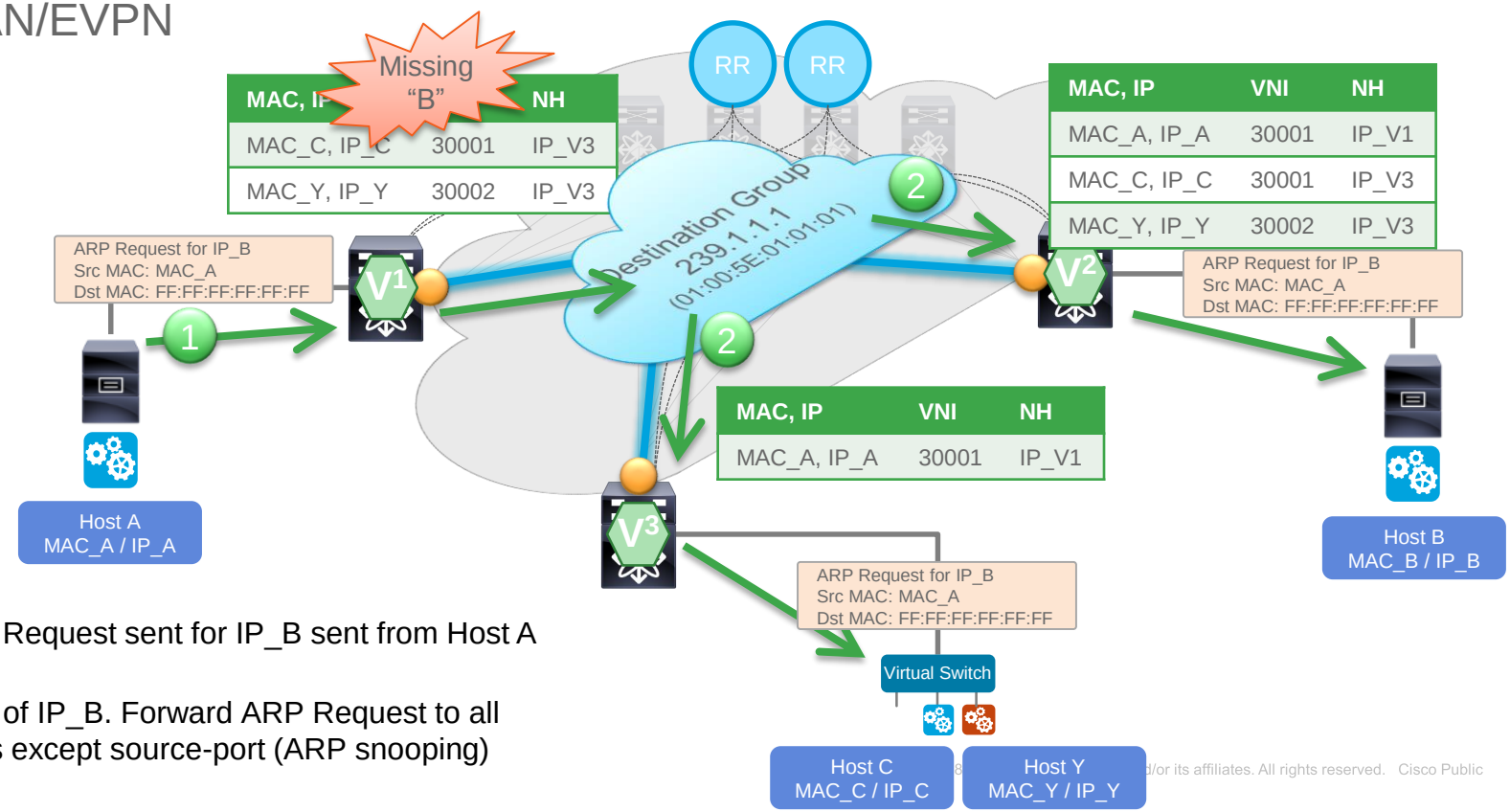


1. ARP Request sent for IP_B sent from Host A

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

ARP Handling on Lookup “Miss” (1)

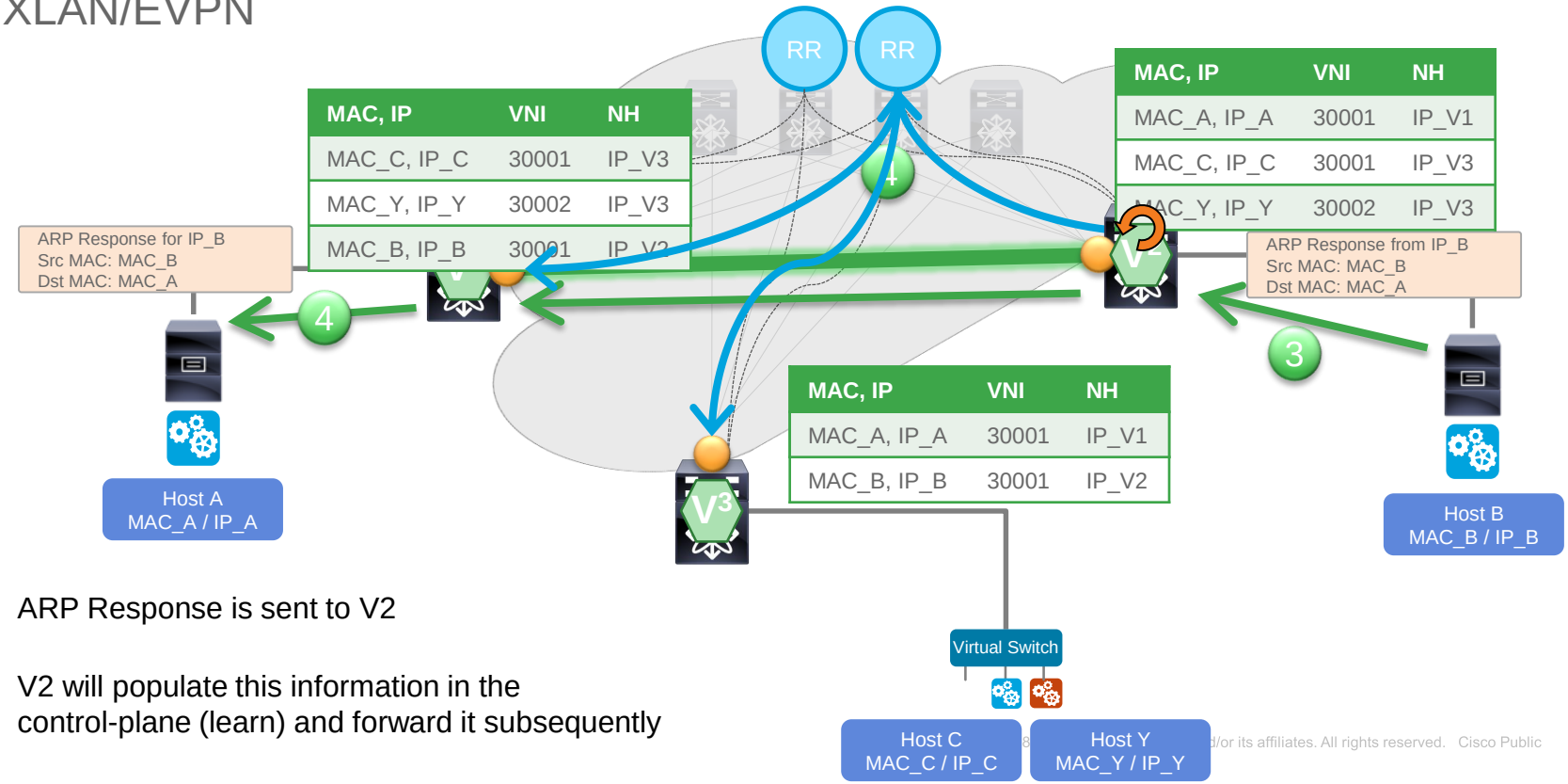
VXLAN/EVPN



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

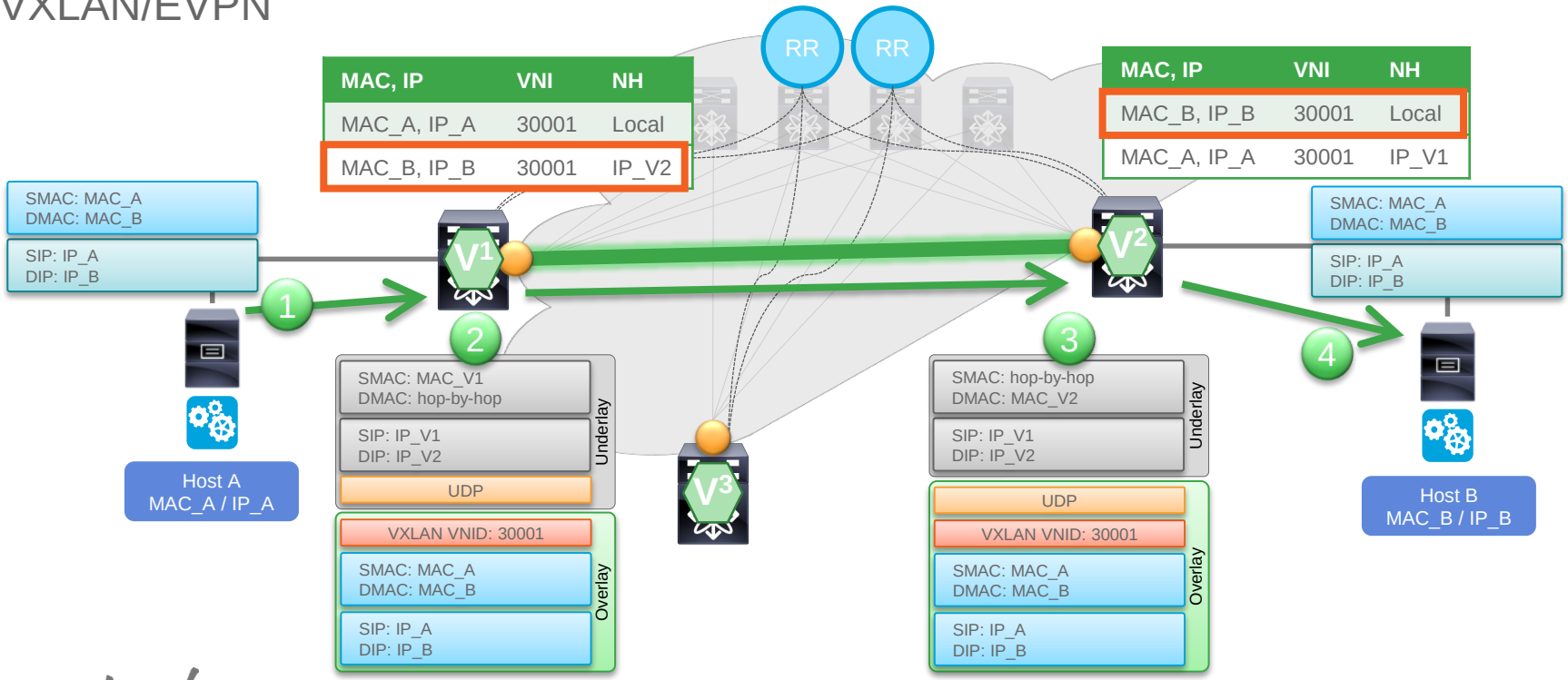


3 ARP Response is sent to V2

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

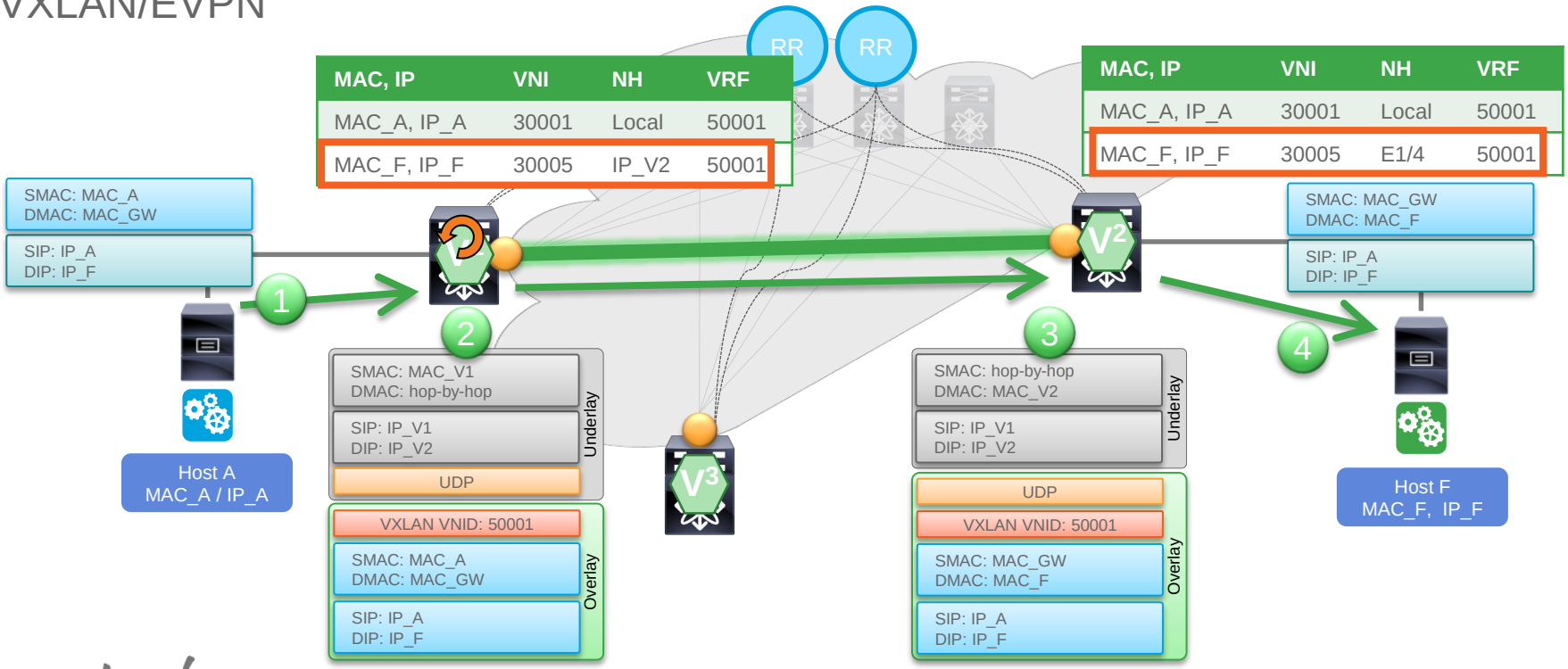
Packet Forwarding (Bridge)

VXLAN/EVPN



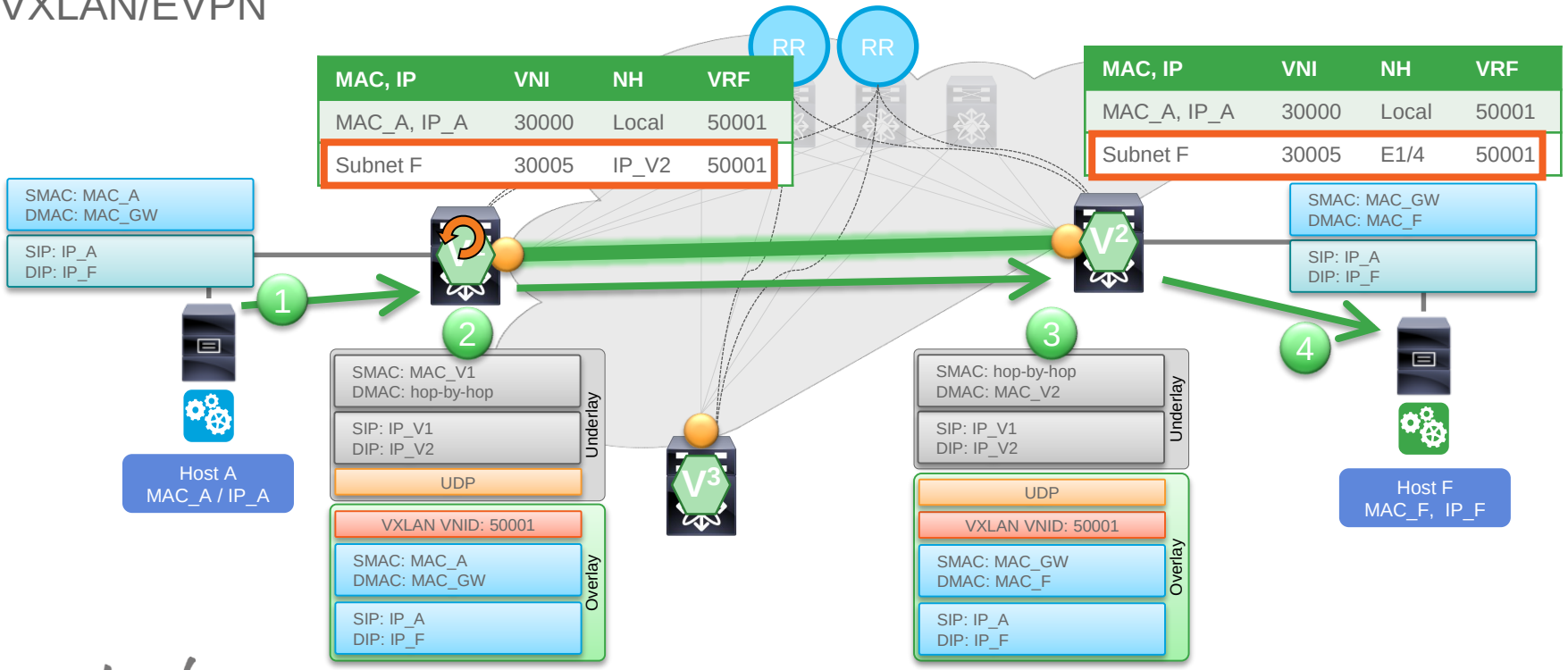
Packet Forwarding (Route)

