Introduction to VXLAN

Maw Khant Lwin
CCIE#60007
CCIE,PCNSE,RHCSA,ITIL,MCSE
Senior Network Engineer (Team Lead)
One Cloud Technology
mawkhant.lwin@1cloudtechnology.com

Introduction to VXLAN

Agenda

- Why VXLAN
- Terminologies (RFC-7348)
- How it Works
- Current Challenges
- Frame Format
- Benefits of using VXLAN
- Multi-Tenancy
- Fabrics with Overlays Management
- Use-cases
- Network Automation with VXLAN

Why VXLAN

- Traditional VLAN (4096 VLANS)
- A Physical Server can have multiple Virtual machines with its own MAC
- STP blocks redundant links
- Virtualization Challenges

Why VXLAN

Traditional VLAN (4096 VLANS)

- Allowing network administrators to apply additional security to network communication
- Making expansion and relocation of a network or a network device easier
- Providing flexibility because administrators are able to configure in a centralized environment while the devices might be located in different geographical locations
- Decreasing the latency and traffic load on the network and the network devices, offering increased performance

Why VXLAN

Traditional VLAN (4096 VLANS)

VLANs also have some disadvantages and limitations as listed below:

- High risk of virus issues because one infected system may spread a virus through the whole logical network

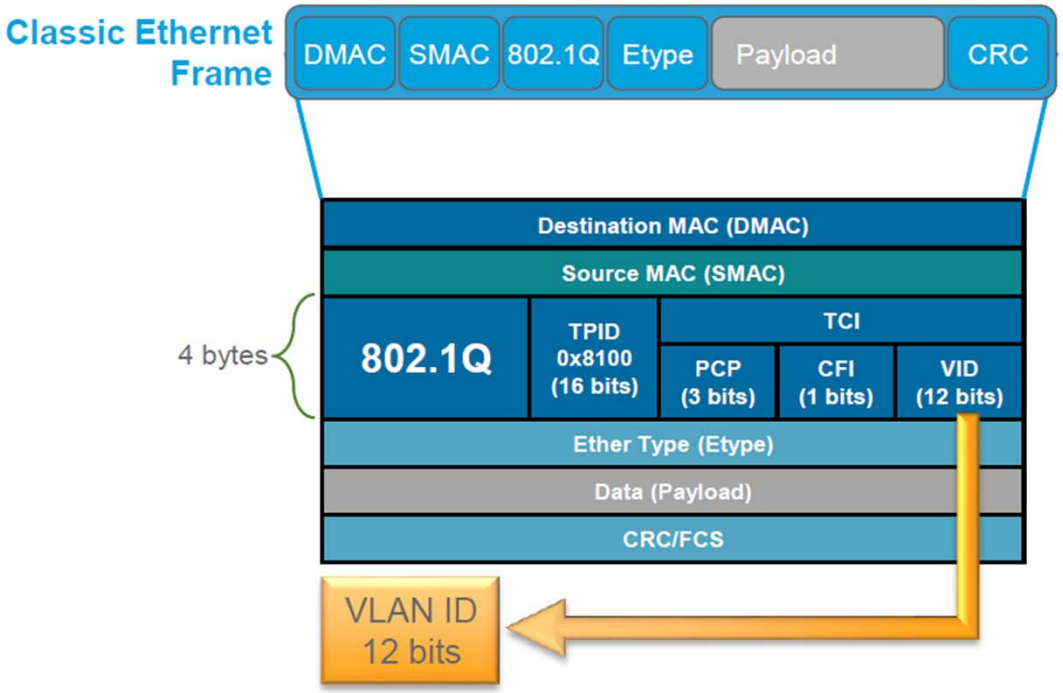
- Equipment limitations in very large networks because additional routers might be needed to control the workload

- More effective at controlling latency than a WAN, but less efficient than a LAN

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

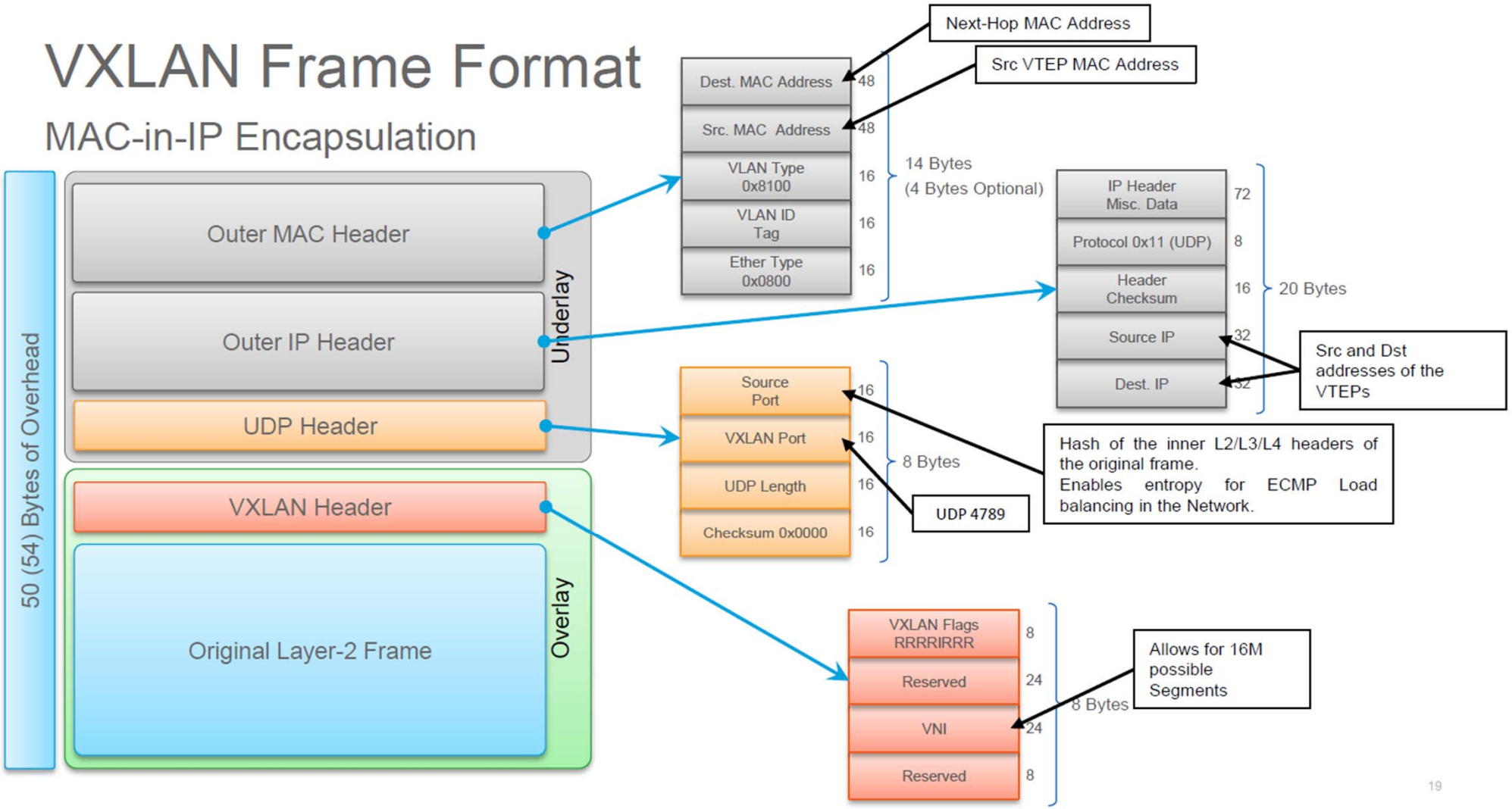


- Traditional VLAN (4096 VLANS)

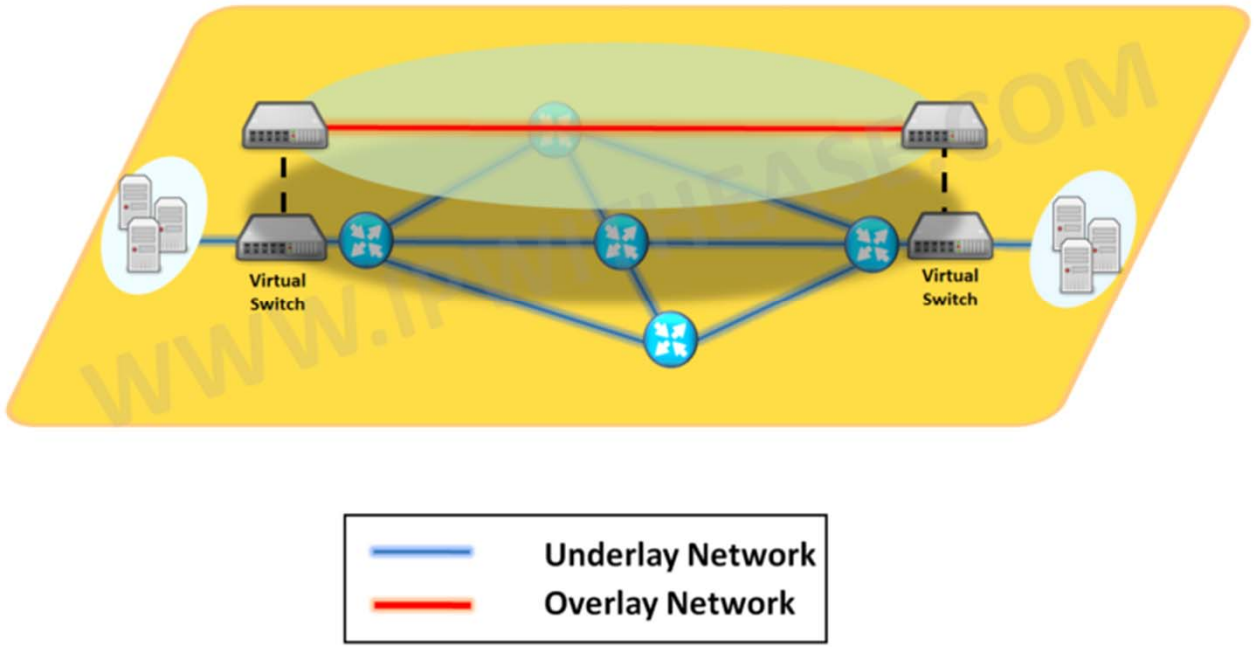
TPID = Tag Protocol Identifier, TCI = Tag Control Information, PCP = Priority Code Point, CFI = Canonical Format Indicator, VID = VLAN Identifier

VXLAN Frame Format

MAC-in-IP Encapsulation

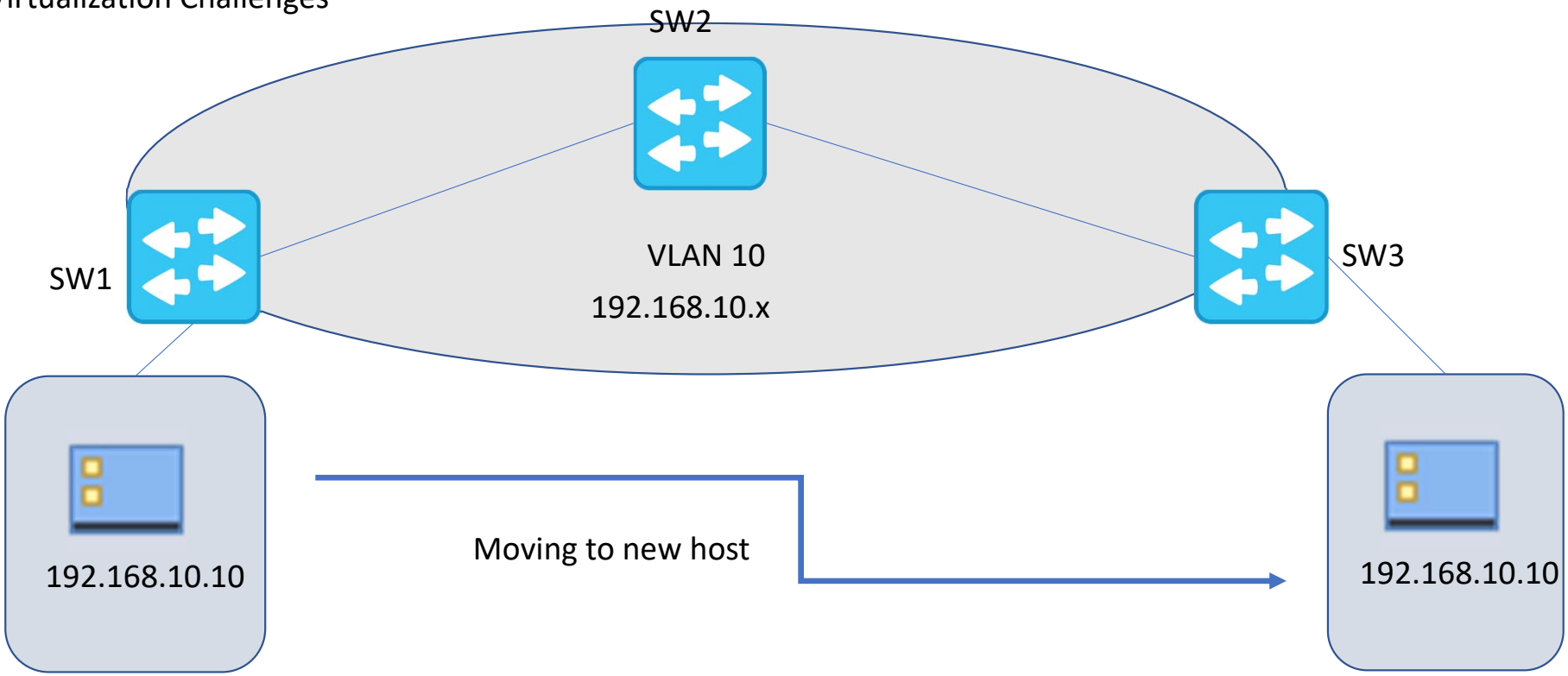


Overlay and Underlay



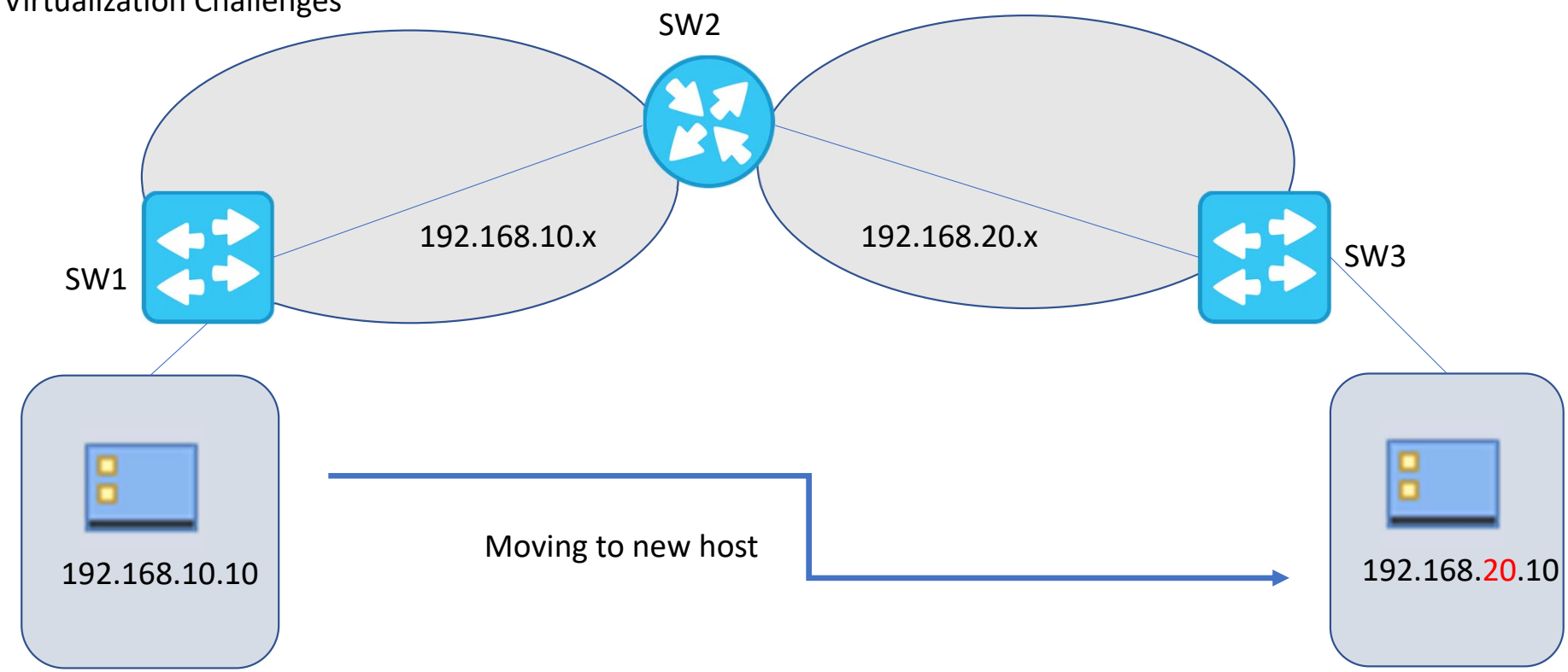
Why VXLAN

Virtualization Challenges



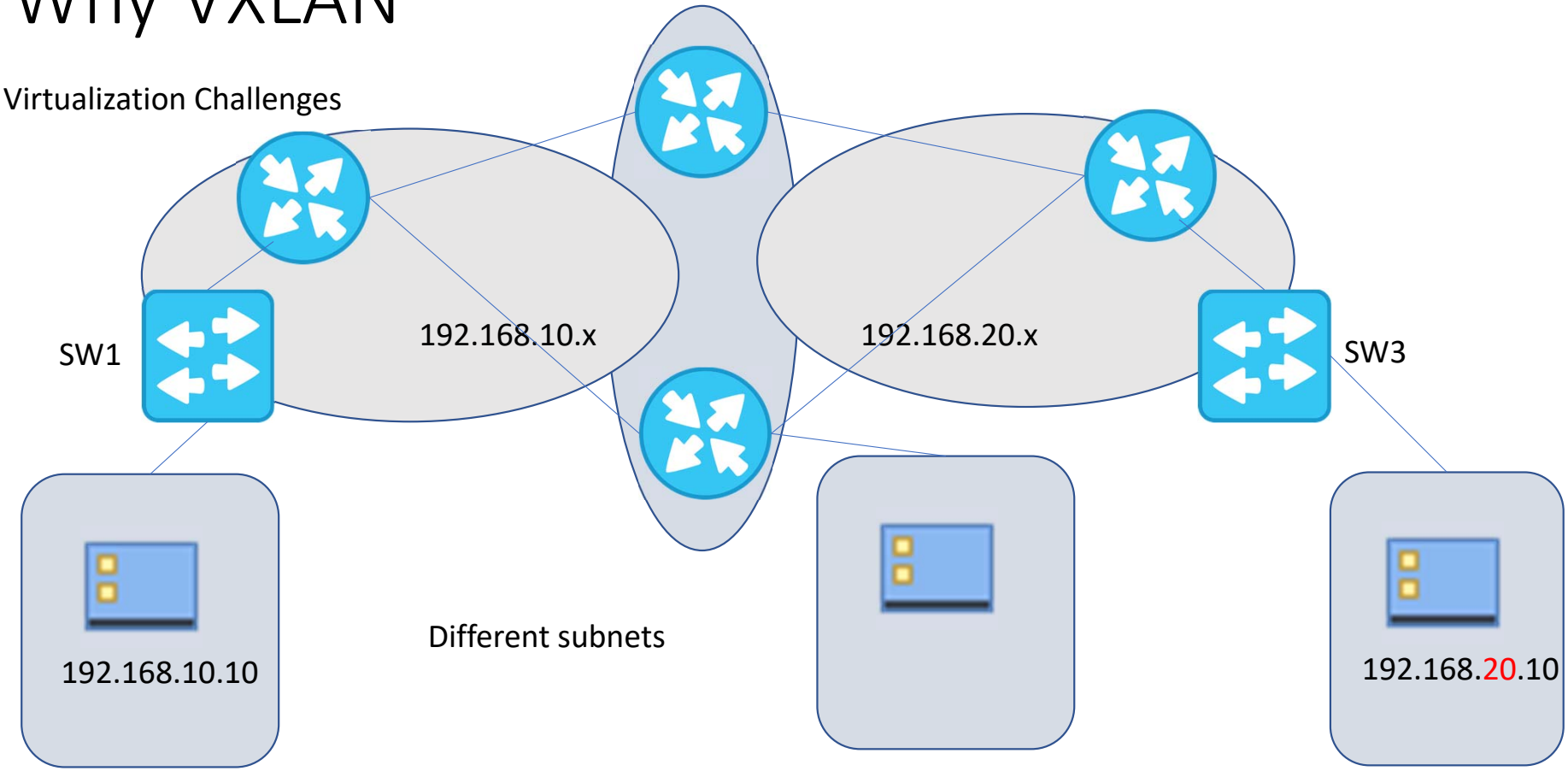
Why VXLAN

Virtualization Challenges



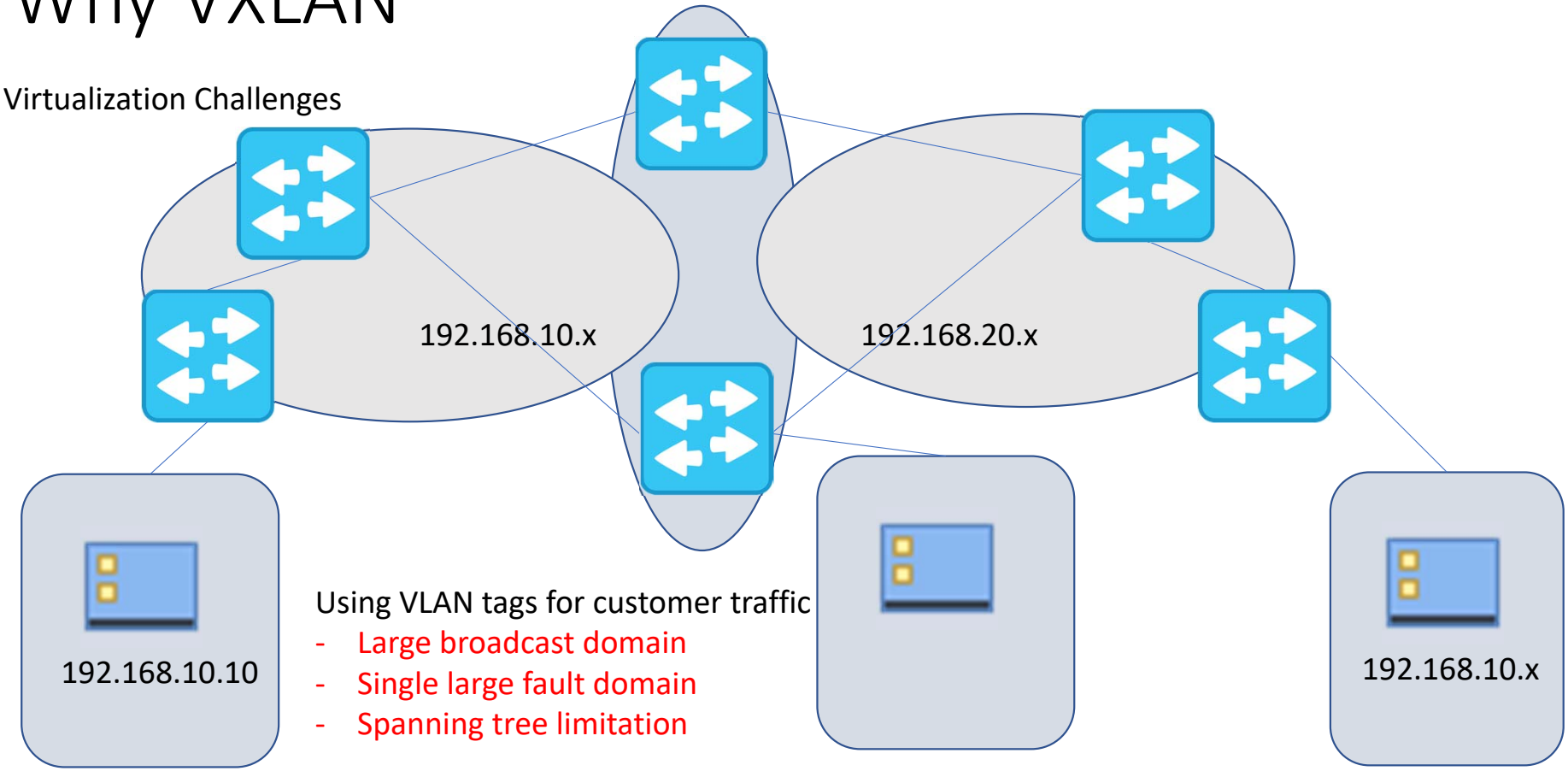
Why VXLAN

Virtualization Challenges



Why VXLAN

Virtualization Challenges

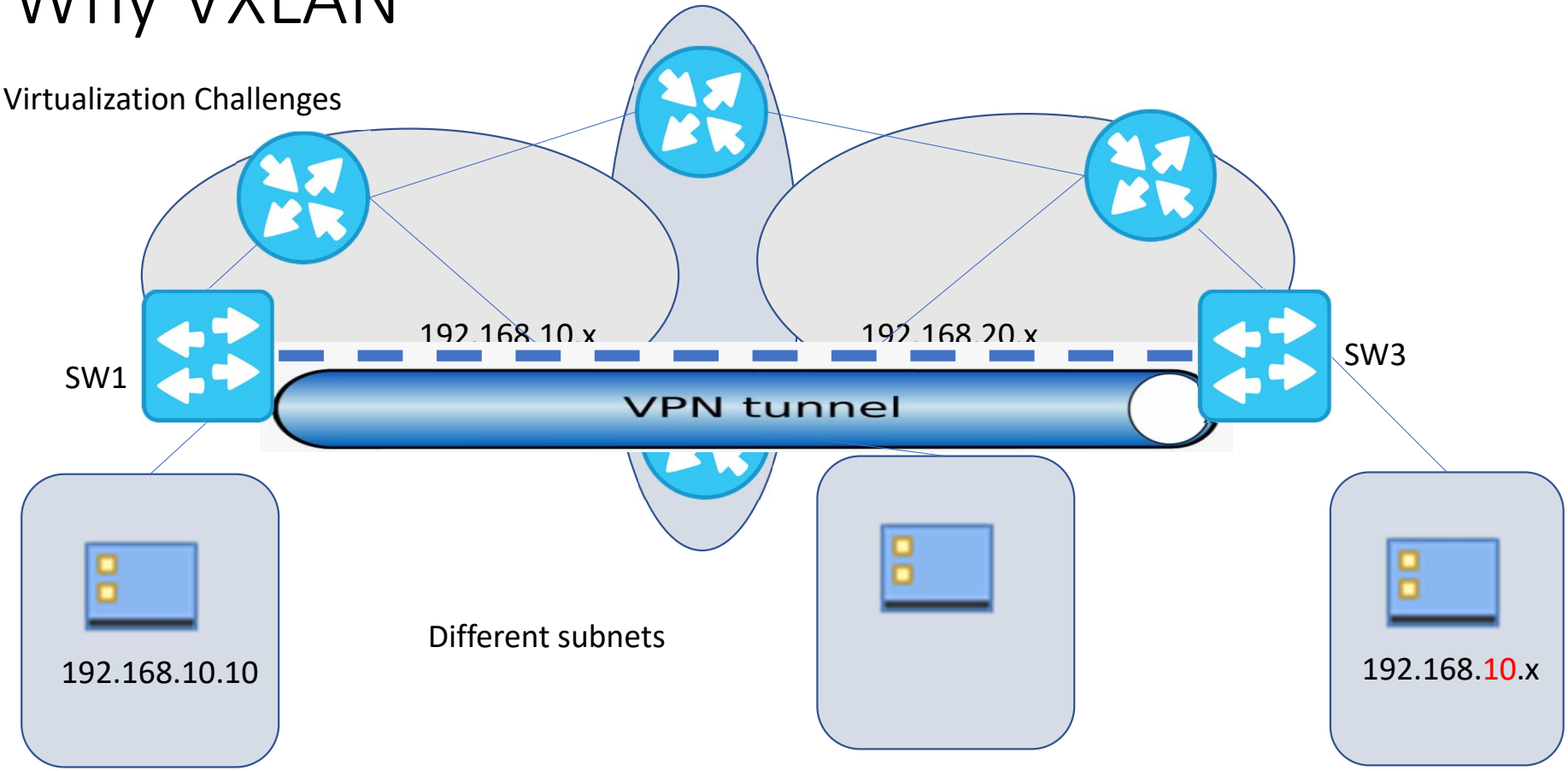


Using VLAN tags for customer traffic

- Large broadcast domain
- Single large fault domain
- Spanning tree limitation

Why VXLAN

Virtualization Challenges

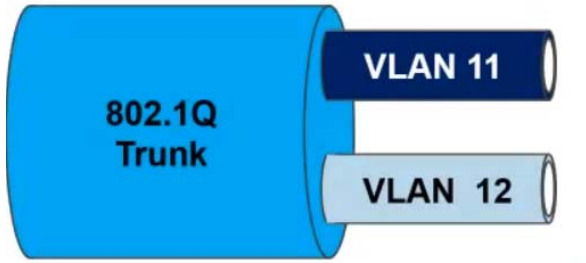


VXLAN Terminology

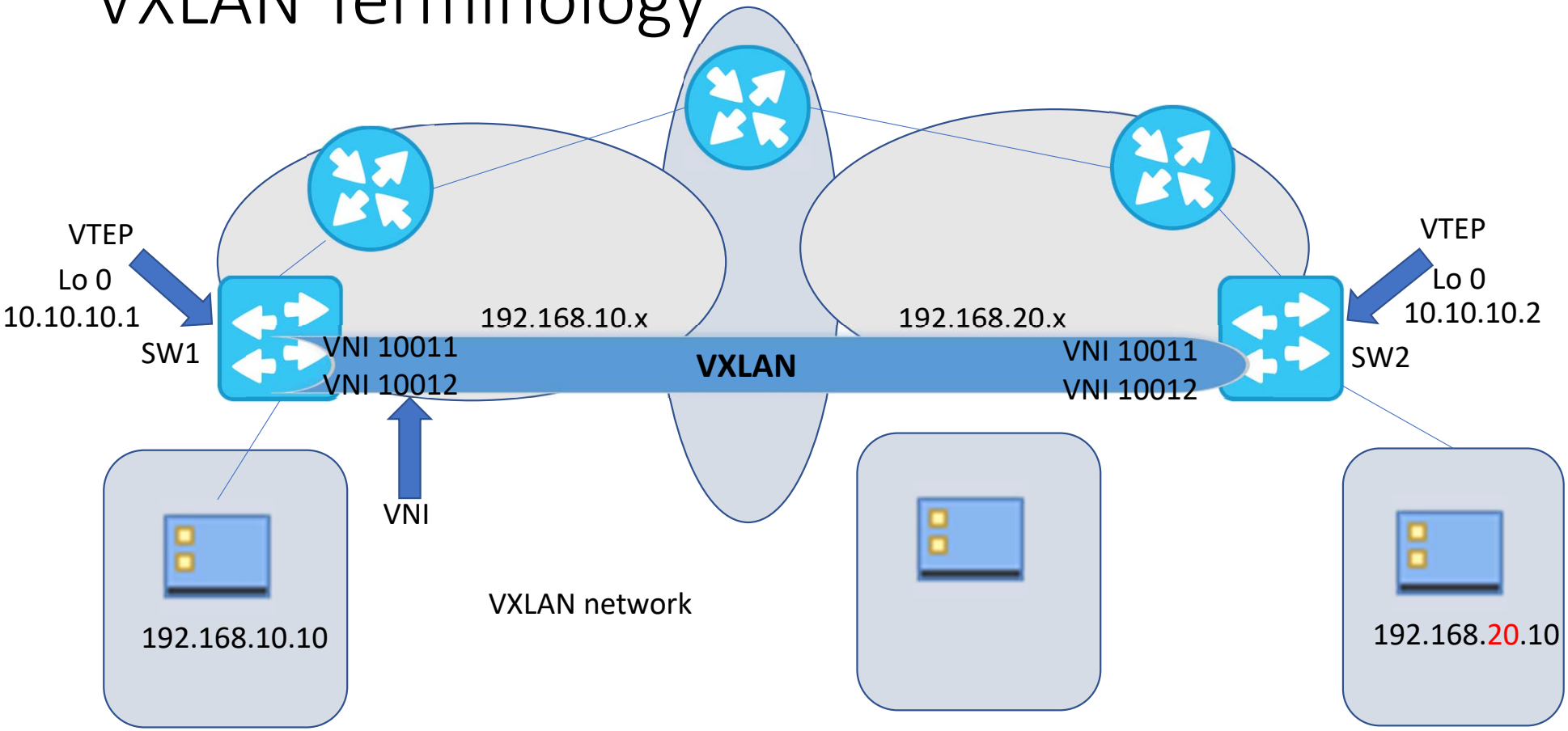
1. Virtual Tunnel End-point (**VTEP**).
- The VTEP acts as the entry point for connecting hosts into the VXLAN overlay network.
- The task of the VTEP is to encap/decap with the appropriate VXLAN header.
- The VTEP component can reside either a software virtual switch or a physical switch.
2. Virtual Tunnel Identifier (**VTI**)
- An IP interface used as the Source IP address for the encapsulated VXLAN traffic
3. Virtual Network Identifier (**VNI**)
- A 24-bit field added within the VXLAN header.
- Identifies the Layer 2 segment of the encapsulated Ethernet frame
4. VXLAN Header
- The IP/UDP and VXLAN header added by the VTEP
- The SRC UDP port of the header is a hash of the inner frame to create entropy for ECMP