VXLAN/EVPN



Packet Forwarding (Route) – Silent Host

VXLAN/EVPN

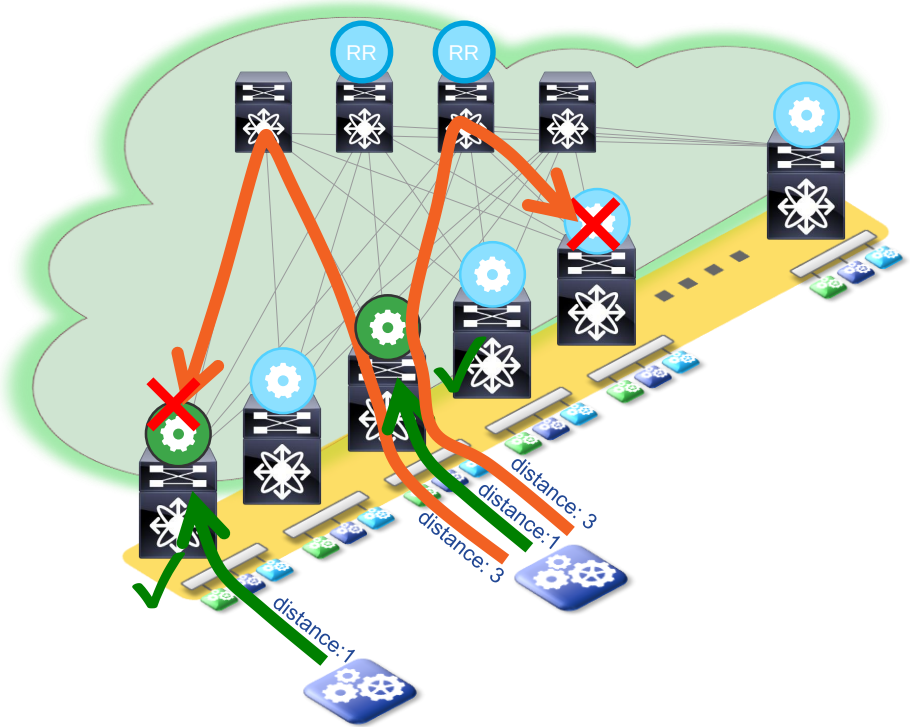


Data Centre Fabric Properties

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

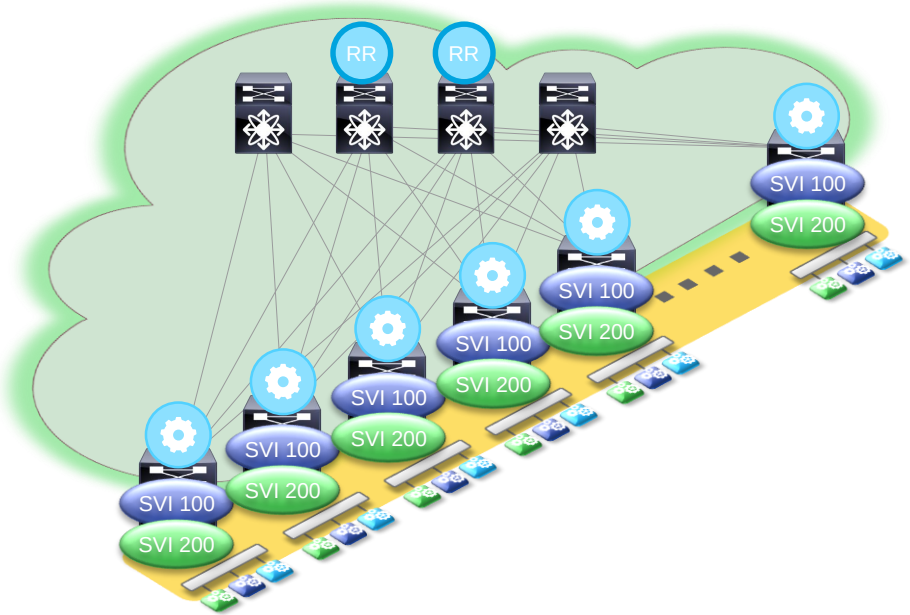
Anycast – One-to-Nearest Association

- a **network** addressing and routing **methodology**
- datagrams sent from a single sender to the topologically **nearest node**
- group of potential receivers, all identified by the **same destination address**



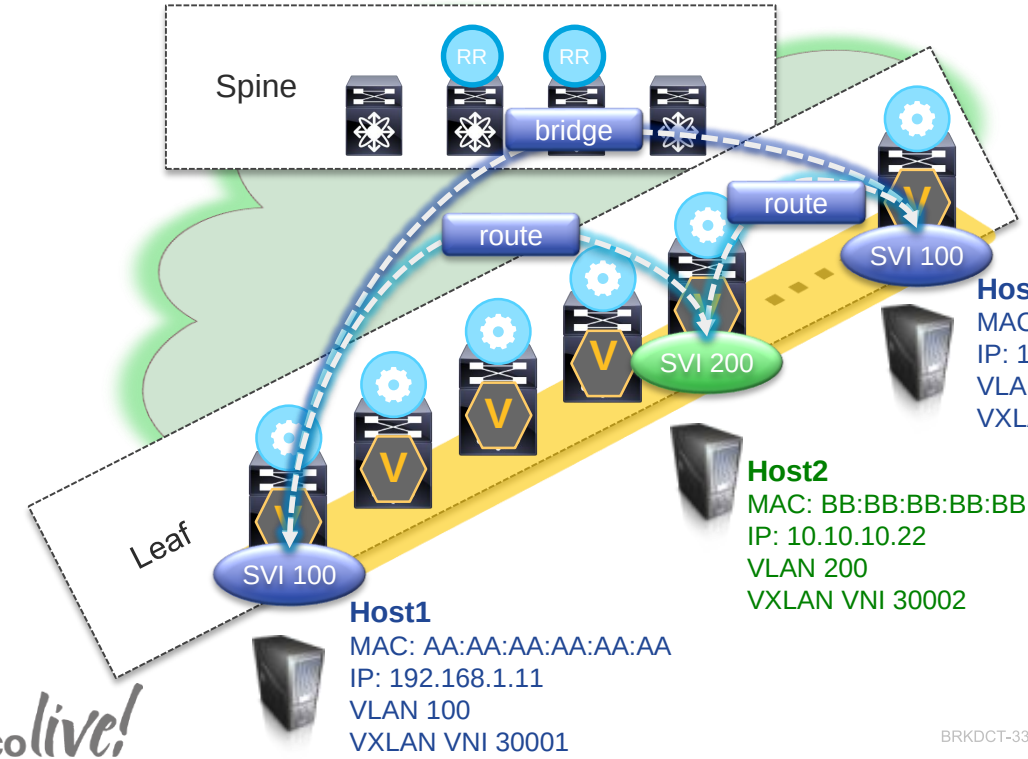
Distributed IP Anycast Gateway

- Distributed Inter-VXLAN Routing at Access Layer (Leaf)
  - All Leafs share same gateway IP and MAC Address for a given Subnet
- Gateway is always active
  - no redundancy protocol, hello exchange etc.
- Distributed state - Smaller ARP tables
  - Only local attached End-Points (Servers)



SVI 100, Gateway IP: 192.168.1.1, Gateway MAC: AG:AG:AG:AG:AG:AG
SVI 200, Gateway IP: 10.10.10.1, Gateway MAC: AG:AG:AG:AG:AG:AG

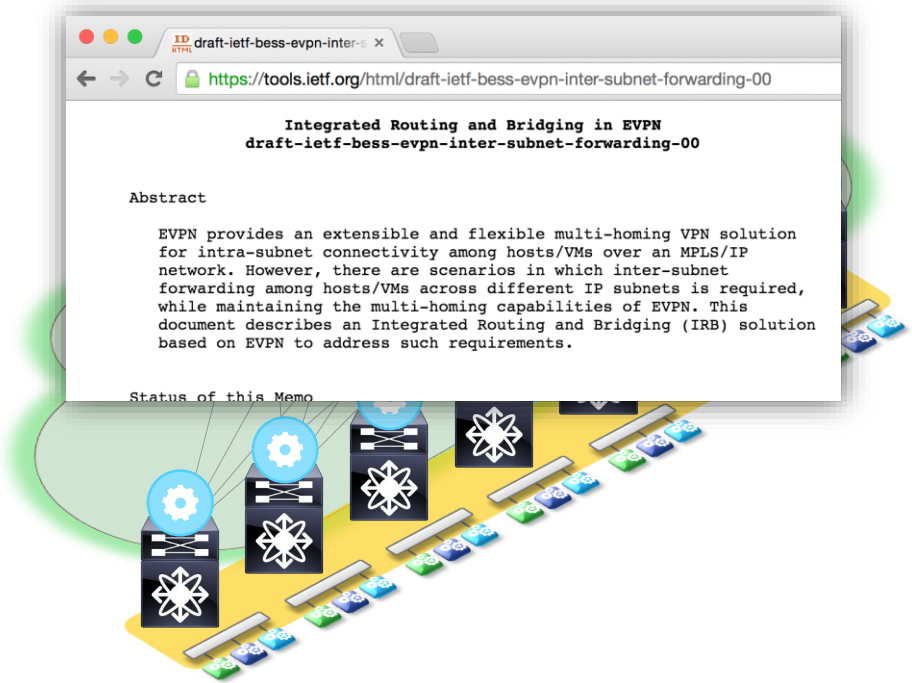
Distributed IP Anycast Gateway



Integrated Routing and Bridging (IRB)

VXLAN/EVPN based overlays follow two slightly different Integrated Routing and Bridging (IRB) semantics

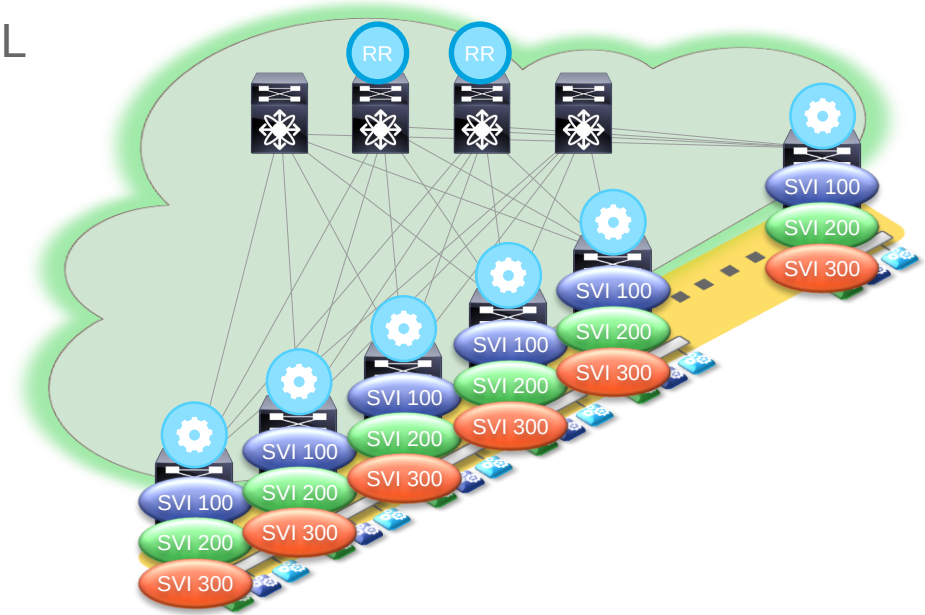
- Asymmetric
  - Uses an “asymmetric path” from the Host towards the egressing port of the VTEP vs. the way back
- Symmetric*
  - Uses an “symmetric path” from the Host towards the egressing port of the VTEP vs. the way back



*Implemented by Cisco’s VXLAN/EVPN

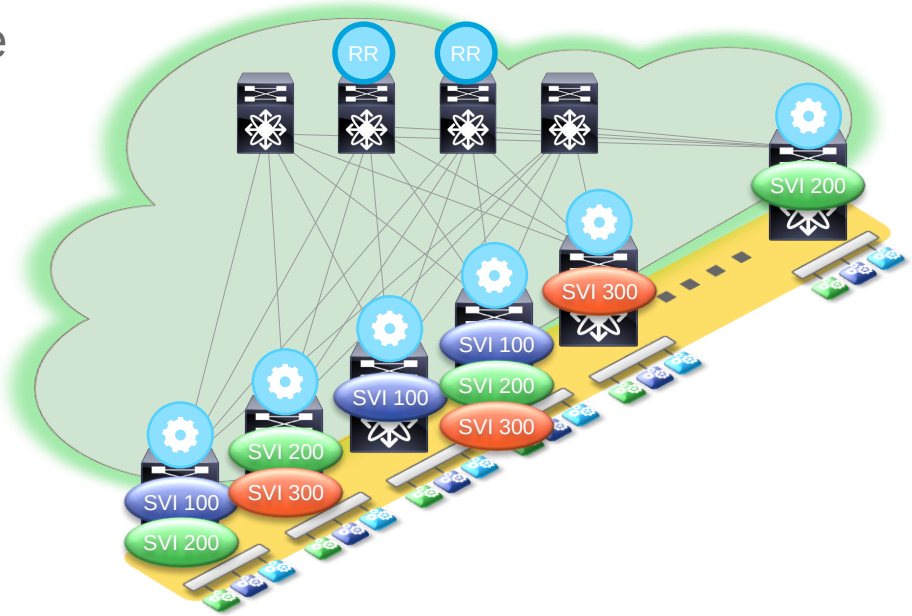
Consistent Configuration

- Logical Configuration (VLAN, VRF, VNI) consistently instantiated on ALL Leafs
- Optimal for Consistency
  - Every VLAN/VNI Everywhere
- Sub-Optimal for Scale
  - Instantiates Resources (VLAN/VNI) even if no End-Point uses it



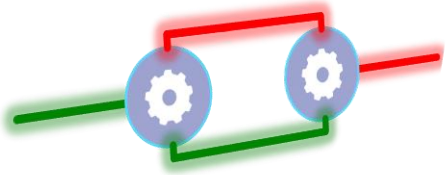
Scoped Configuration

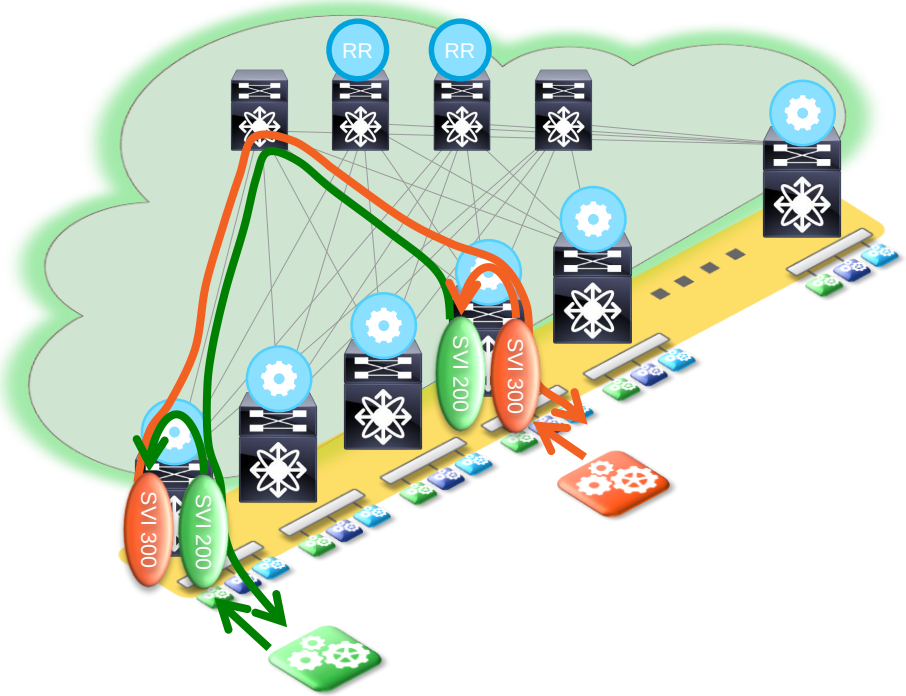
- Logical Configuration (VLAN, VRF, VNI) scoped to Leafs with respective connected End-Points
- Optimal for Scale
  - Instantiates Resources (VLAN/VNI) where End-Points are connected
- Consistency with End-Points
  - Configuration Consistency depends on End-Points