VXLAN Terminology

802.1Q Frame Format

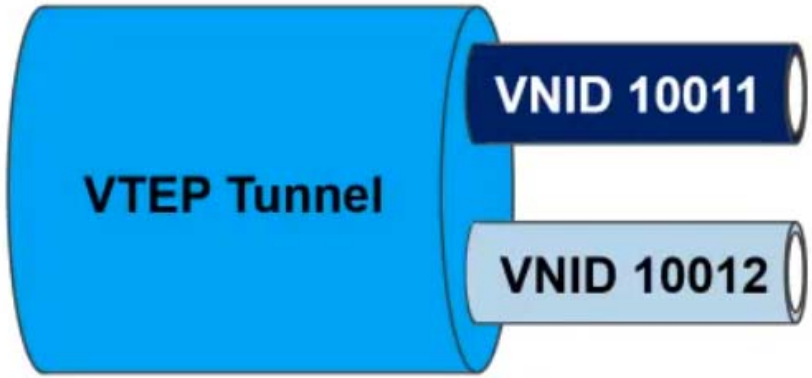


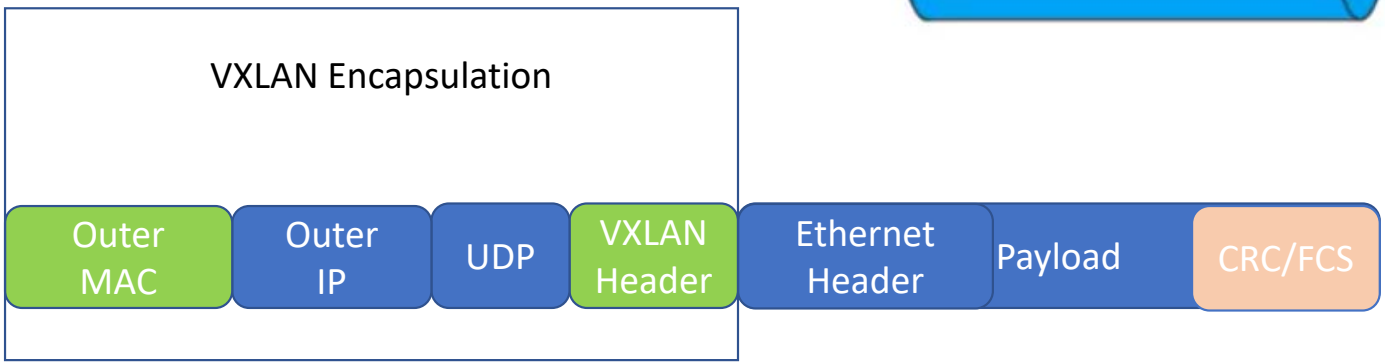
VXLAN Terminology



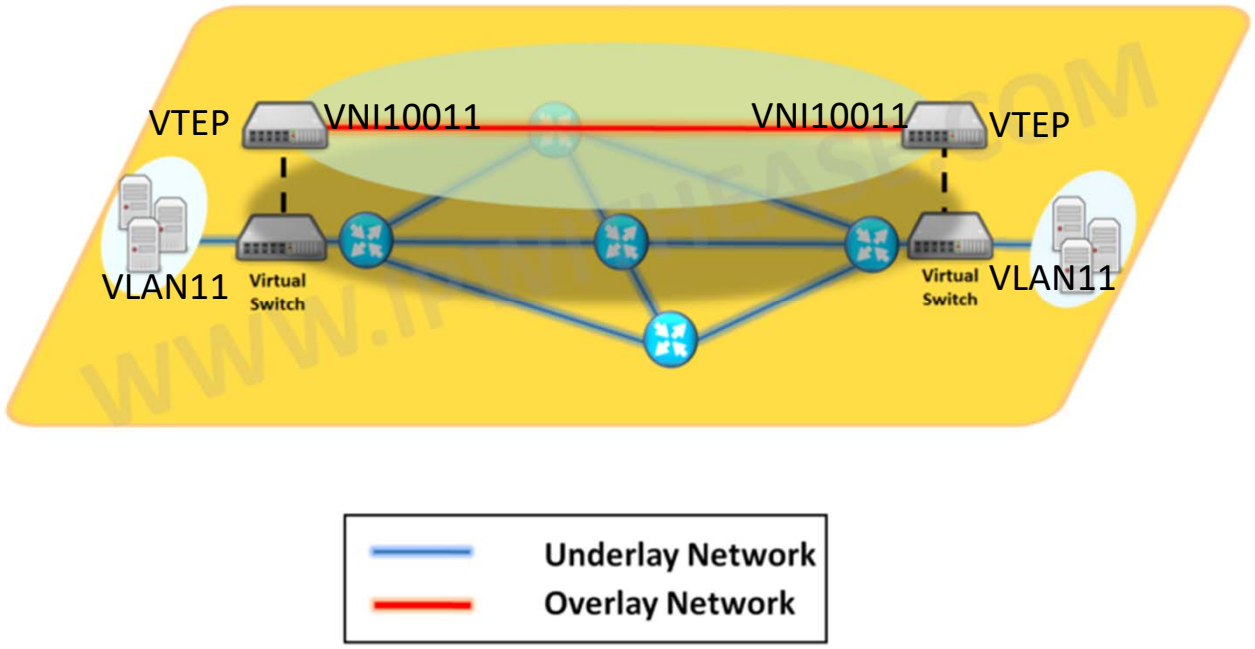
VXLAN Terminology

VXLAN frame Headers

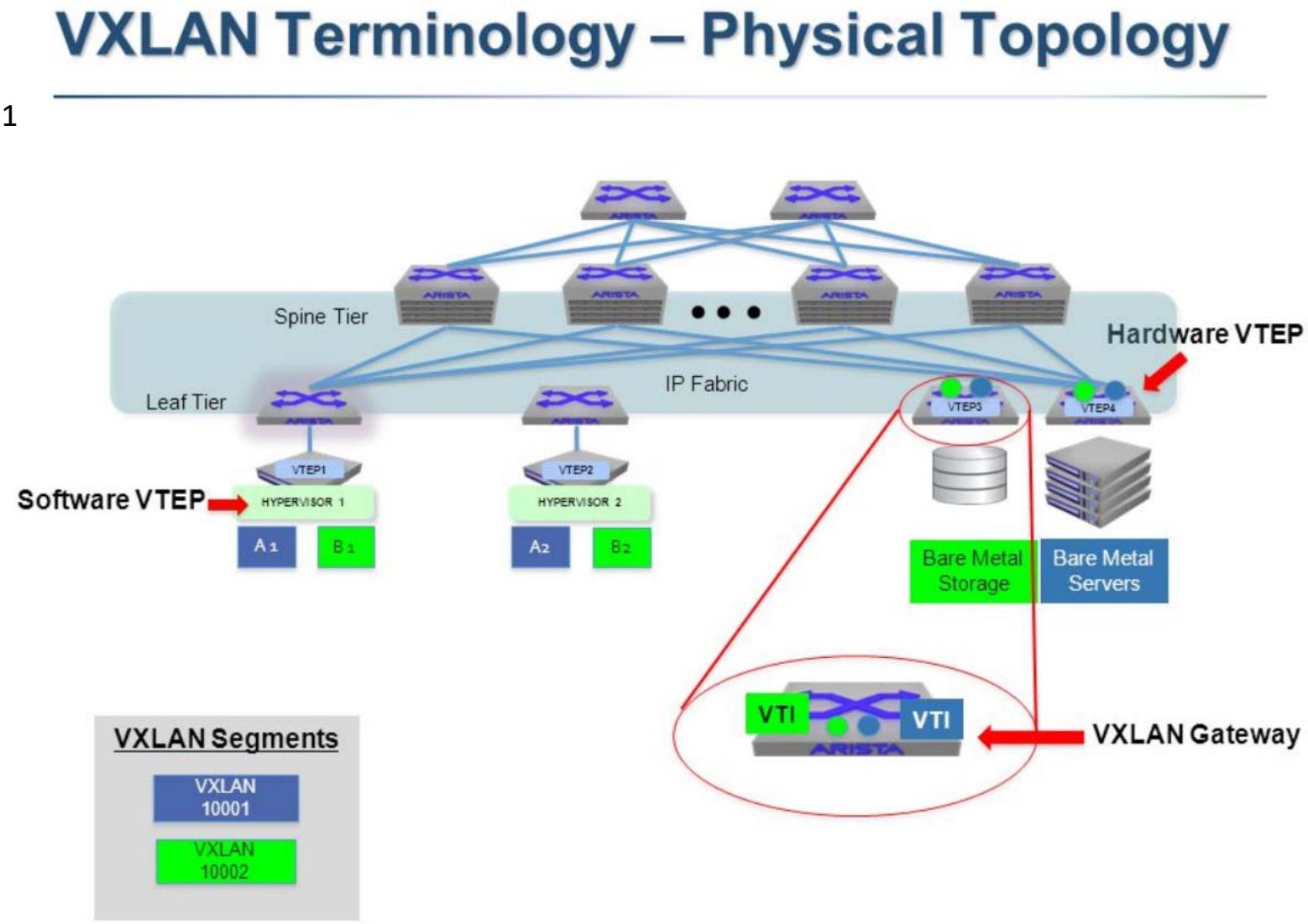




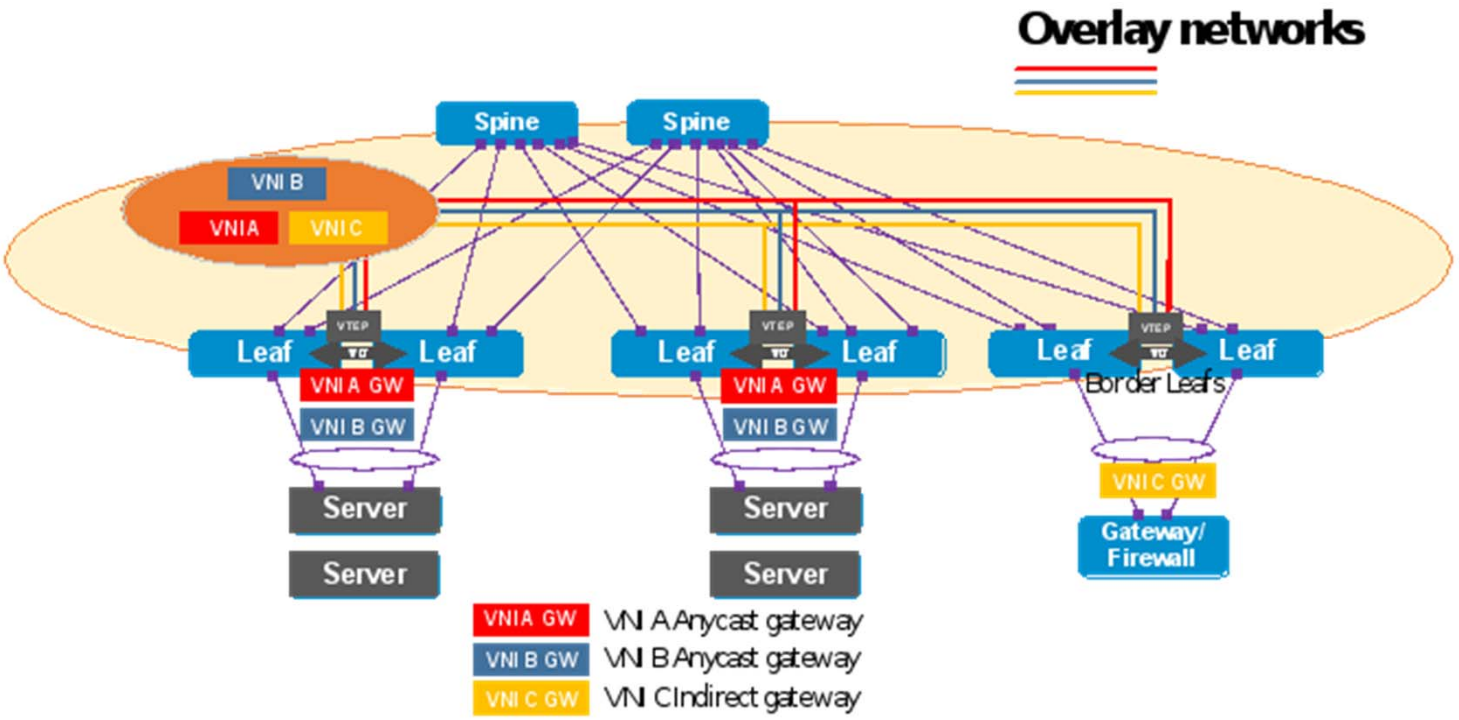
Overlay and Underlay



Example Diagram - 1



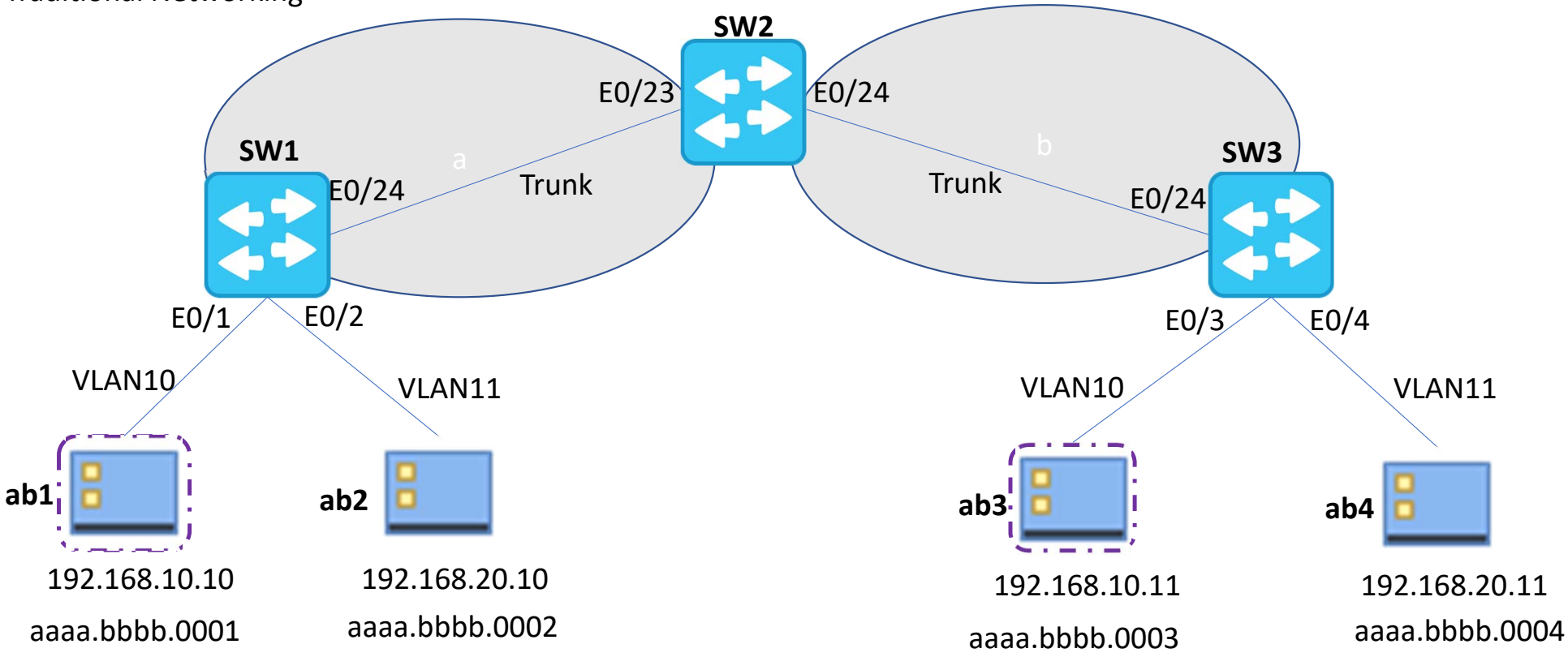
Example Diagram - 2



How VXLAN Works

1. **ab1** wants to send traffic to **ab3**

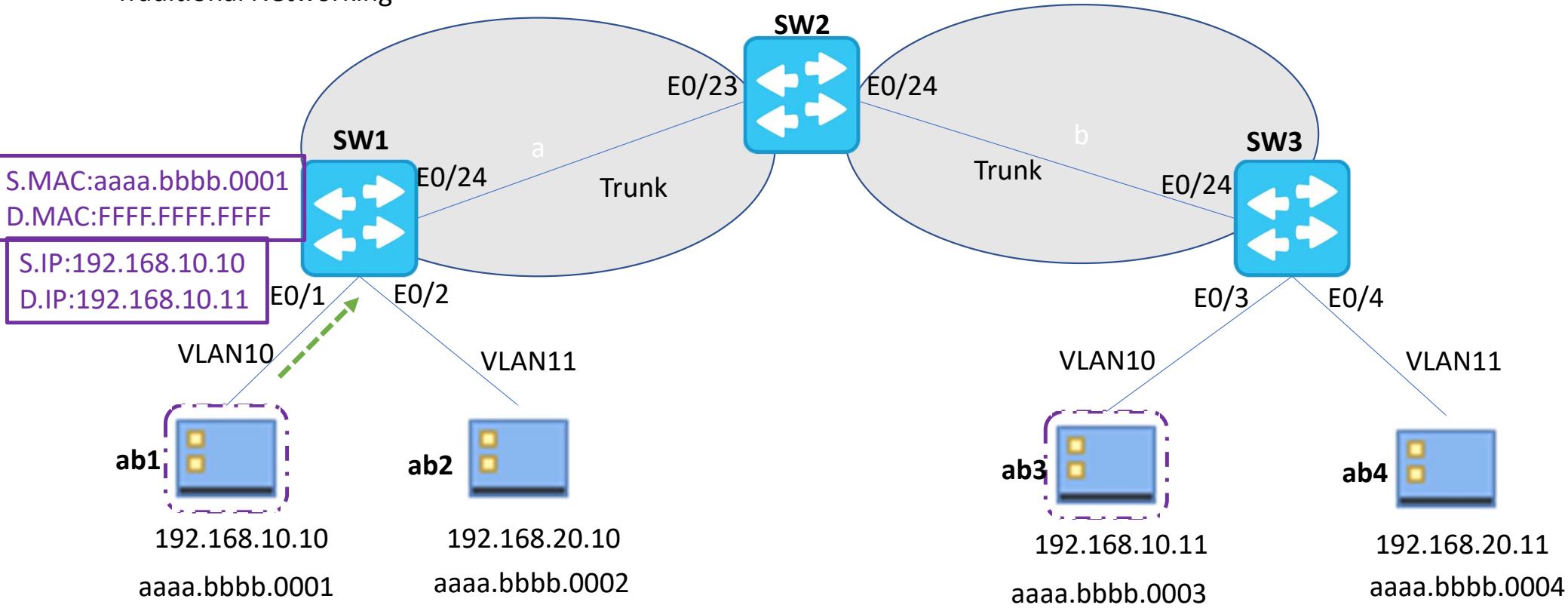
Traditional Networking



How VXLAN Works

2. ARP request to discover
The MAC of **ab3**.

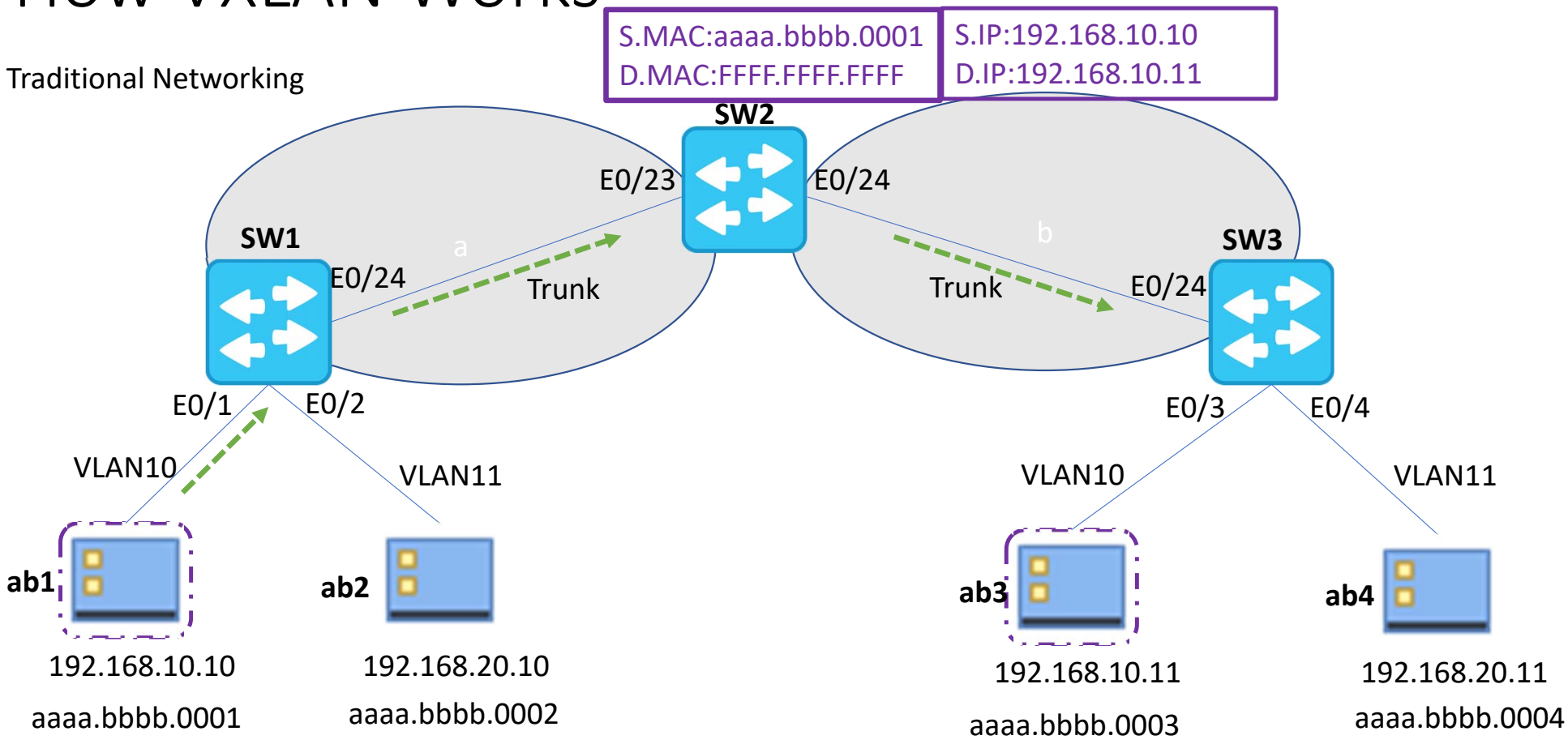
Traditional Networking



How VXLAN Works

Traditional Networking

3. SW1 floods the request out all ports except the one it was received.



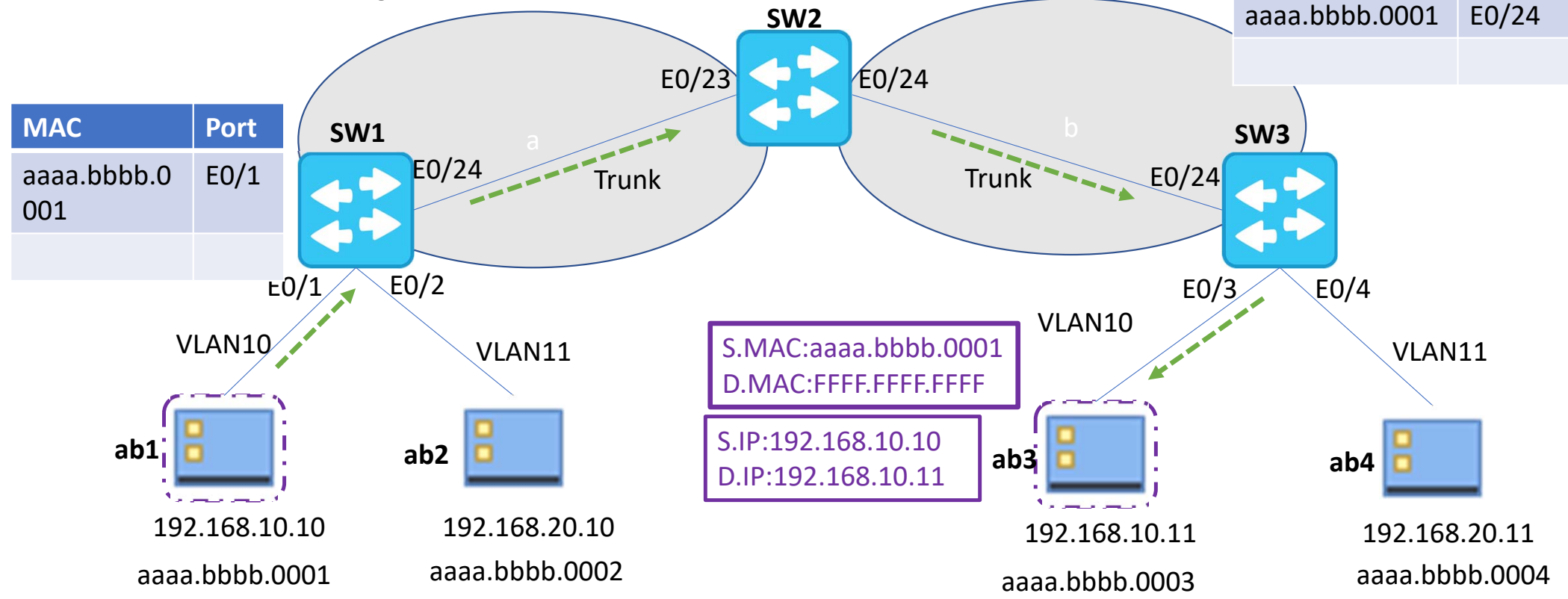
How VXLAN Works

Traditional Networking

4. SW1 updates switching table with **ab1** MAC

MAC	Port