Asymmetric IRB

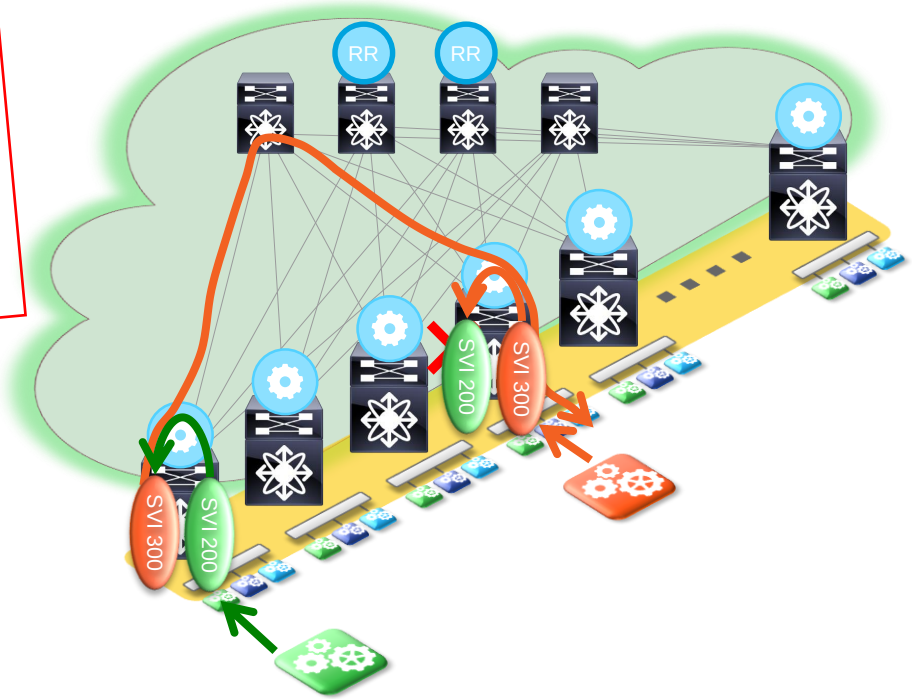
- Similar to todays Inter-VLAN routing
- Requires to follow a consistent configuration of VLAN and L2VNI across all Switches
- Post routed traffic will leverage destination Layer 2 Segment (L2VNI), same as for bridged traffic



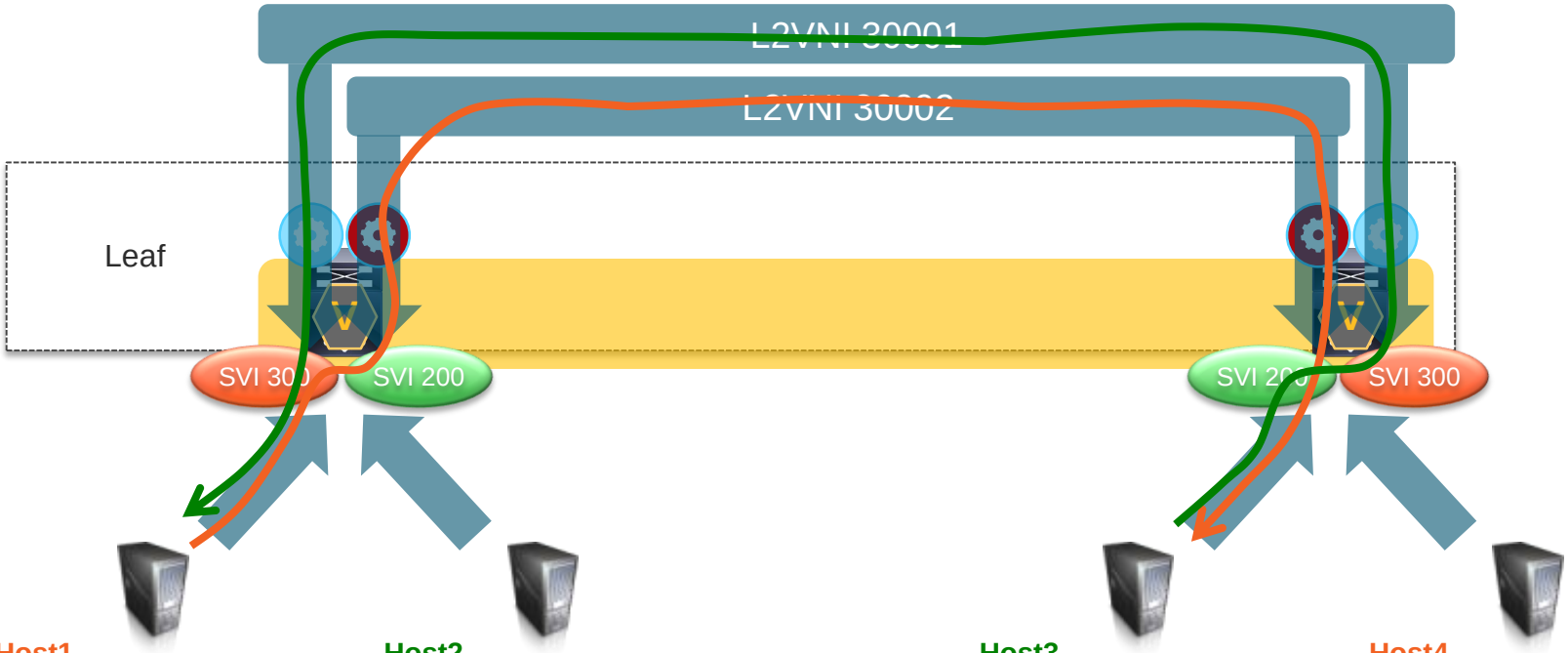


Asymmetric IRB

What happens if NOT all SVIs are on all VTEPs ?
(aka Scoped Configuration)



Asymmetric IRB



Host1
MAC: AA:AA:AA:AA:AA:AA
IP: 192.168.1.11
VLAN 300
VXLAN VNI 30003

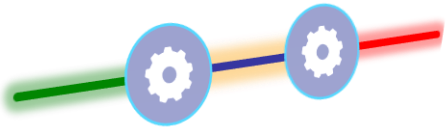
Host2
MAC: BB:BB:BB:BB:BB:BB
IP: 10.10.10.22
VLAN 200
VXLAN VNI 30002

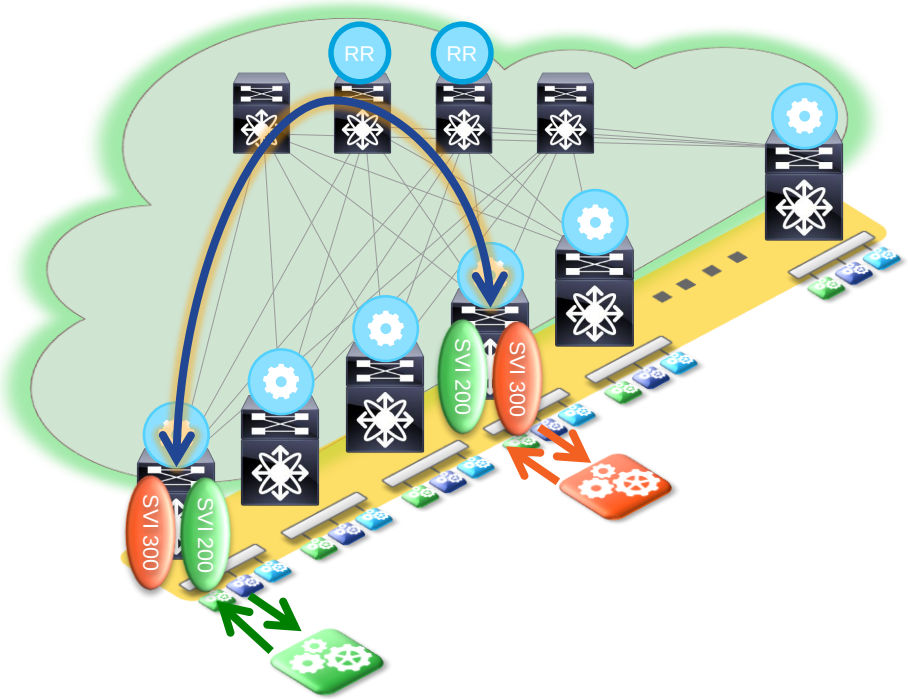
Host3
MAC: CC:CC:CC:CC:CC:CC
IP: 10.10.10.33
VLAN 200
VXLAN VNI 30002

Host4
MAC: DD:DD:DD:DD:DD:DD
IP: 192.168.1.44
VLAN 300
VXLAN VNI 30003

Symmetric IRB

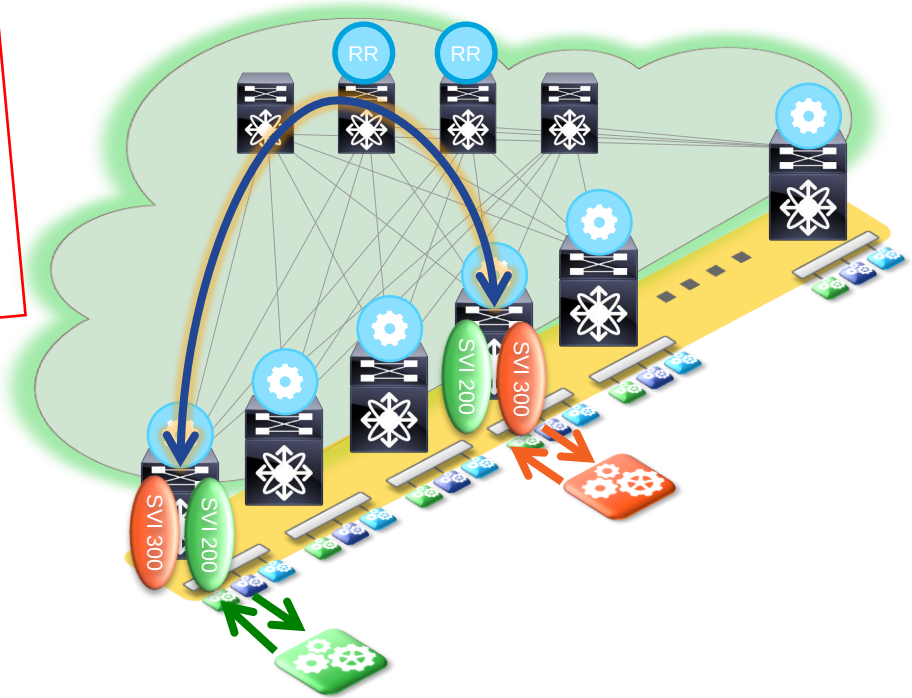
- Similar to Transit Routing Segments
- Scoped Configuration of VLAN/L2VNI; only required where End-Points (Server) reside
- New VNI (L3VNI) introduced per virtual routing and forwarding (VRF) context
- Routed traffic uses transit VNI (L3VNI), while bridged traffic uses L2VNI



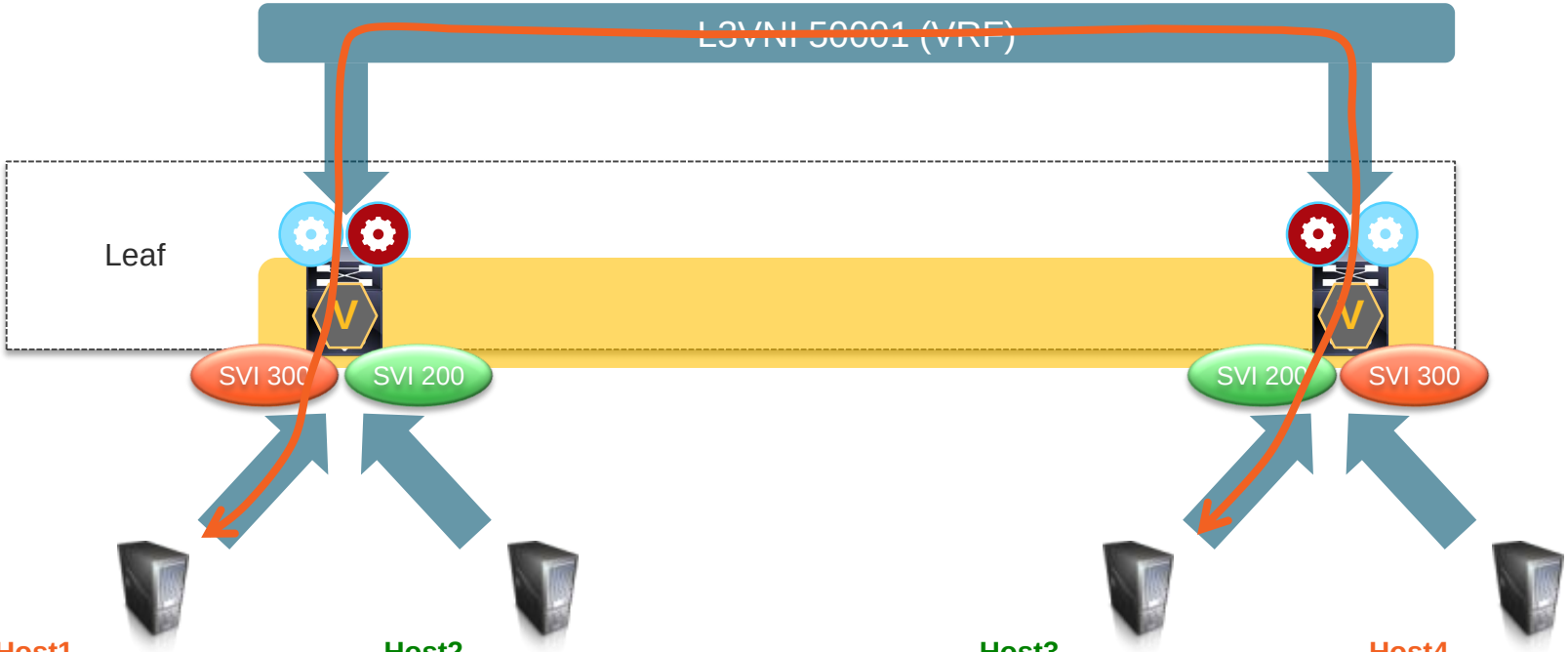


Symmetric IRB

What happens if NOT all SVIs are on all VTEPs ?
(aka Scoped Configuration)



Symmetric IRB



Host1
MAC: AA:AA:AA:AA:AA:AA
IP: 192.168.1.11
VLAN 300
VXLAN VNI 30003

Host2
MAC: BB:BB:BB:BB:BB:BB
IP: 10.10.10.22
VLAN 200
VXLAN VNI 30002

Host3
MAC: CC:CC:CC:CC:CC:CC
IP: 10.10.10.33
VLAN 200
VXLAN VNI 30002

Host4
MAC: DD:DD:DD:DD:DD:DD
IP: 192.168.1.44
VLAN 300
VXLAN VNI 30003

Data Centre Fabric Properties

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

Agenda

- Introduction to Data Centre Fabrics
- VXLAN with BGP EVPN
  - Overview
  - Underlay
  - Control & Data Plane
  - **Multi-Tenancy**
- “Stories” and Use-Cases
- Fabric Management & Automation

What is Multi-Tenancy

- A mode of operation, where multiple independent instances (tenant) operate in a shared environment.

- Each instance (i.e. VRF/VLAN) is logically isolated, but physically integrated.