aaaa.bbbb.0001	E0/24

MAC	Port
aaaa.bbbb.0001	E0/1



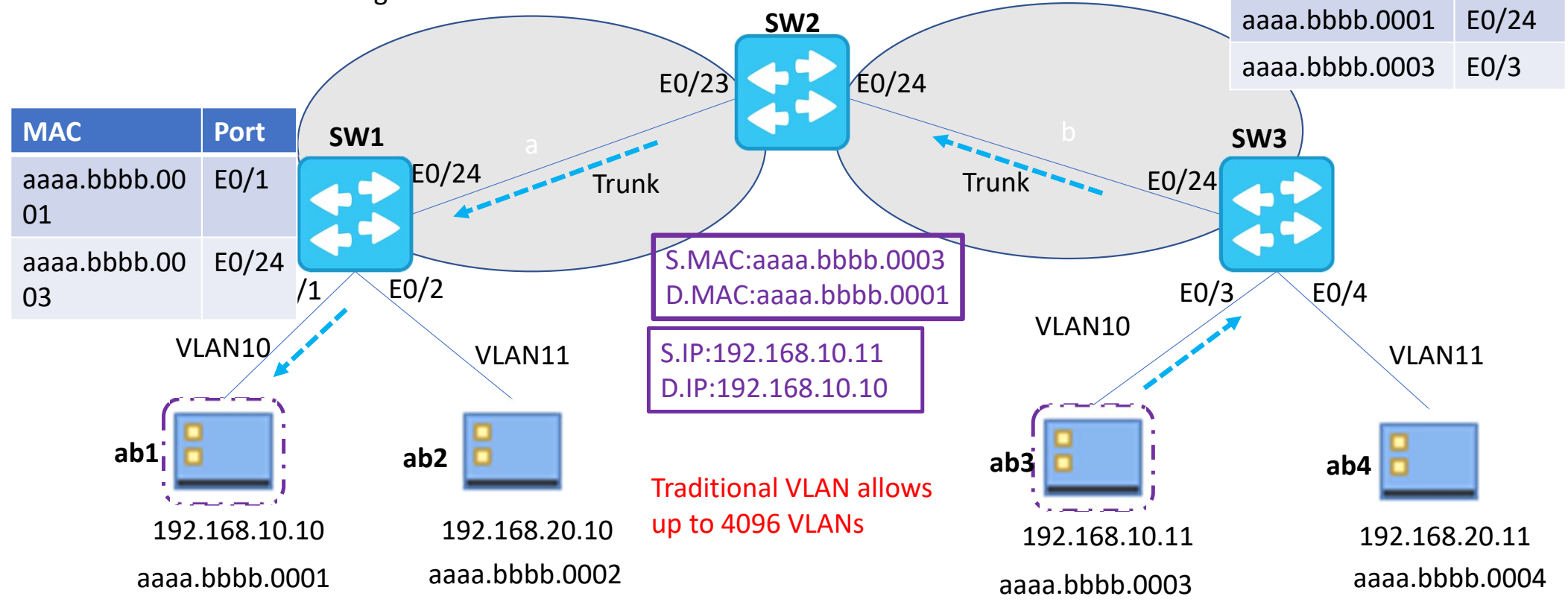
How VXLAN Works

5. SW2 forwards response and record **ab3** MAC

Traditional Networking

MAC	Port
aaaa.bbbb.0001	E0/24
aaaa.bbbb.0003	E0/3

MAC	Port
aaaa.bbbb.0001	E0/1
aaaa.bbbb.0003	E0/24

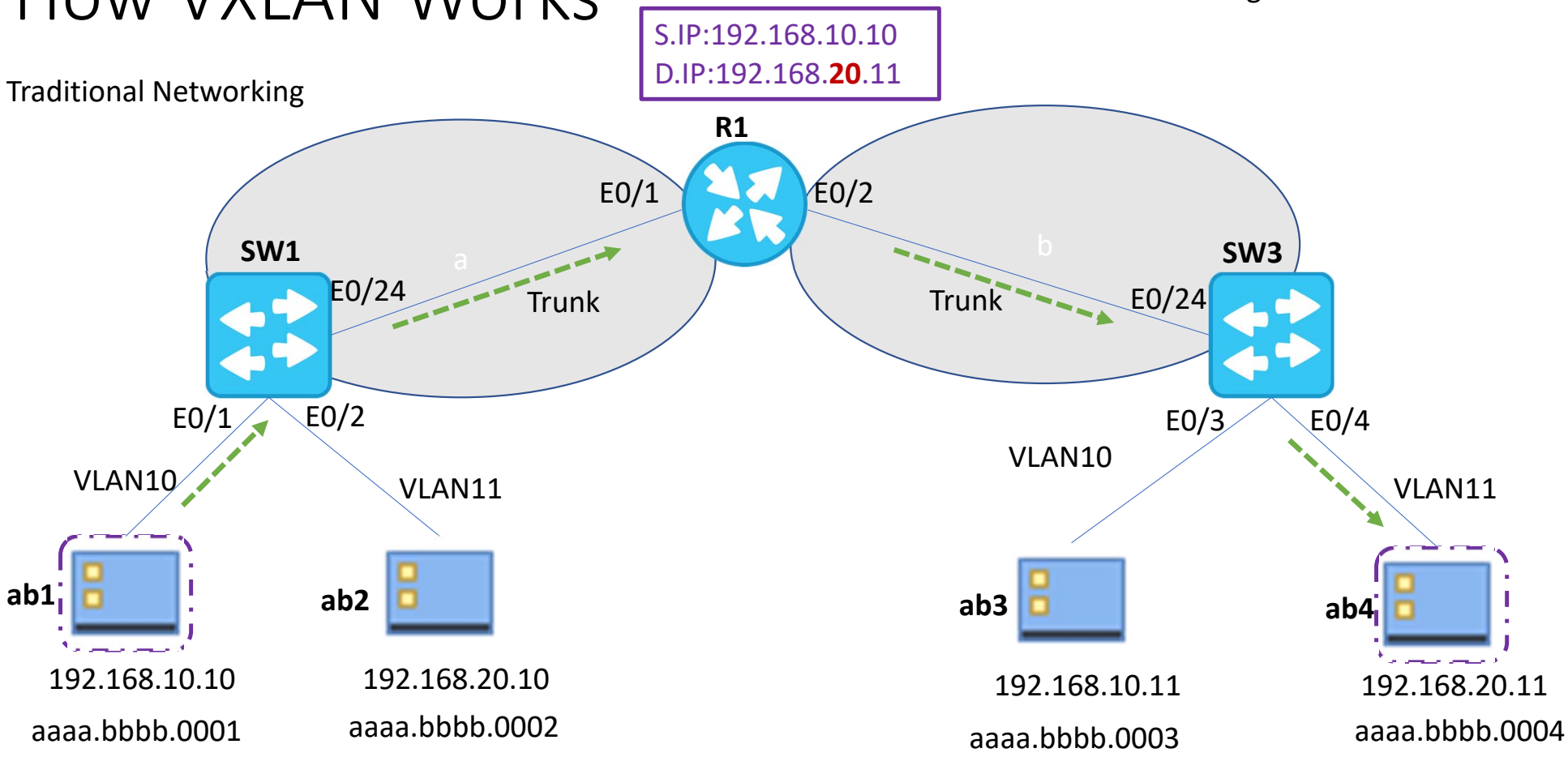


Traditional VLAN allows up to 4096 VLANs

How VXLAN Works

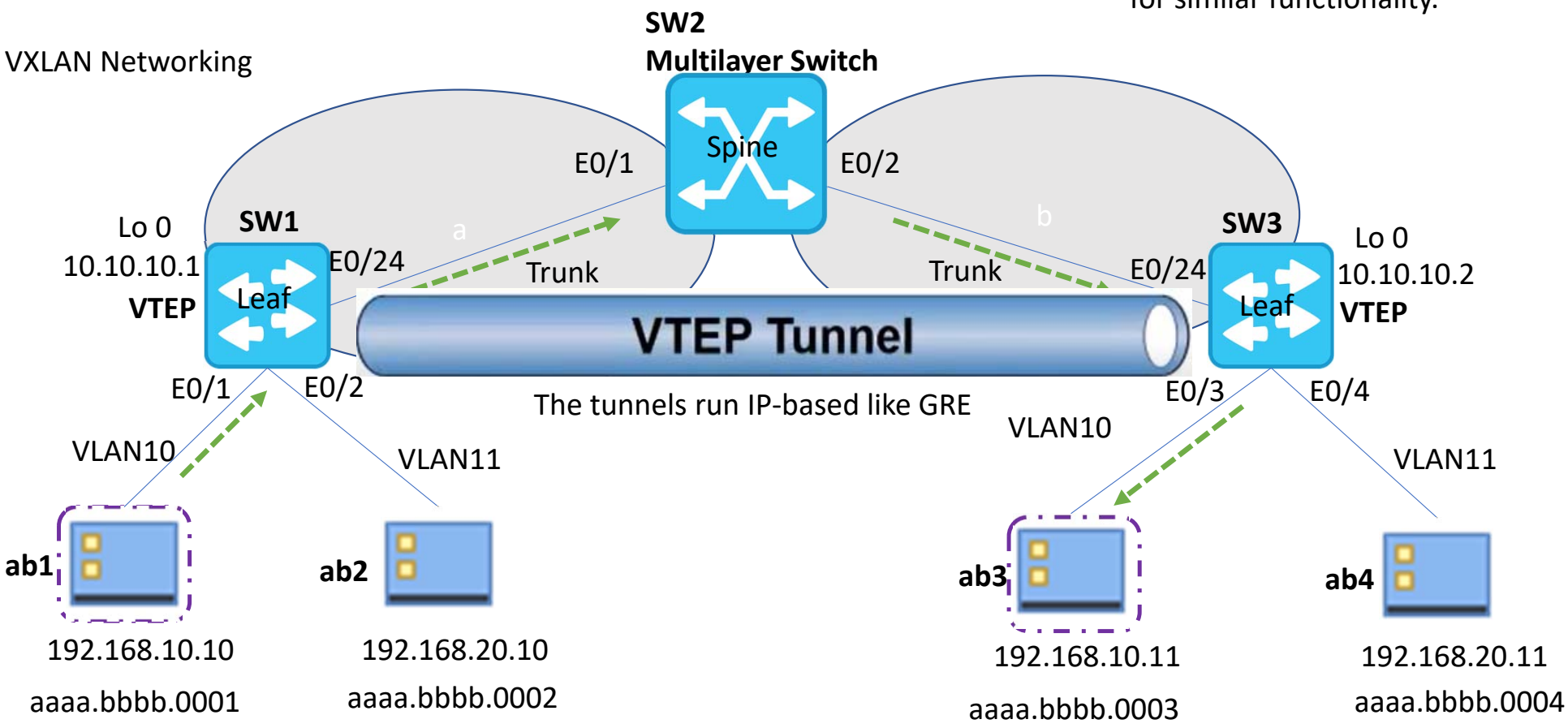
6. From ab1 to **ab4** traffic,
L3 routing is needed