Where can we apply Multi-Tenancy

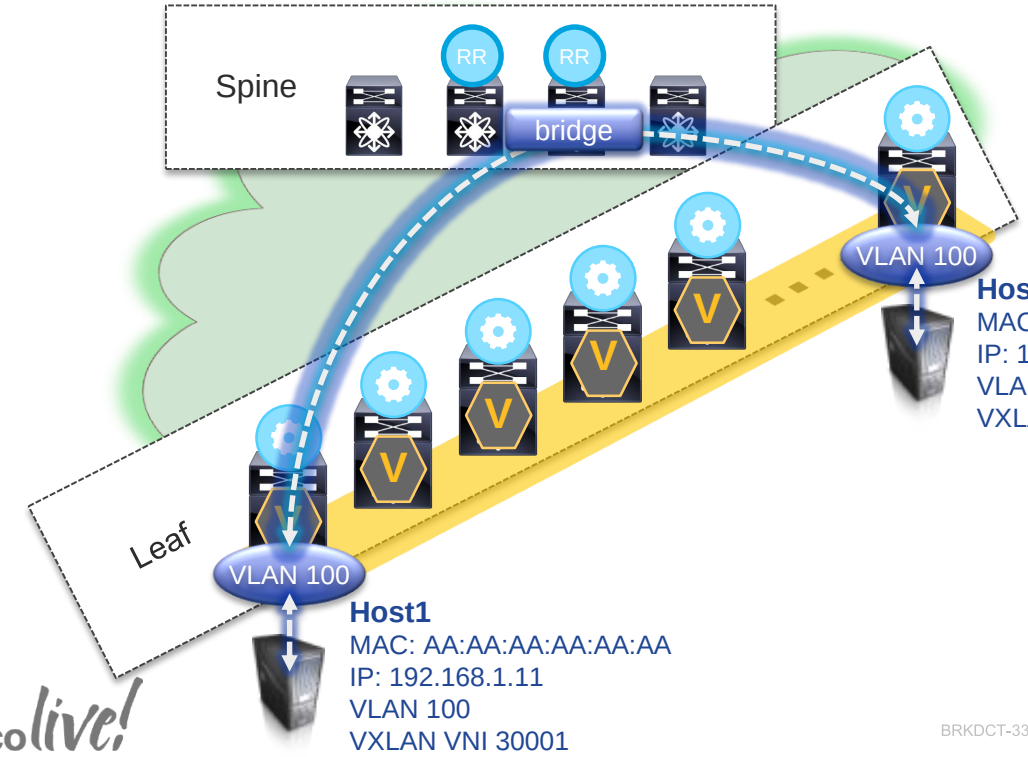
Multi-Tenancy at Layer-2

- Per-Switch VLAN-to-VNI mapping
- Per-Port VLAN Significance

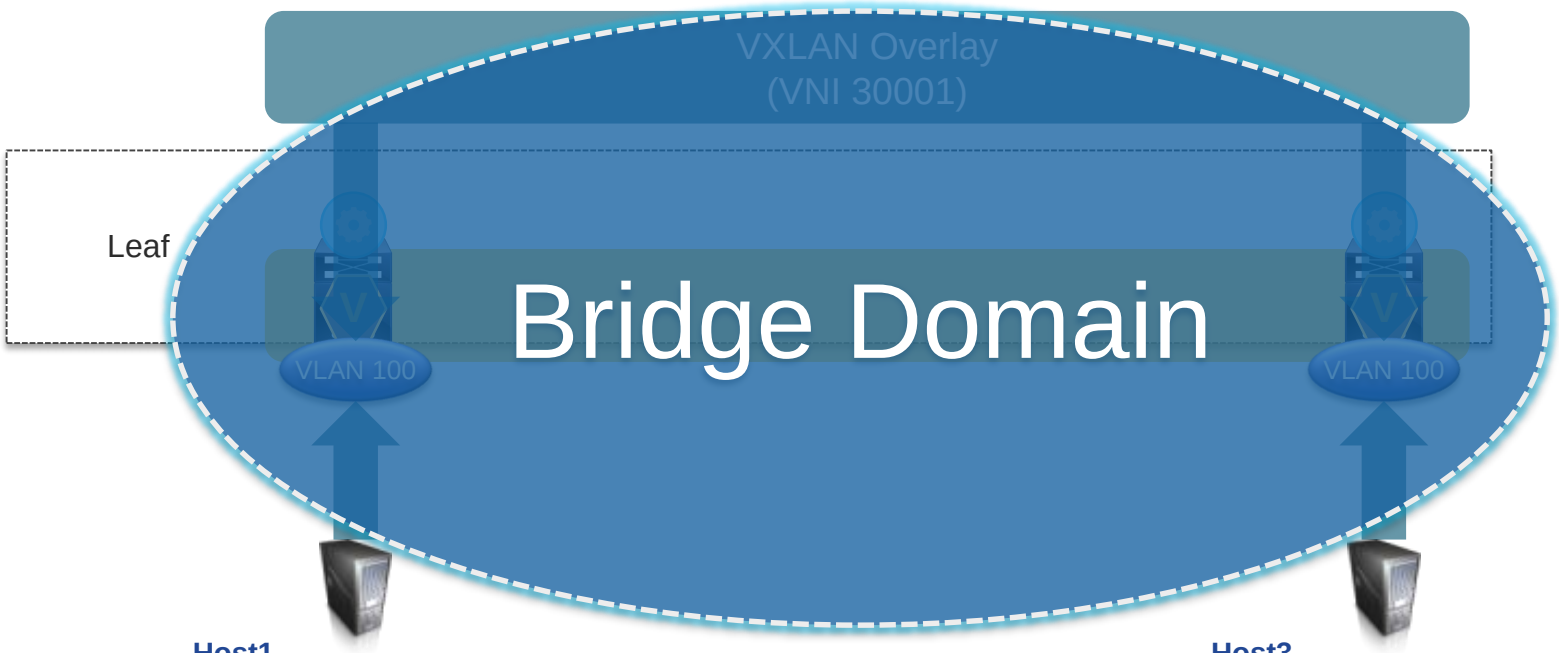
Multi-Tenancy at Layer-3

- VRF-to-VNI mapping
- MP-BGP for scaling with VPNs

Layer-2 Multi-Tenancy



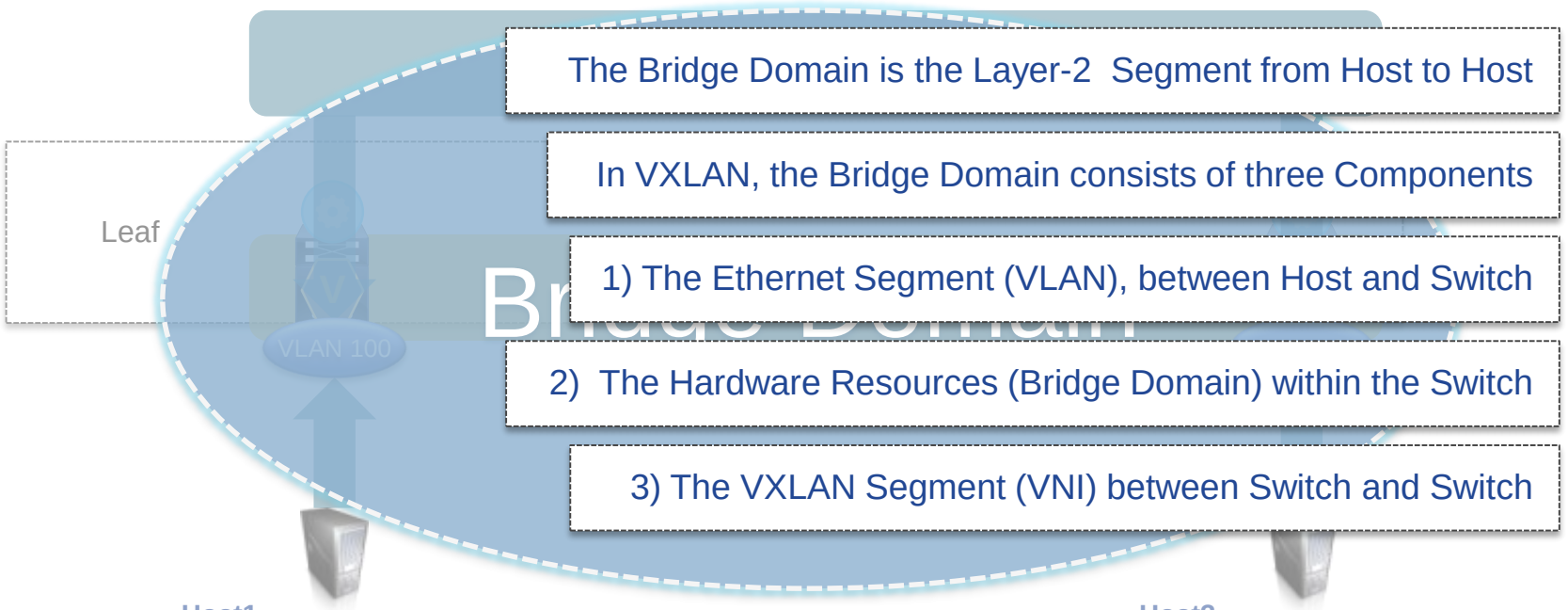
Layer-2 Multi-Tenancy – Bridge Domains



Host1
MAC: AA:AA:AA:AA:AA:AA
IP: 192.168.1.11
VLAN 100
VXLAN VNI 30001

Host3
MAC: CC:CC:CC:CC:CC:CC
IP: 192.168.1.33
VLAN 100
VXLAN VNI 30001

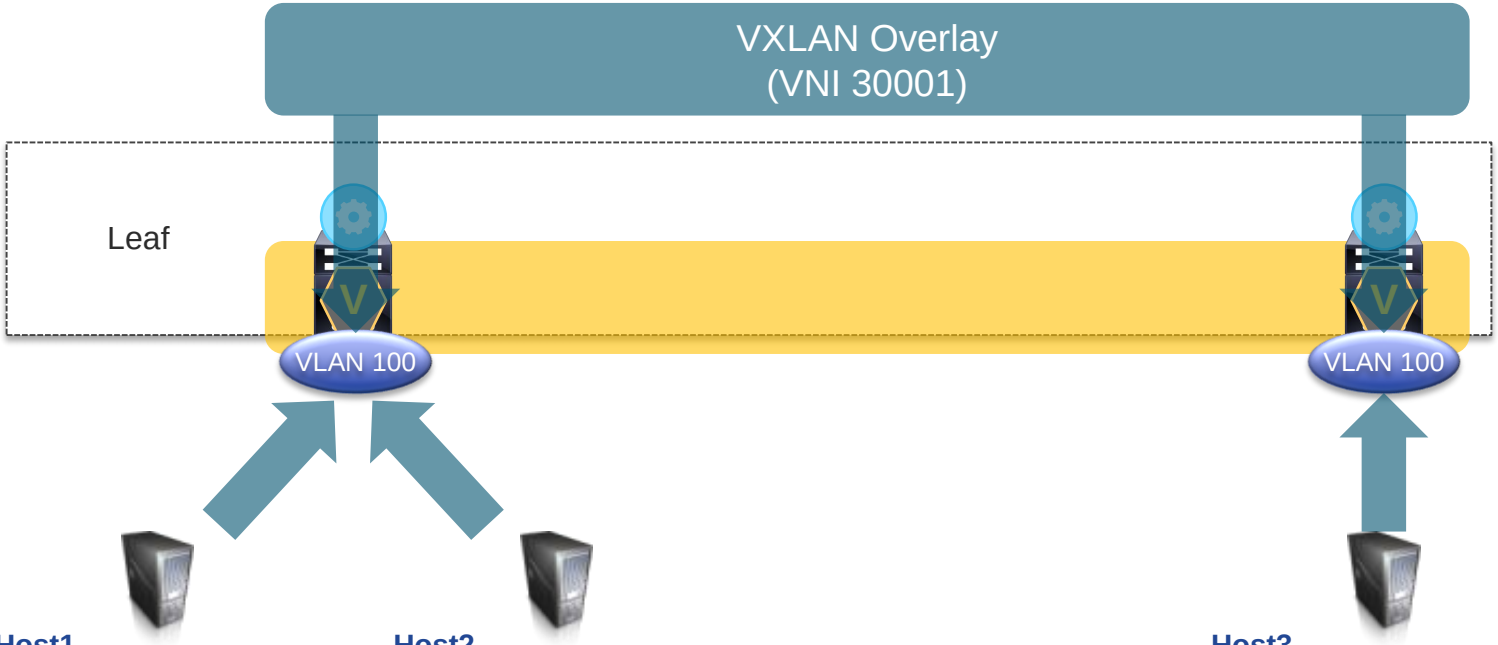
Layer-2 Multi-Tenancy – Bridge Domains



Host1
MAC: AA:AA:AA:AA:AA:AA
IP: 192.168.1.11
VLAN 100
VXLAN VNI 30001

Host3
MAC: CC:CC:CC:CC:CC:CC
IP: 192.168.1.33
VLAN 100
VXLAN VNI 30001

VLAN-to-VNI Mapping



Host1
MAC: AA:AA:AA:AA:AA:AA
IP: 192.168.1.11
VLAN 100
VXLAN VNI 30001

Host2
MAC: BB:BB:BB:BB:BB:BB
IP: 192.168.1.22
VLAN 100
VXLAN VNI 30001

Host3
MAC: CC:CC:CC:CC:CC:CC
IP: 192.168.1.33
VLAN 100
VXLAN VNI 30001

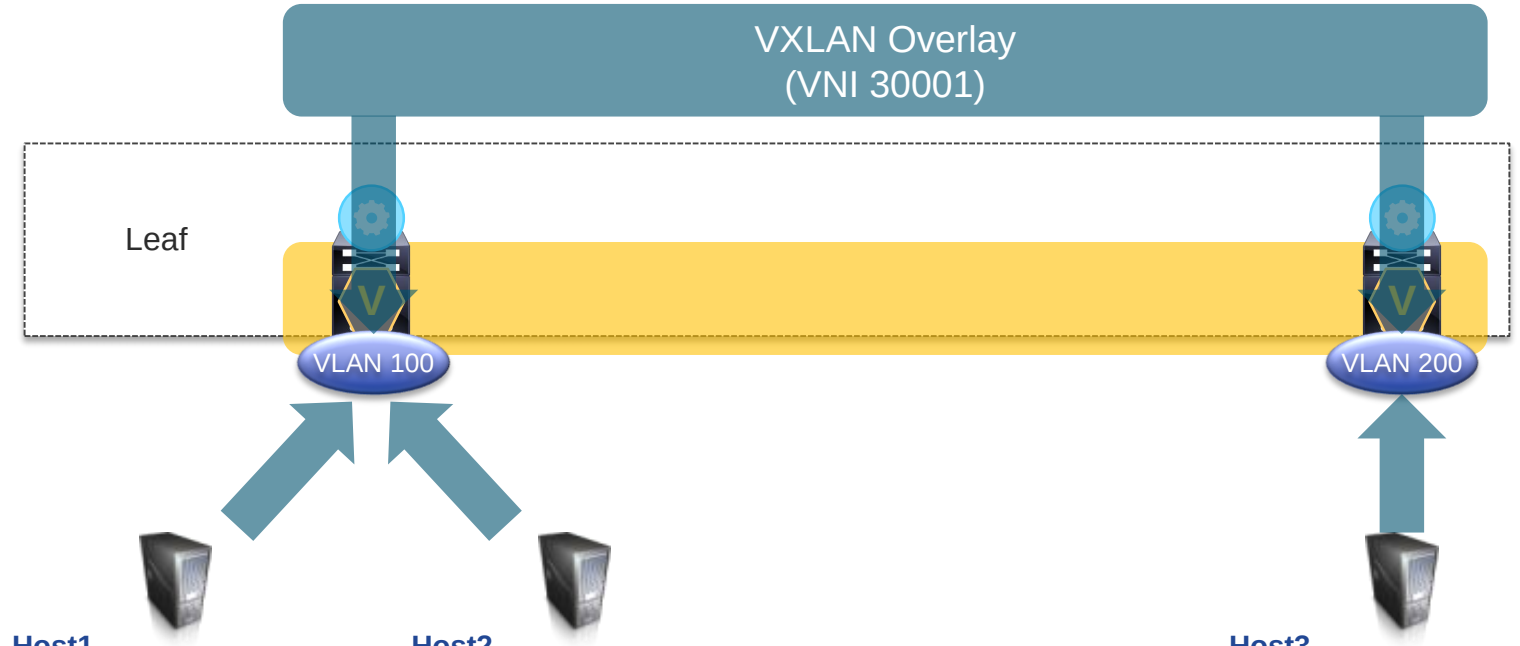
FYI

CLI Modes - VLAN based (per-Switch)

```
Leaf#1
vlan 100
 vn-segment 30001

Leaf#2
vlan 100
 vn-segment 30001
```

- VLAN to VNI configuration on a per-switch basis
- VLAN becomes “Switch Local Identifier”
- VNI becomes “Network Global Identifier”

Per-Switch VLAN-to-VNI Mapping



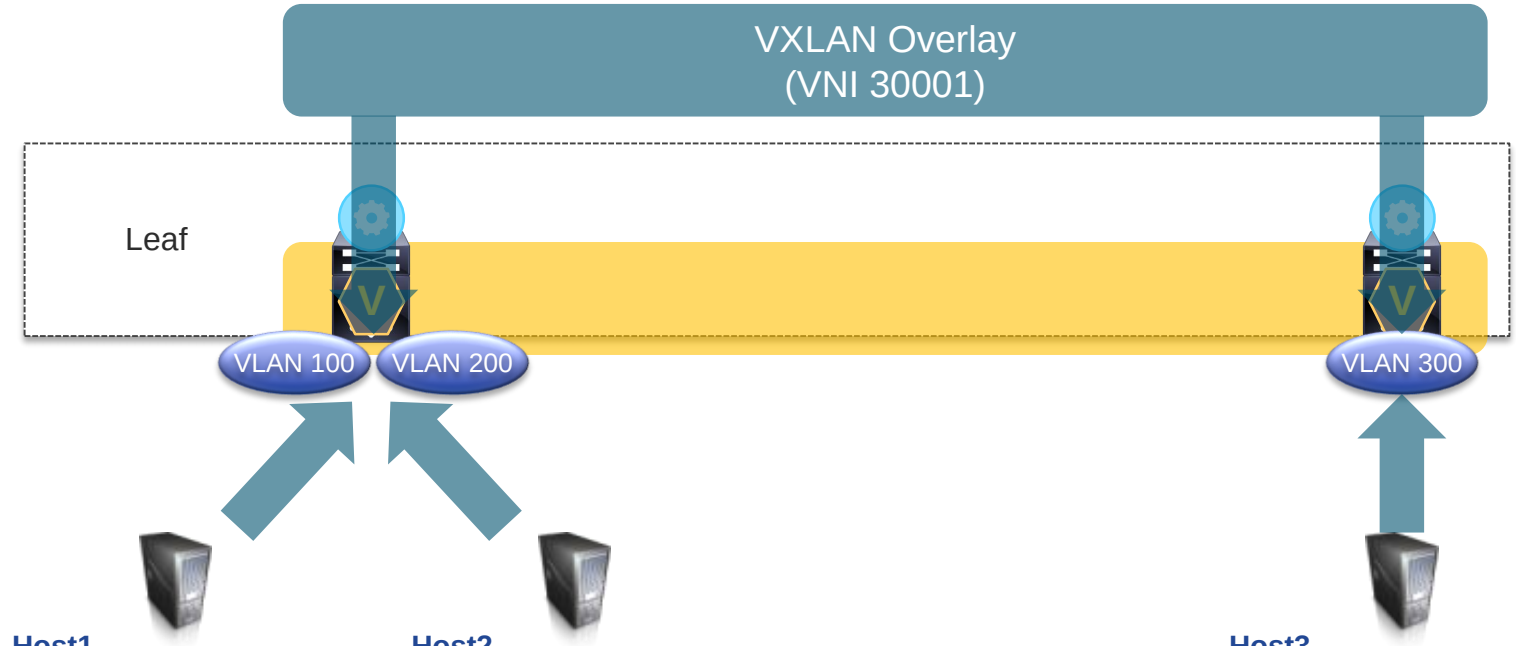
FYI

CLI Modes - VLAN Based (per-Switch)

```
Leaf#1
vlan 100
 vn-segment 30001

Leaf#2
vlan 200
 vn-segment 30001
```

- VLAN to VNI configuration on a per-switch basis
- VLAN becomes “Switch Local Identifier”
- VNI becomes “Network Global Identifier”
- 4k VLAN limitation has been removed

Per-Port VLAN-to-VNI Mapping



FYI

CLI Modes - VLAN Based (per-Port)

```
Leaf#1
vlan 2500
 vn-segment 30001

interface Ethernet 1/8
 switchport mode trunk
 switchport vlan mapping enable
 switchport vlan mapping  100 2500

interface Ethernet 1/9
 switchport mode trunk
 switchport vlan mapping enable
 switchport vlan mapping  200 2500
```

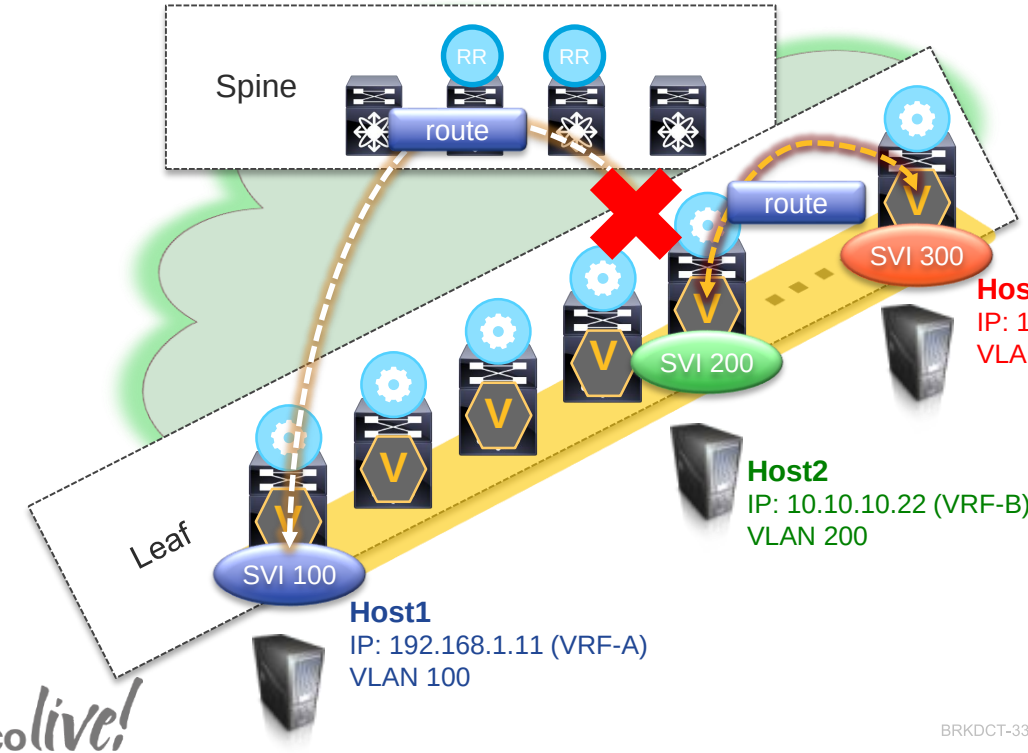
FYI

CLI Modes - Bridge-Domain Based (per-Port)

```
Leaf#1
bridge-domain 100
 member vni 30001

encapsulation profile vni VLAN100-30001
 dot1q 100 vni 30001

encapsulation profile
 dot1q 200 vni 30001
```

```
interface Ethernet 1/8
 no switchport
 service instance 1 vni
 encapsulation profile VLAN100-30001 default

interface Ethernet 1/9
 no switchport
 service instance 1 vni
 encapsulation profile VLAN200-30001 default
```

Layer-3 Multi-Tenancy



VRF-A (VNI 50001)
VRF-B (VNI 50002)

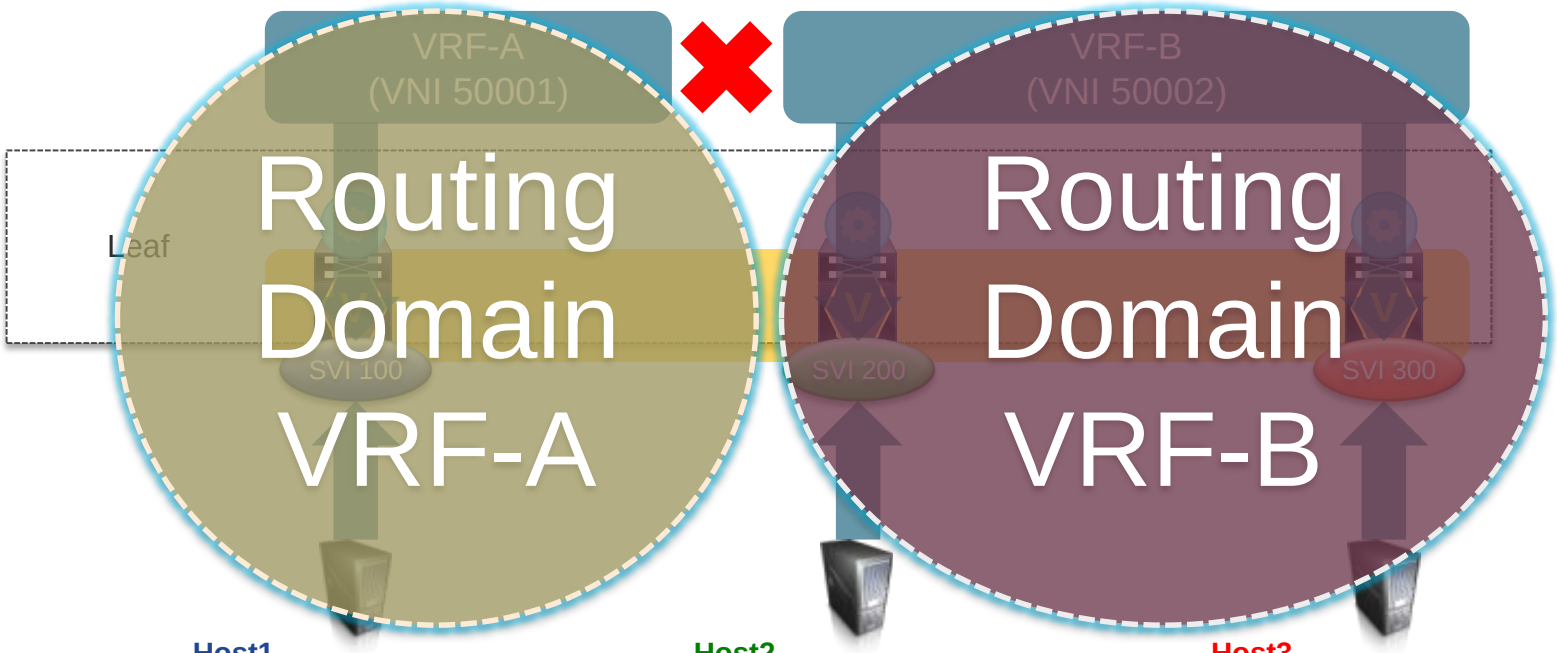
SVI 100, Gateway IP: 192.168.1.1 (VRF-A)
SVI 200, Gateway IP: 10.10.10.1 (VRF-B)
SVI 300, Gateway IP: 172.16.1.1 (VRF-B)

Host3
IP: 172.16.1.33 (VRF-B)
VLAN 300

Host2
IP: 10.10.10.22 (VRF-B)
VLAN 200

Host1
IP: 192.168.1.11 (VRF-A)
VLAN 100

Layer-3 Multi-Tenancy – VRF-VNI or L3VNI

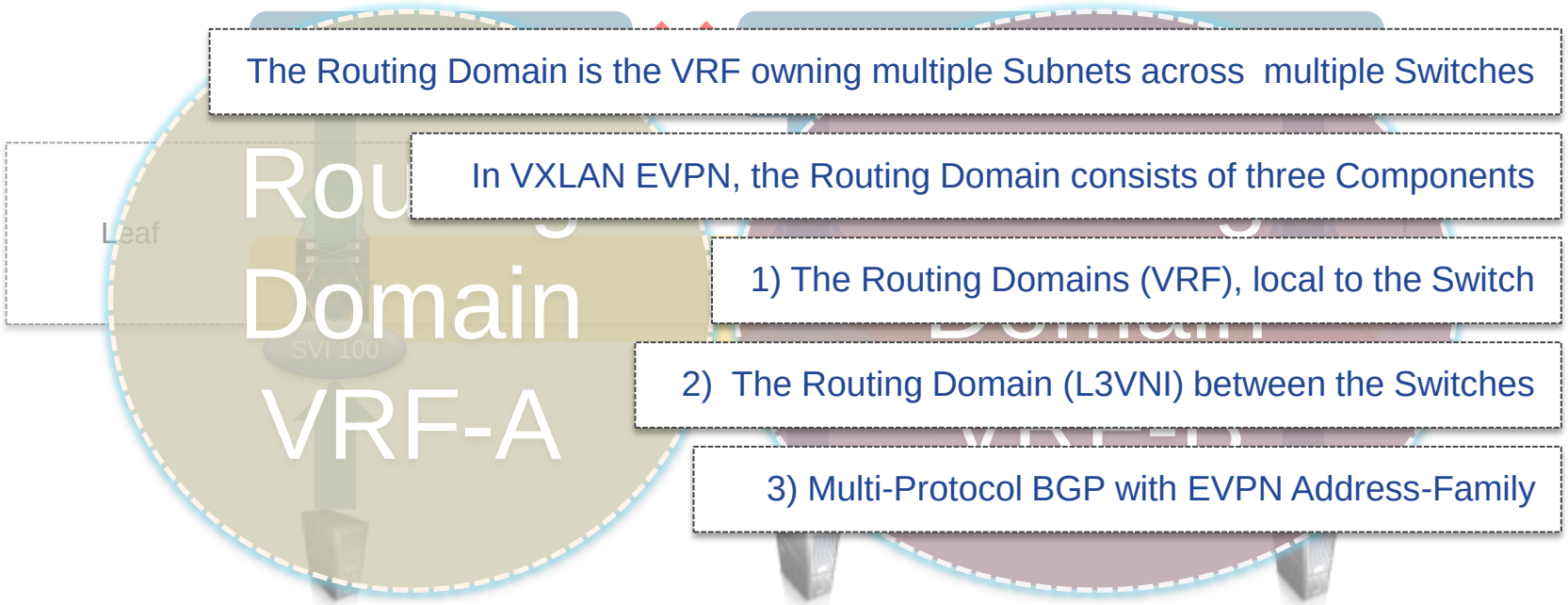


Host1
IP: 192.168.1.11 (VRF-A)
VLAN 100

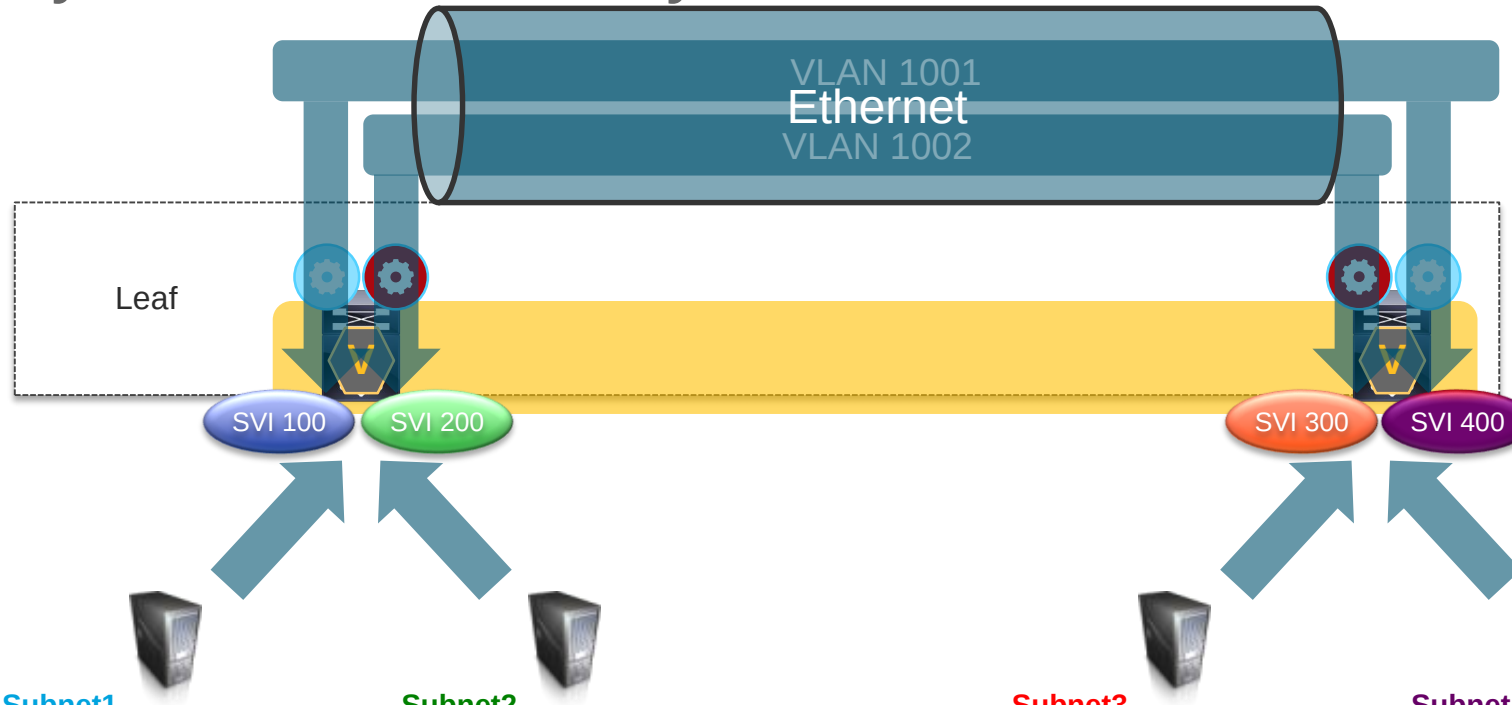
Host2
IP: 10.10.10.22 (VRF-B)
VLAN 200

Host3
IP: 172.16.1.33 (VRF-B)
VLAN 300

Layer-3 Multi-Tenancy – VRF-VNI or L3VNI



Layer-3 Multi-Tenancy – VRF-Lite



Subnet1
IP: 192.168.1.0/24 (VRF-A)
VLAN 100

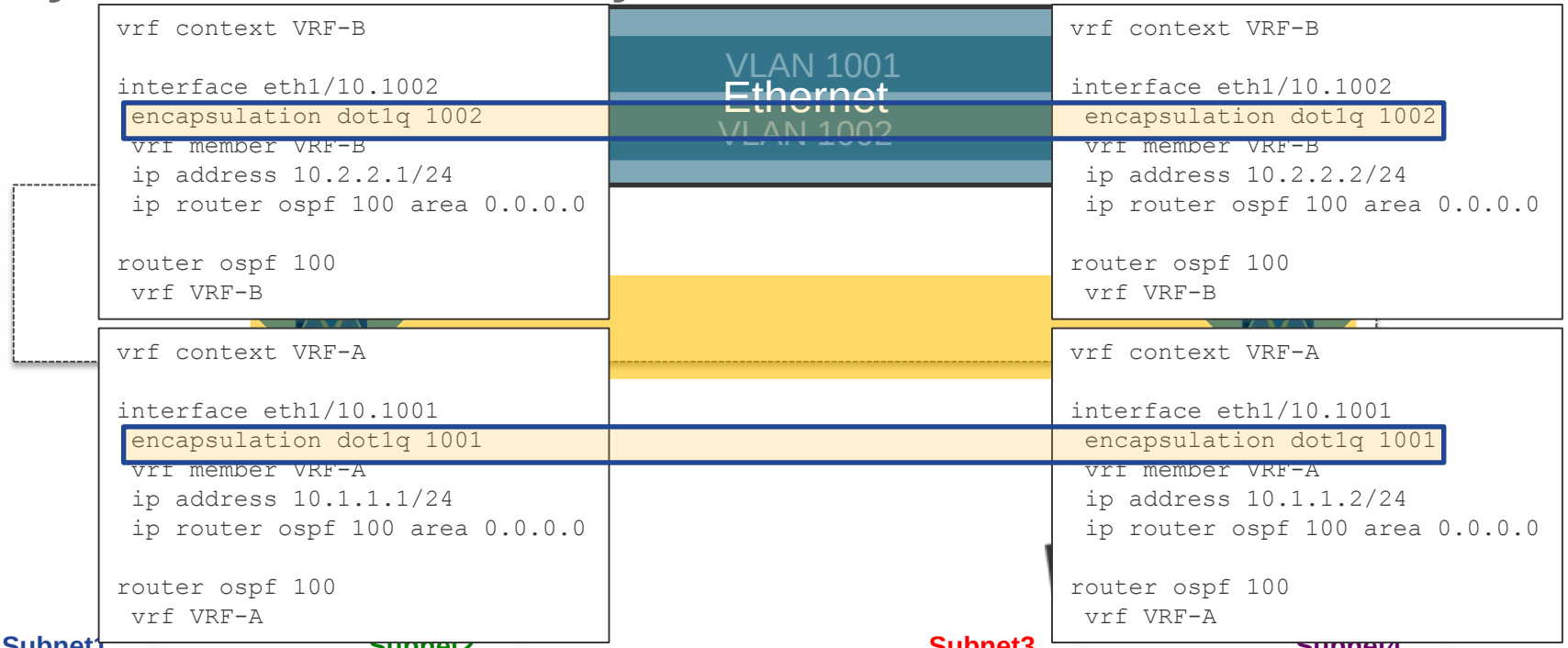
Subnet2
IP: 10.10.10.0/24 (VRF-B)
VLAN 200