Traditional Networking



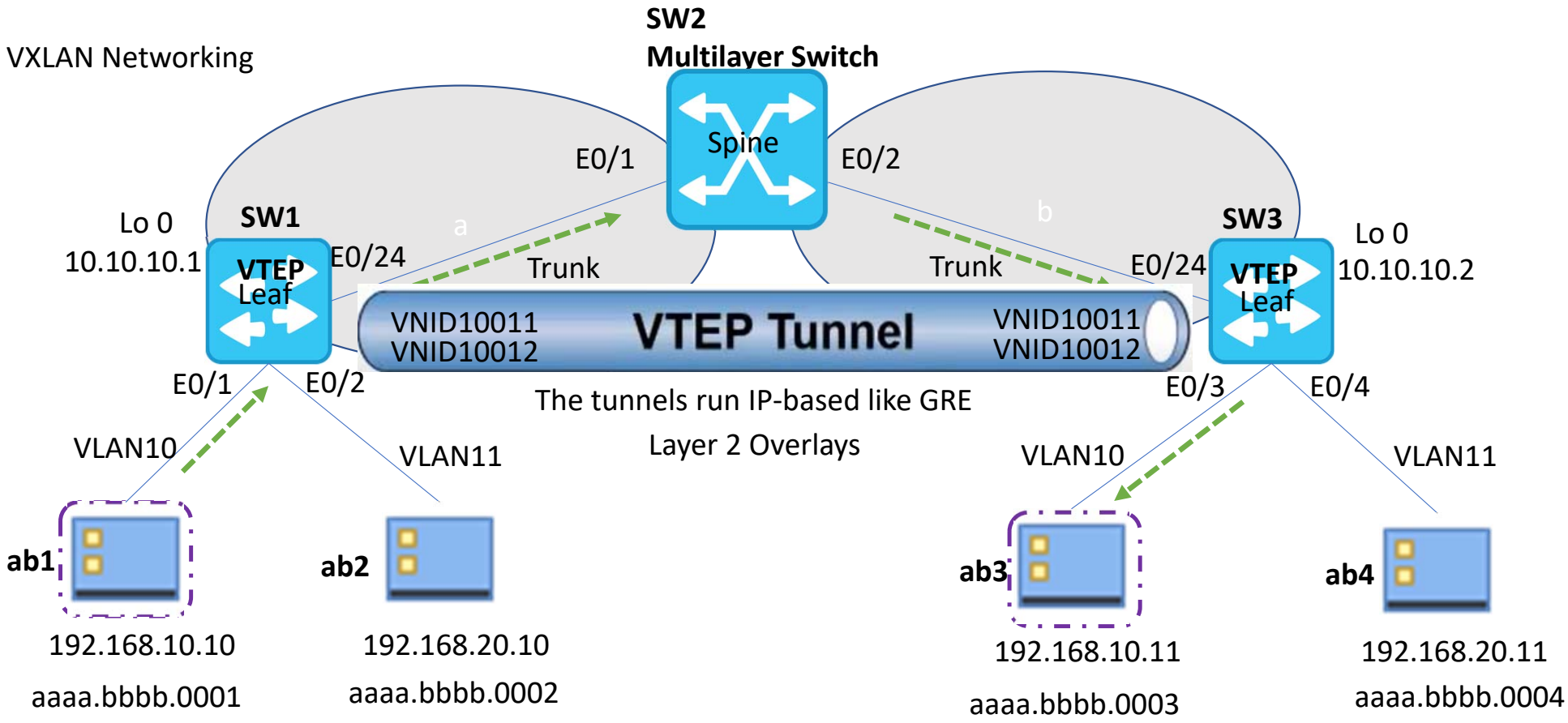
How VXLAN Works

VXLAN replaces directly connected Physical trunk links with tunnels for similar functionality.