Subnet3
IP: 172.16.1.0/24 (VRF-B)
VLAN 300

Subnet4
IP: 10.44.44.0/24 (VRF-A)
VLAN 400

FYI

Layer-3 Multi-Tenancy – VRF-Lite



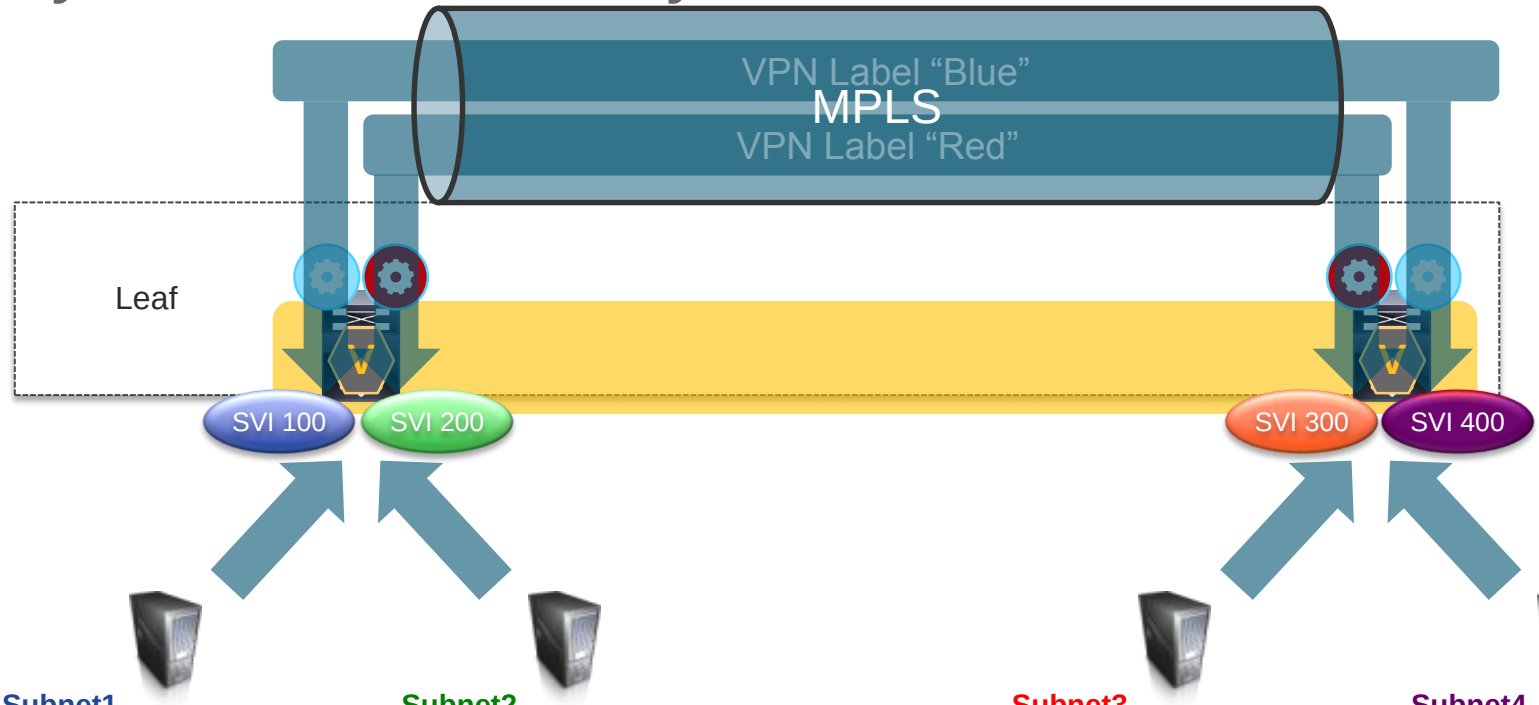
Subnet1
IP: 192.168.1.0/24 (VRF-A)
VLAN 100

Subnet2
IP: 10.10.10.0/24 (VRF-B)
VLAN 200

Subnet3
IP: 172.16.1.0/24 (VRF-B)
VLAN 300

Subnet4
IP: 10.44.44.0/24 (VRF-A)
VLAN 400

Layer-3 Multi-Tenancy – MPLS L3VPN



Subnet1
IP: 192.168.1.0/24 (VRF-A)
VLAN 100

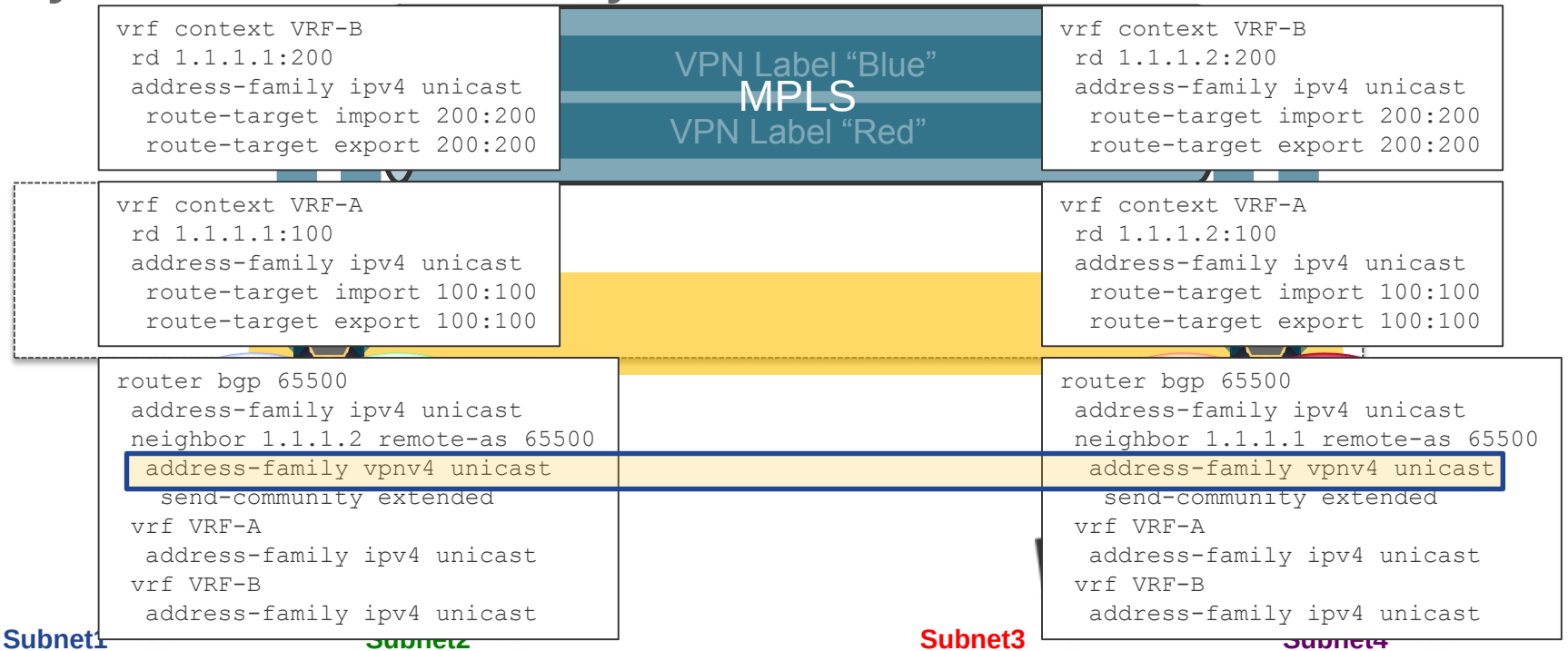
Subnet2
IP: 10.10.10.0/24 (VRF-B)
VLAN 200

Subnet3
IP: 172.16.1.0/24 (VRF-B)
VLAN 300

Subnet4
IP: 10.44.44.0/24 (VRF-A)
VLAN 400

FYI

Layer-3 Multi-Tenancy – MPLS L3VPN



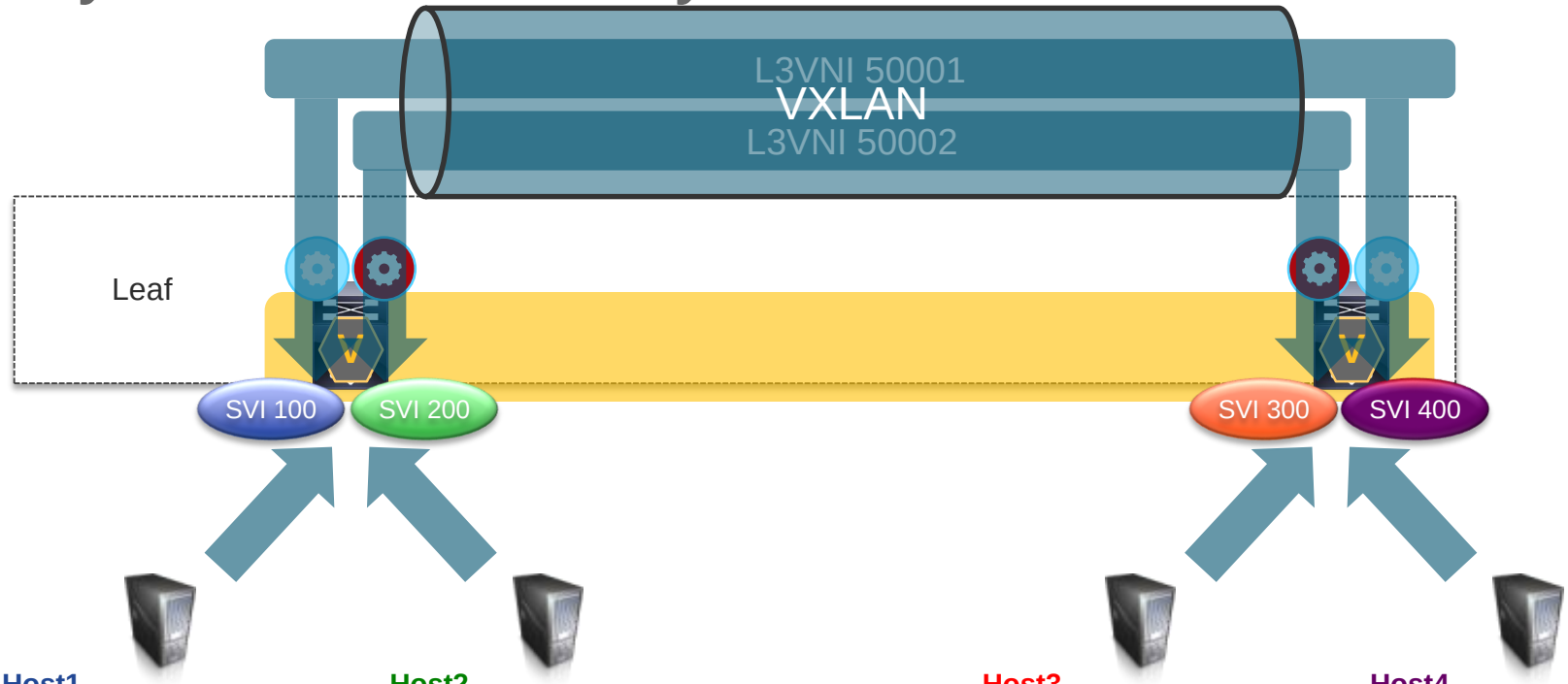
Subnet1
IP: 192.168.1.0/24 (VRF-A)
VLAN 100

Subnet2
IP: 10.10.10.0/24 (VRF-B)
VLAN 200

Subnet3
IP: 172.16.1.0/24 (VRF-B)
VLAN 300

Subnet4
IP: 10.44.44.0/24 (VRF-A)
VLAN 400

Layer-3 Multi-Tenancy – VXLAN EVPN



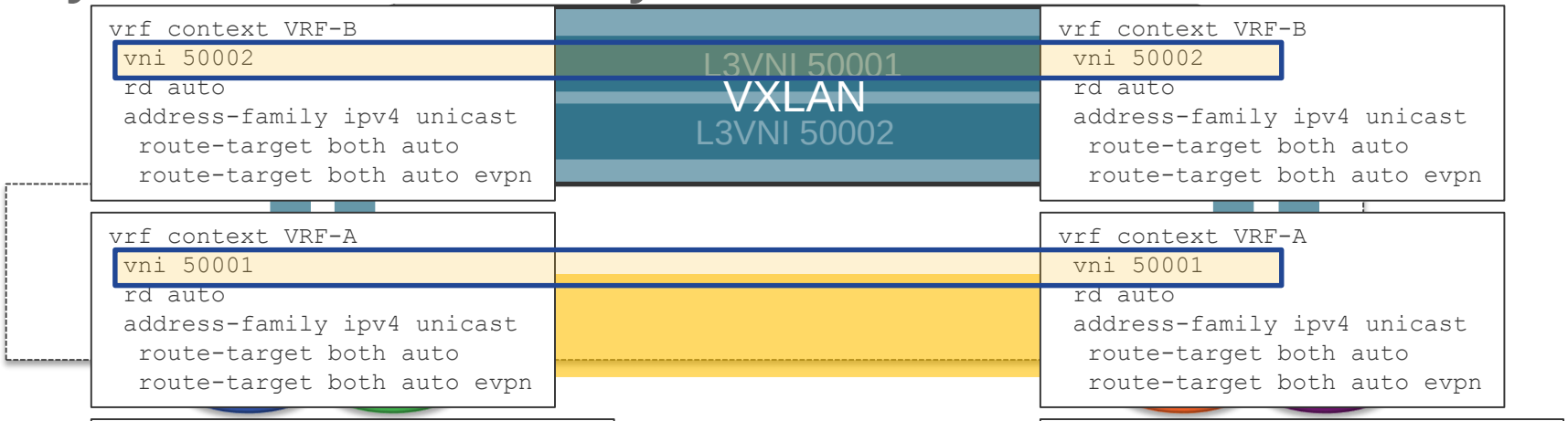
Host1
MAC: AA:AA:AA:AA:AA:AA
IP: 192.168.1.11 (VRF-A)
VLAN 100
VXLAN VNI 30001

Host2
MAC: BB:BB:BB:BB:BB:BB
IP: 10.10.10.22 (VRF-B)
VLAN 200
VXLAN VNI 30002

Host3
MAC: CC:CC:CC:CC:CC:CC
IP: 172.16.1.33 (VRF-B)
VLAN 300
VXLAN VNI 30003

Host4
MAC: DD:DD:DD:DD:DD:DD
IP: 10.44.44.44 (VRF-A)
VLAN 400
VXLAN VNI 30004

Layer-3 Multi-Tenancy – VXLAN EVPN



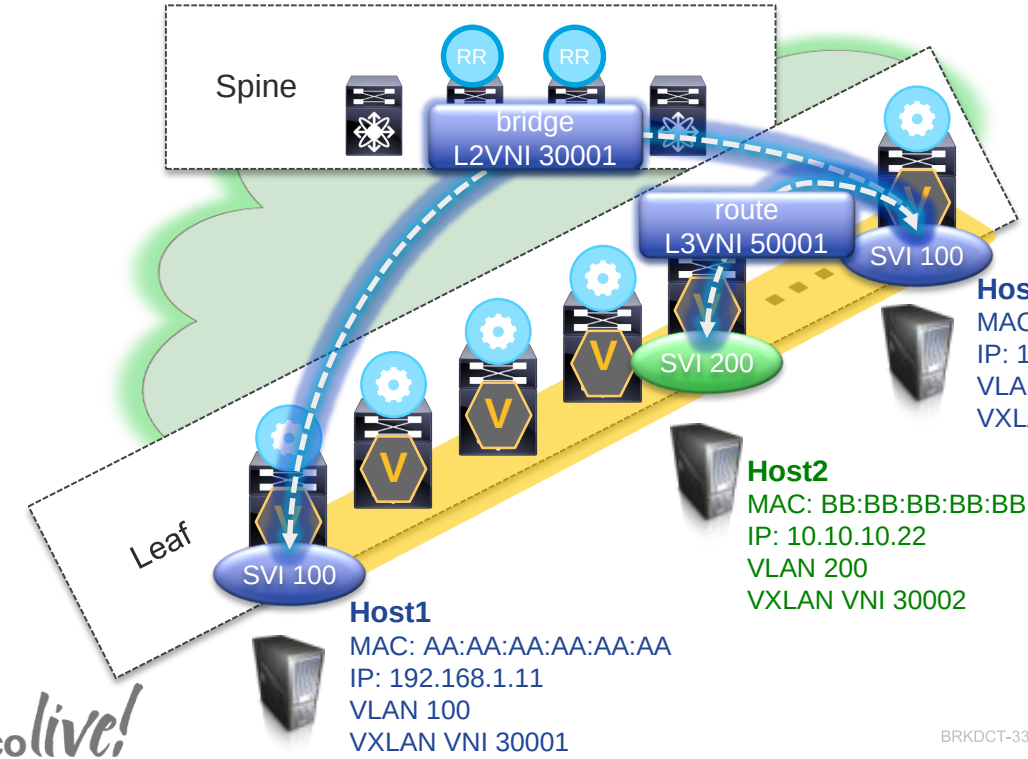
Host1
MAC: A
IP: 192.
VLAN 1
VXLAN VNI 30001

BB:BB
-B)
VXLAN VNI 30002

Host3
MAC: CC:
IP: 172.16
VLAN 300
VXLAN VNI 30003

D:DD
)
VXLAN VNI 30004

Integrated Route & Bridge + Multi-Tenancy



VRF-A (VNI 50001)

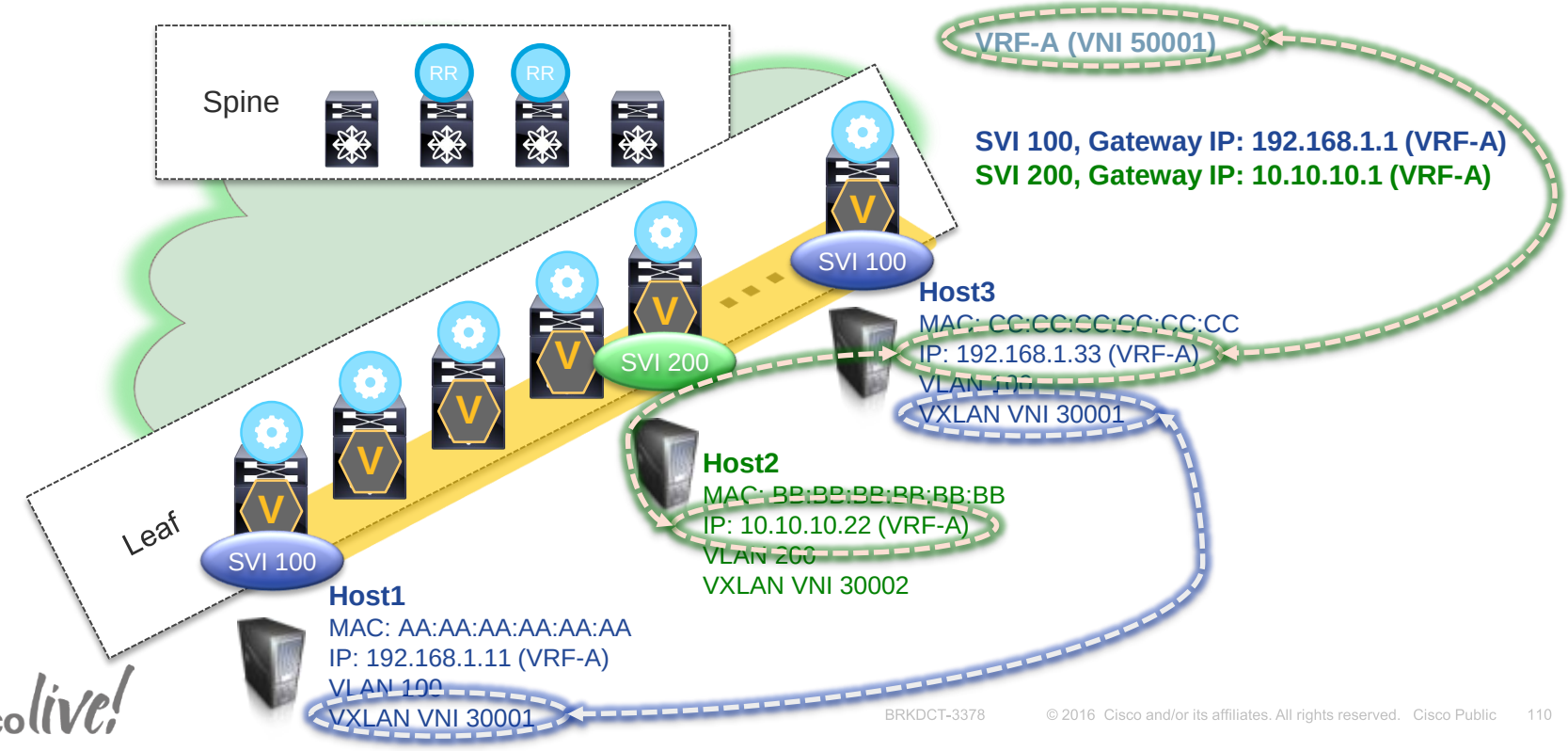
SVI 100, Gateway IP: 192.168.1.1 (VRF-A)
SVI 200, Gateway IP: 10.10.10.1 (VRF-A)

Host3
MAC: CC:CC:CC:CC:CC:CC
IP: 192.168.1.33
VLAN 100
VXLAN VNI 30001

Host2
MAC: BB:BB:BB:BB:BB:BB
IP: 10.10.10.22
VLAN 200
VXLAN VNI 30002

Host1
MAC: AA:AA:AA:AA:AA:AA
IP: 192.168.1.11
VLAN 100
VXLAN VNI 30001

Integrated Route & Bridge + Multi-Tenancy



Data Centre Fabric Properties

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

Agenda

- Introduction to Data Centre Fabrics
- VXLAN with BGP EVPN
  - Overview
  - Underlay
  - Control & Data Plane
  - Multi-Tenancy
- **“Stories” and Use-Cases**
- Fabric Management & Automation

“Stories” and Use-Cases

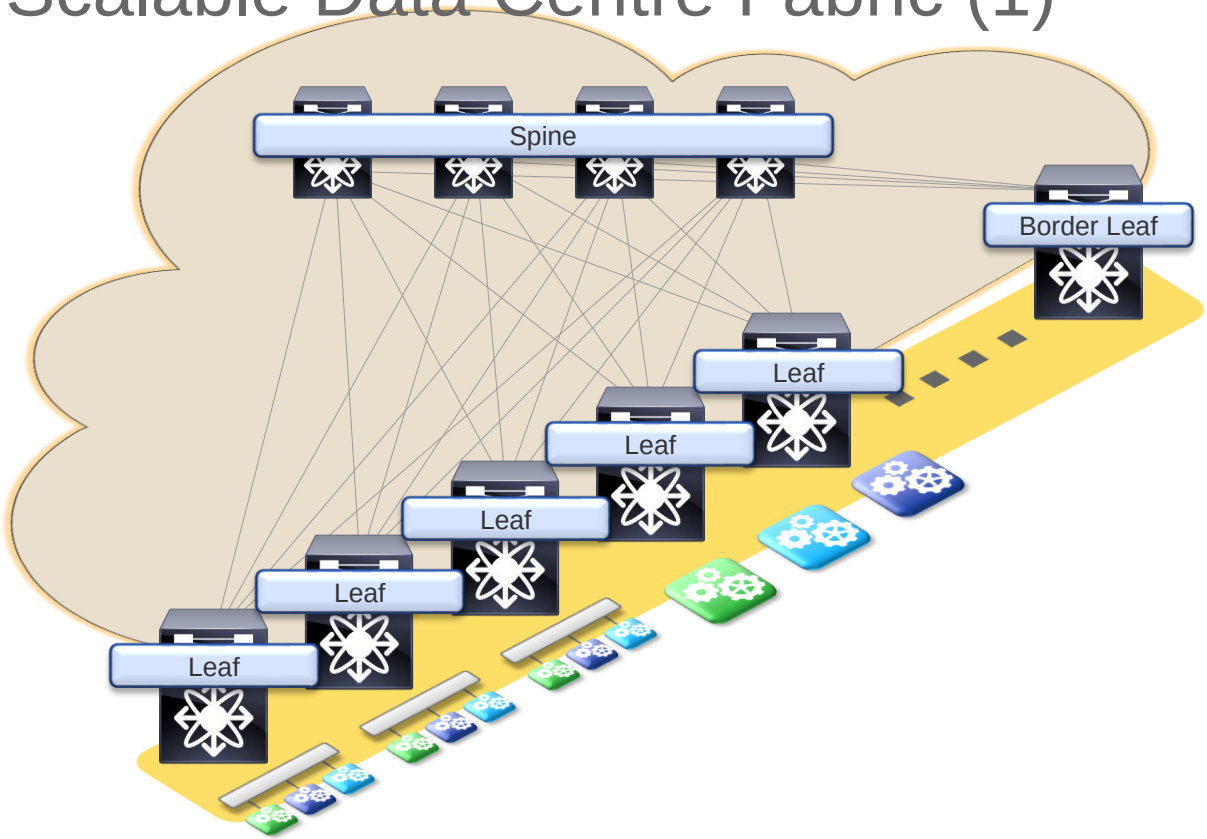
VXLAN applicability evolves as the Control Plane evolves!

- 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)

Story #1: Scalable Data Centre Fabric

- VXLAN based Data Centre Fabric
- BGP EVPN Control-Protocol (Overlay)
- OSPF for Underlay Routing (Unicast)
- PIM ASM with Anycast-RP for BUM Replication (Underlay)
- Distributed IP Anycast Gateway

Story #1: Scalable Data Centre Fabric (1)



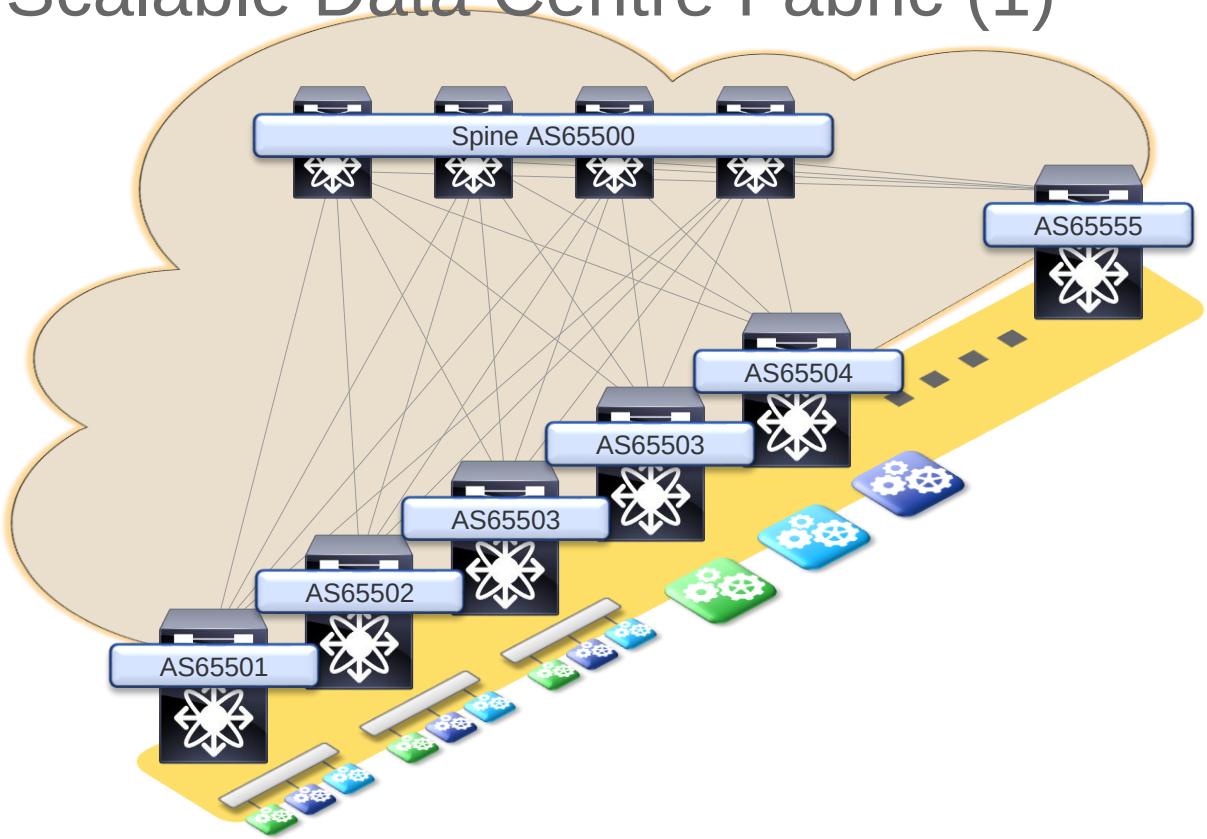
p2p Agg: 10.1.1.0/24
RID Agg: 10.10.10.0/24
VTEP Agg: 10.200.200.0/24
RP Agg: 10.254.254.0/24

Story #2: Scalable Data Centre Fabric

- VXLAN based Data Centre Fabric
- BGP EVPN Control-Protocol (Overlay)
- eBGP for Underlay Routing (Unicast)
- eBGP Multi-AS Design
- Ingress Replication for BUM (Underlay)
- Distributed IP Anycast Gateway

Note: Configurations do NOT claim completeness

Story #2: Scalable Data Centre Fabric (1)



p2p Agg: 10.1.1.0/24
RID Agg: 10.10.10.0/24
VTEP Agg: 10.200.200.0/24
RP Agg: 10.254.254.0/24

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?

Note sure if it is an Elephant

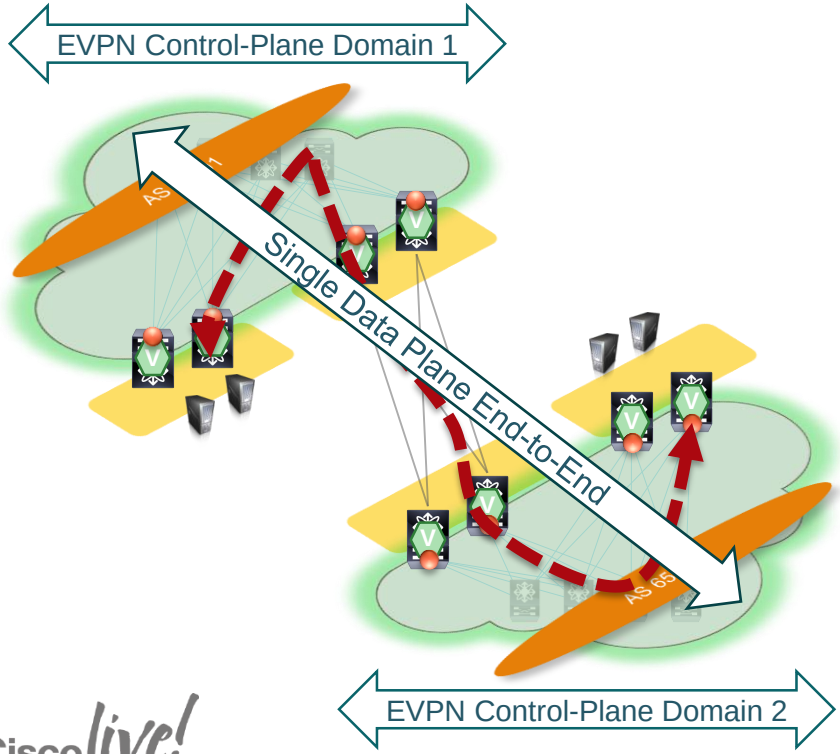
VXLAN for Interconnecting Networks

Story #3: Inter-Fabric Connectivity

Note: Check Release Notes for Feature Support

- Option 1: End-to-End Fabric Stretch
- Option 2: Fabric-DCI-Fabric (2-box)
- Option 3: Fabric-DCI-Fabric L3-DCI (1-box)
- Option 4: Fabric-DCI-Fabric L2-DCI (1-box)

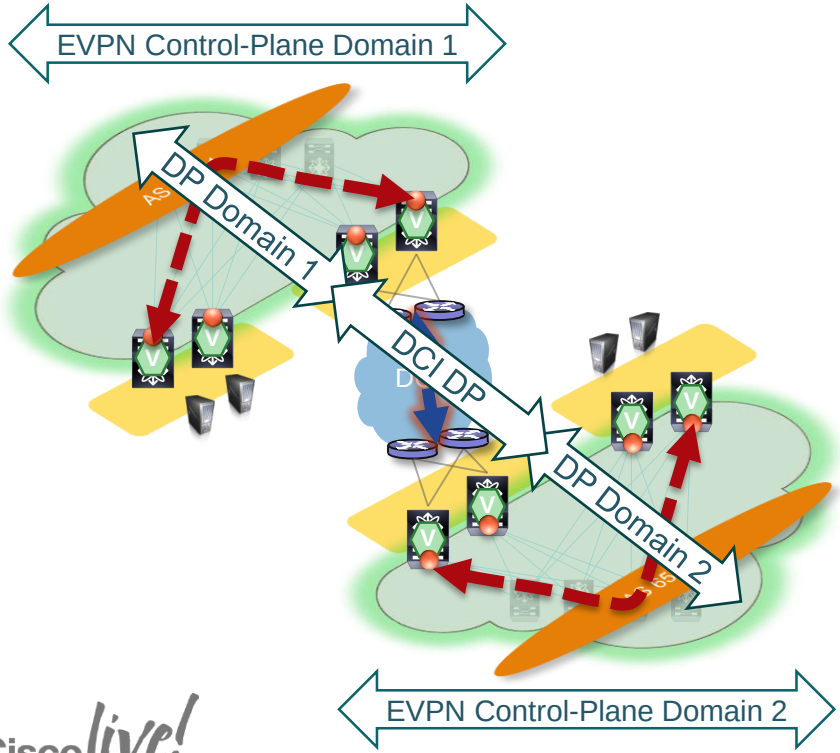
Inter-Fabric Connectivity (Option 1)



- Multiple BGP-EVPN Control-Plane Domains
- End-to-End reachability for VTEP
- End-to-End reachability for BUM Replication
  - Multicast / Ingress Replication
- End-to-End Data-Plane encapsulation

VXLAN Encapsulation

Inter-Fabric Connectivity (Option 2)

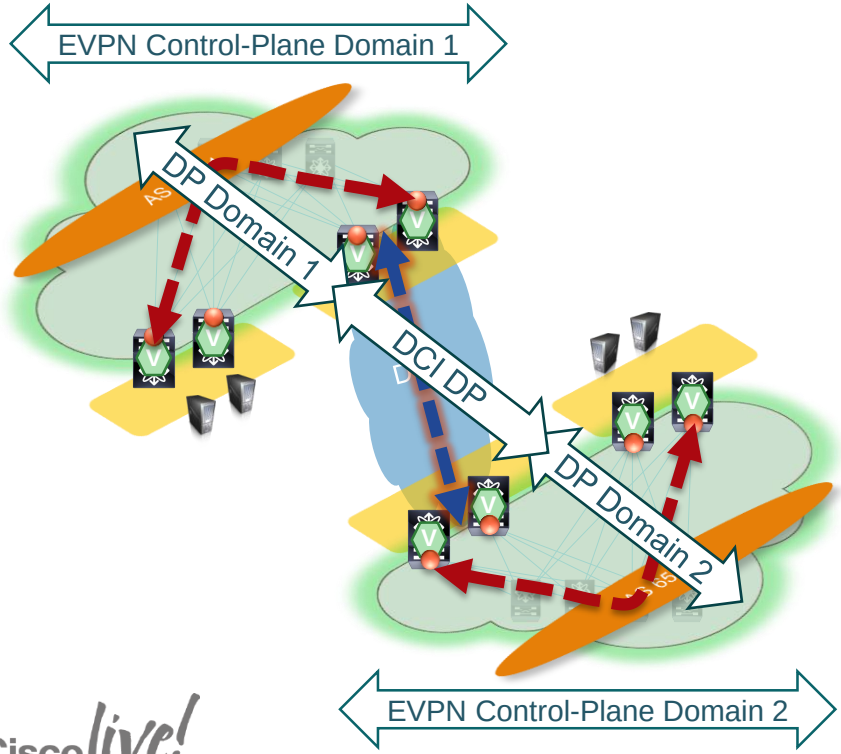


- Multiple BGP-EVPN Control-Plane Domains
- Normalisation via Ethernet (MPLS, VRF-lite & IEEE 802.1Q Trunk) at the Border
- Separate Data-Plane (DP) encapsulation per Domain
  - Multicast / Ingress Replication

VXLAN Encapsulation

DCI Encapsulation

Inter-Fabric Connectivity (Option 3 / Option 4)



- Multiple BGP-EVPN Control-Plane Domains
- Integrated Hand-Off with Data-Plane separation
  - Option 3 – L3 DCI
    - L3-LISP, MPLS, EVPN
  - Option 4 – L2 DCI
    - OTV, L2-LISP, EVPN
  - Separate Data-Plane (DP) encapsulation per Domain
    - Multicast / Ingress Replication

VXLAN Encapsulation

DCI Encapsulation

Inter-Fabric Connectivity

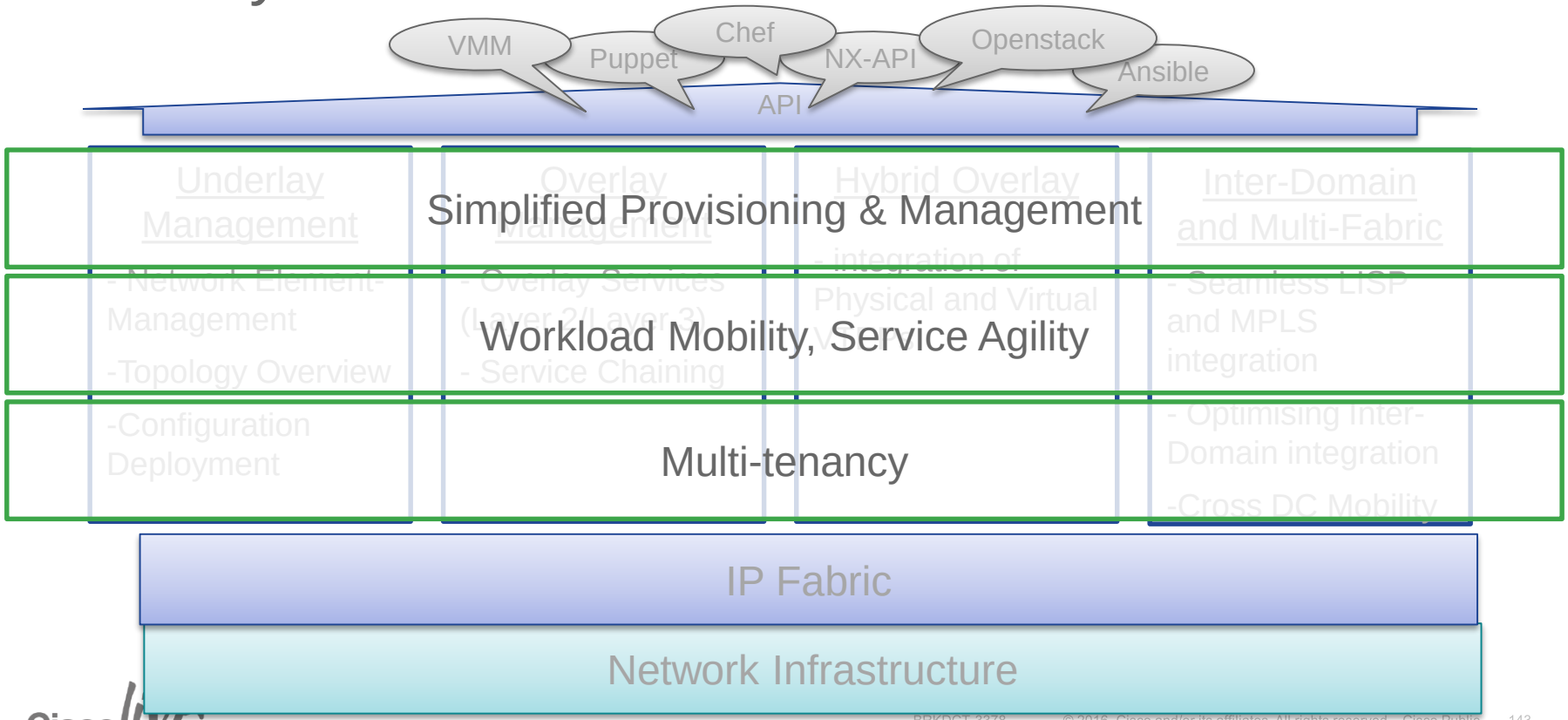
	Option 1	Option 2	Option 3/4
Underlay Control Plane	Unified Underlay Domain	Separated Underlay Domains	Separated Underlay Domains
Overlay Control Plane	Separated Overlay Control-Plane Domains		
Overlay Data Plane	Single Data-Plane	Separated Data-Planes	Separated Data-Planes
BUM Replication in DCI	Unified Underlay Domain (All Multicast or All Ingress Replication)	Dependency on DCI	Choice (Unicast/Multicast)
ARP Flood Suppression (DCI)	yes	yes	yes
Unknown Unicast Flood Suppression (DCI)	no	yes	yes
Broadcast Suppression/Limit (DCI)	no	yes	yes
Layer-2 Loop Prevention	Loop mitigation (Edge Protection)	VPC at Border	Loop mitigation (At DCI)

Fabric Management & Automation

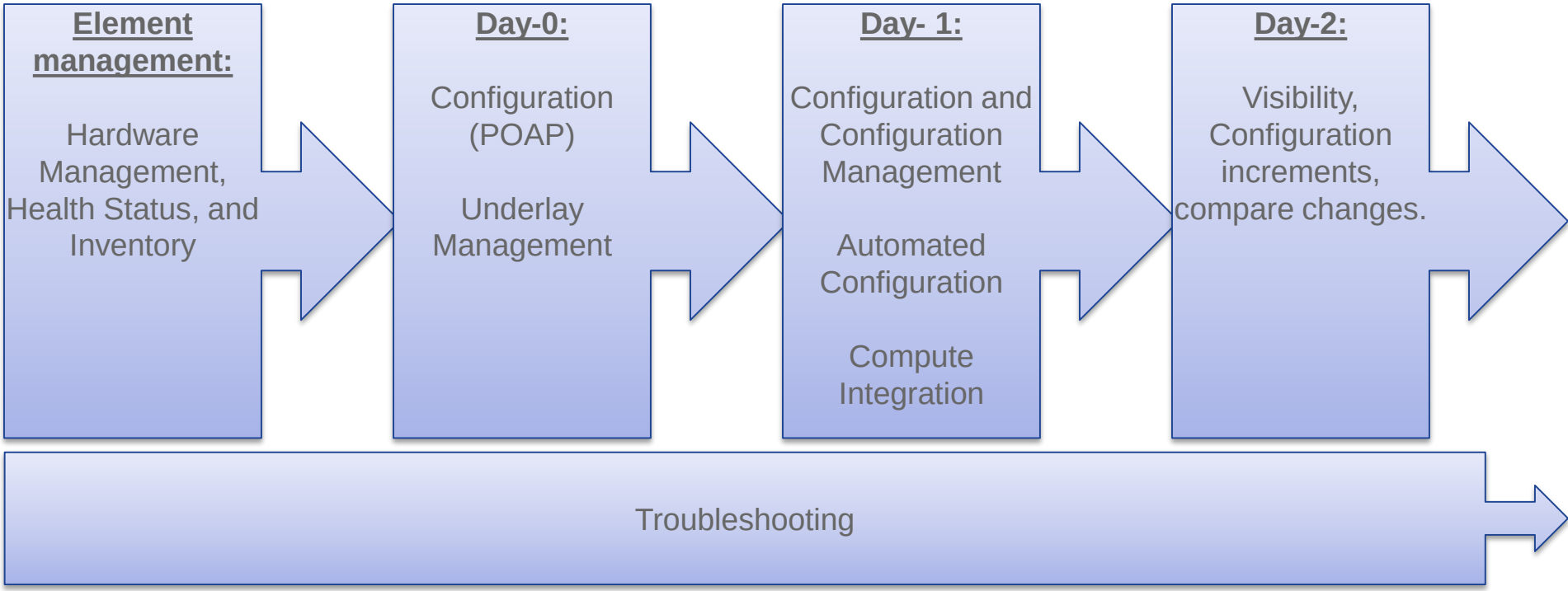
How to Achieve Data Centre Automation

- Simplify
  - Do not start with the most difficult task (low hanging Fruits)
- Standardise
  - 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

Anatomy of Data Centre Automation



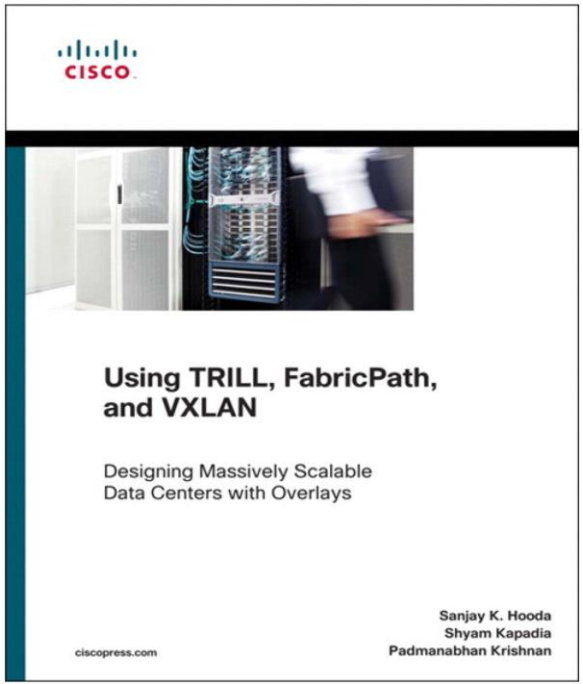
Fabric Management & Operations



Device Auto-Configuration (POAP)
Day 0, Day 0.5 and Day 1

1. Easy way to unbox, rack the device, and not enter any base CLI configuration. Just rack, power, and plug into the management network.
2. Provides a standard and consistent configuration across of the data centre network devices.
3. Provides a standard and consistent images to deploy to all of the data centre devices.

Recommended Reading



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

- Sanjay K. Hooda
- Shyam Kapadia
- Padmanabhan Krishnan

ISBN-10: 1-58714-393-3

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

Recommended Viewing

Cisco Programmable Fabric Using VXLAN with BGP EVPN LiveLessons

- David Jansen
- Lukas Krattiger

ISBN-10: 0-13-427229-3

ISBN-13: 978-0-13-427229-0

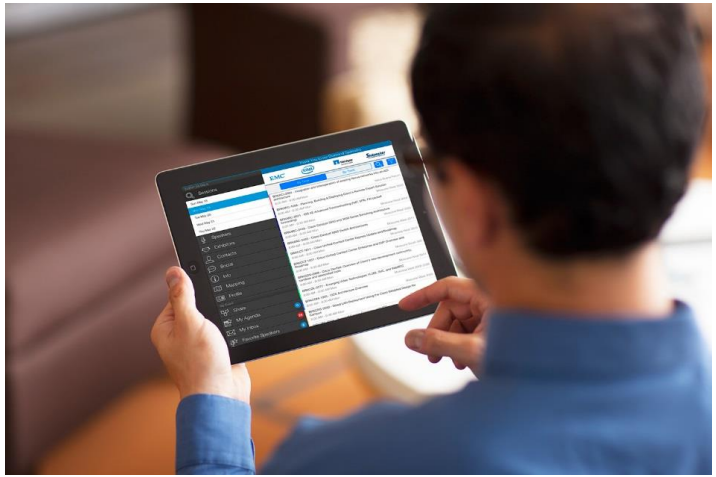
Q & A

Complete Your Online Session Evaluation

Give us your feedback and receive a **Cisco 2016 T-Shirt** by completing the Overall Event Survey and 5 Session Evaluations.

- Directly from your mobile device on the Cisco Live Mobile App
- By visiting the Cisco Live Mobile Site http://showcase.genie-connect.com/ciscolivemelbourne2016/
- Visit any Cisco Live Internet Station located throughout the venue

T-Shirts can be collected Friday 11 March at Registration



Learn online with Cisco Live!
Visit us online after the conference for full access to session videos and presentations.
www.CiscoLiveAPAC.com

Thank you


CISCO