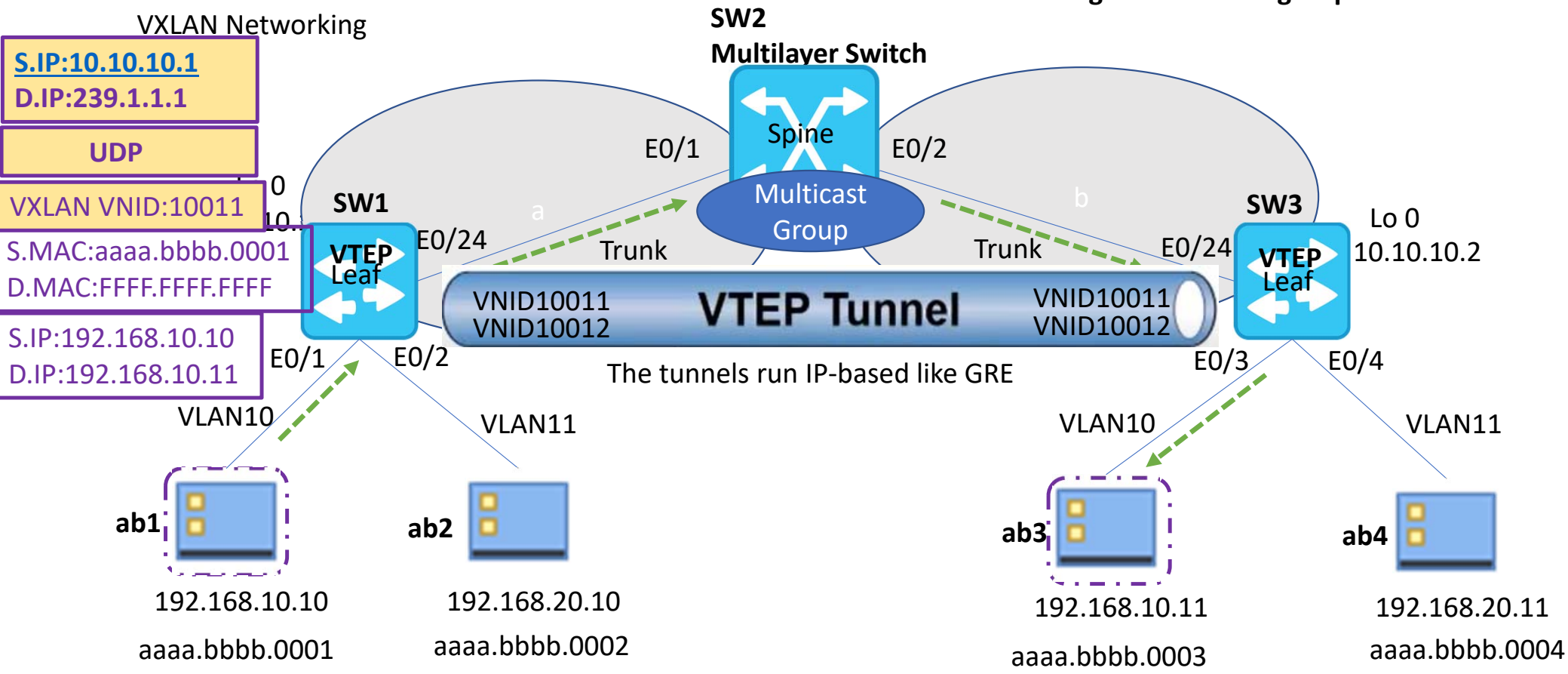
1. ab1 wants to send traffic to ab3

How VXLAN Works



How VXLAN Works

2. SW1 receives ARP request, adds VXLAN and tunnel headers and forwards to the assigned multicast group

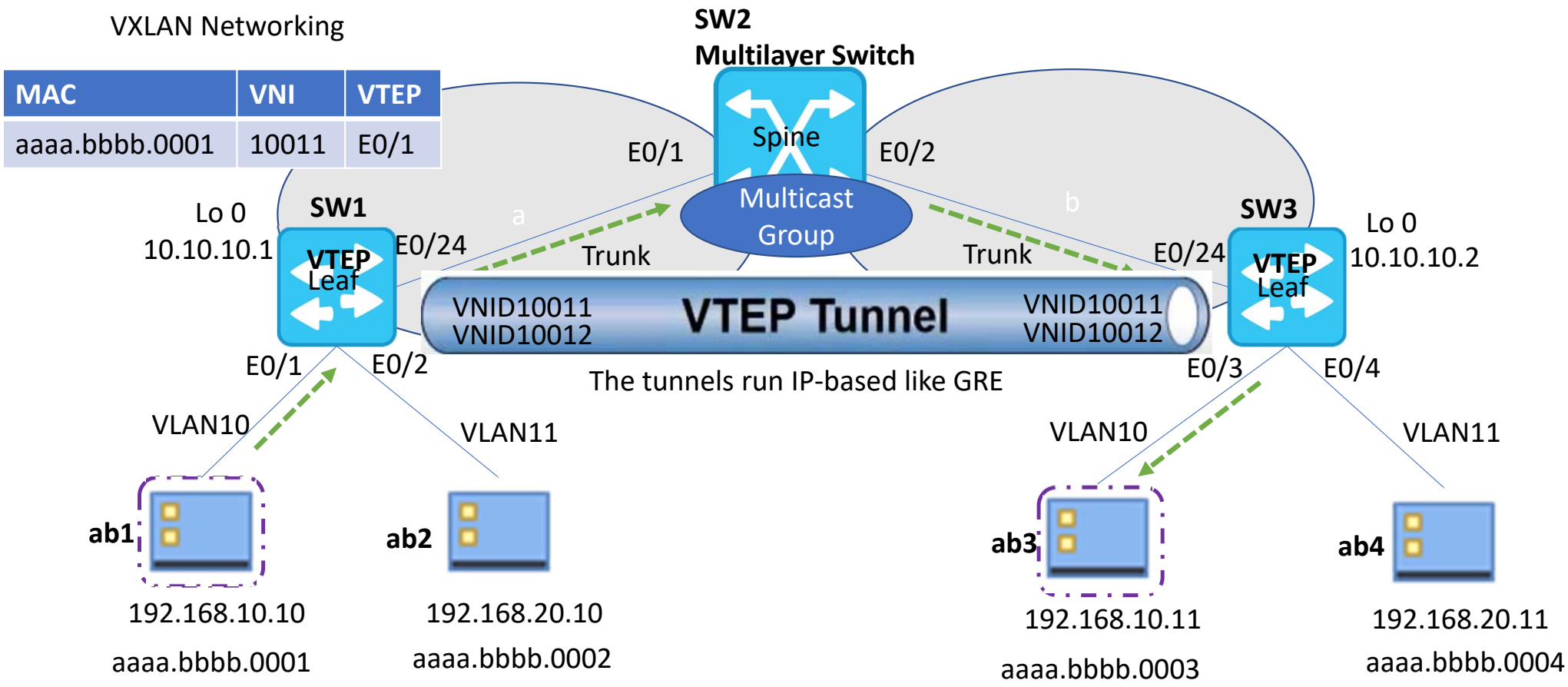


How VXLAN Works

3. Multicast group replicates the packet to all VTEPs that are part of Layer 2 VNID

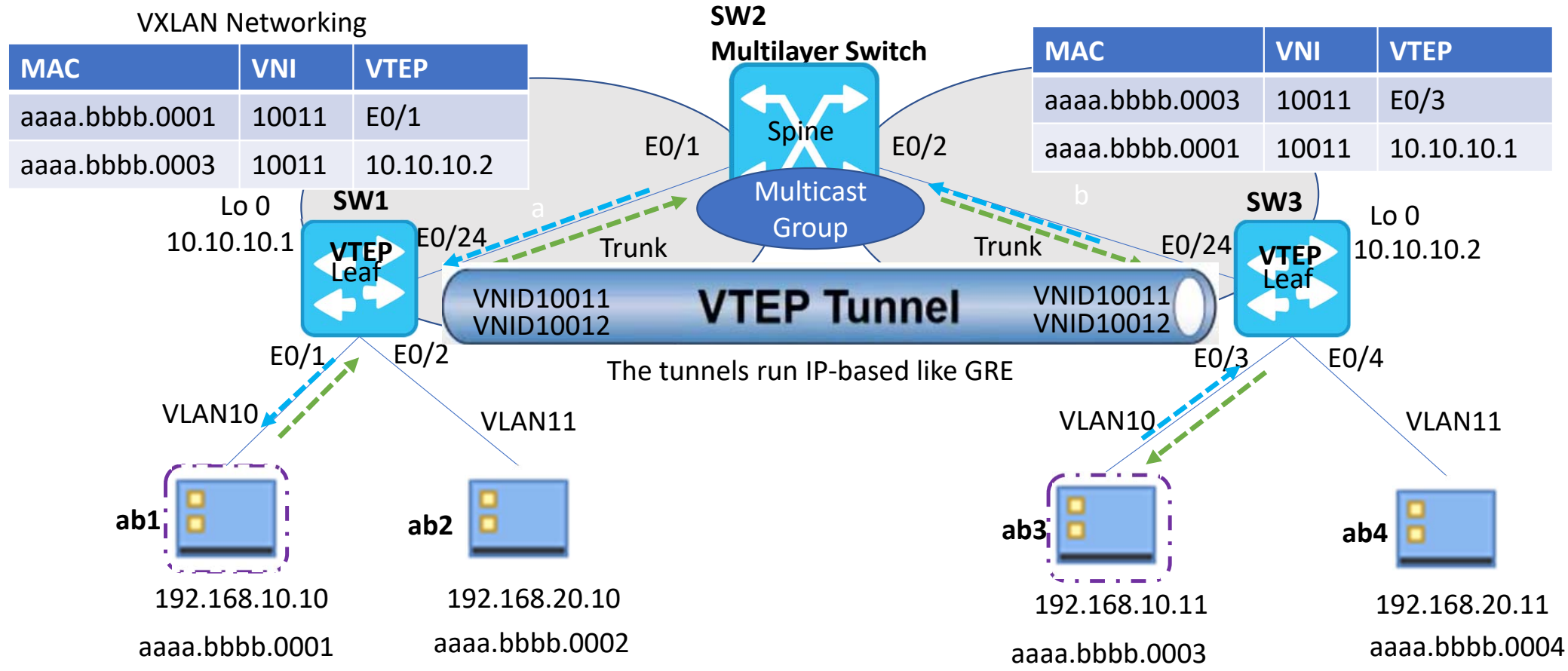
VXLAN Networking

MAC	VNI	VTEP
aaaa.bbbb.0001	10011	E0/1



How VXLAN Works

4. Now all traffic between ab1 and ab3 will flow directly through overlay tunnel.

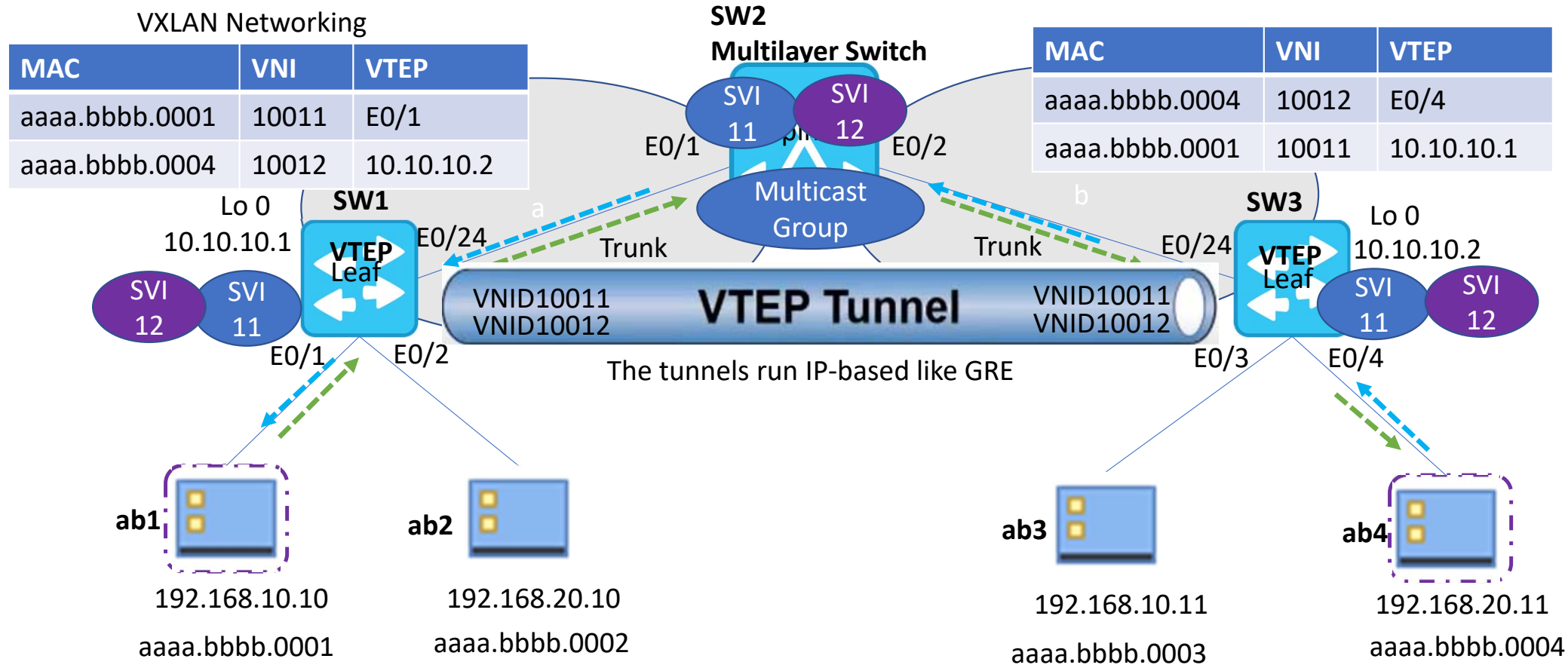


MAC	VNI	VTEP
aaaa.bbbb.0001	10011	E0/1
aaaa.bbbb.0003	10011	10.10.10.2

MAC	VNI	VTEP
aaaa.bbbb.0003	10011	E0/3
aaaa.bbbb.0001	10011	10.10.10.1

How VXLAN Works

Traffic flow from ab1 to ab4 is to be routed via L3 VNI

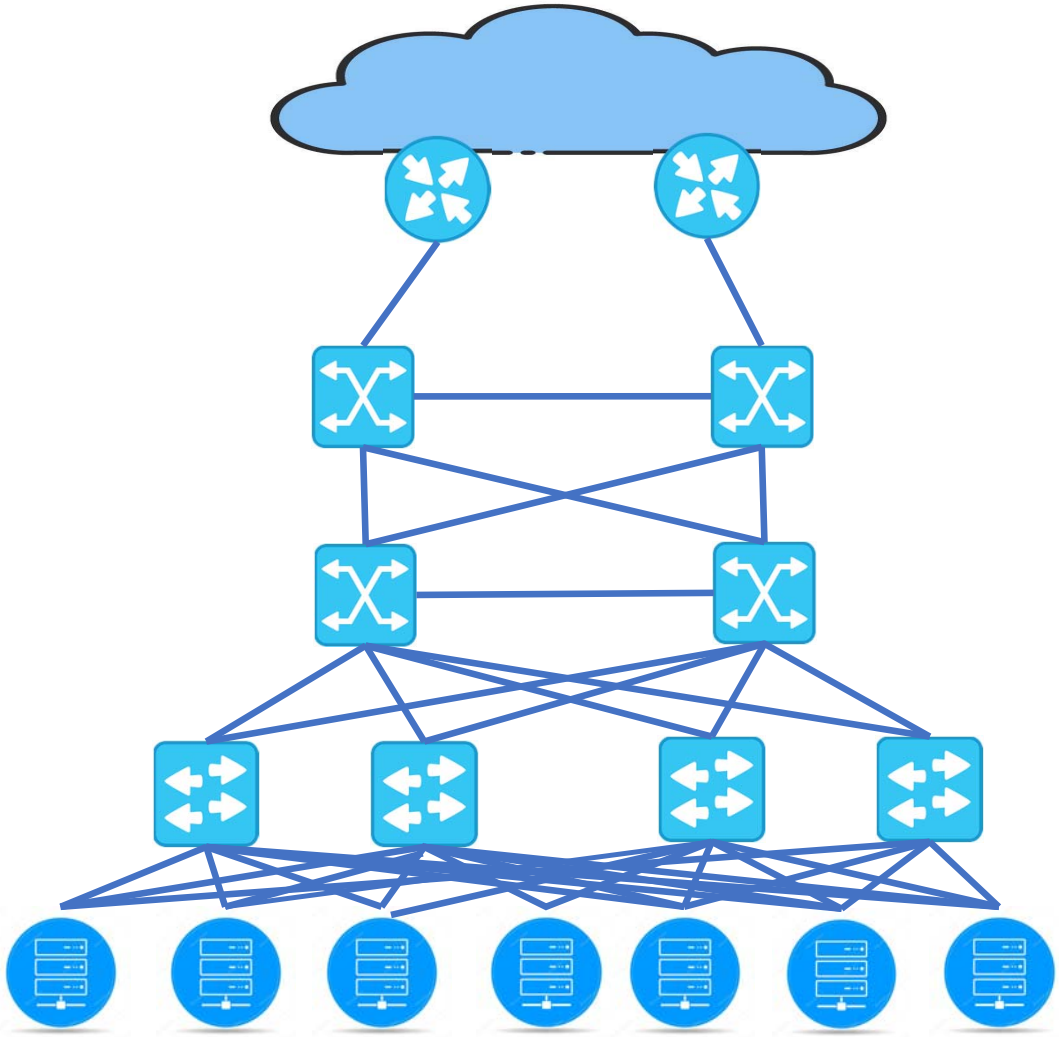


Current Challenges

- Over-Subscription
- Scalability
- Cost
- Mobility
- Latency
- Manageability

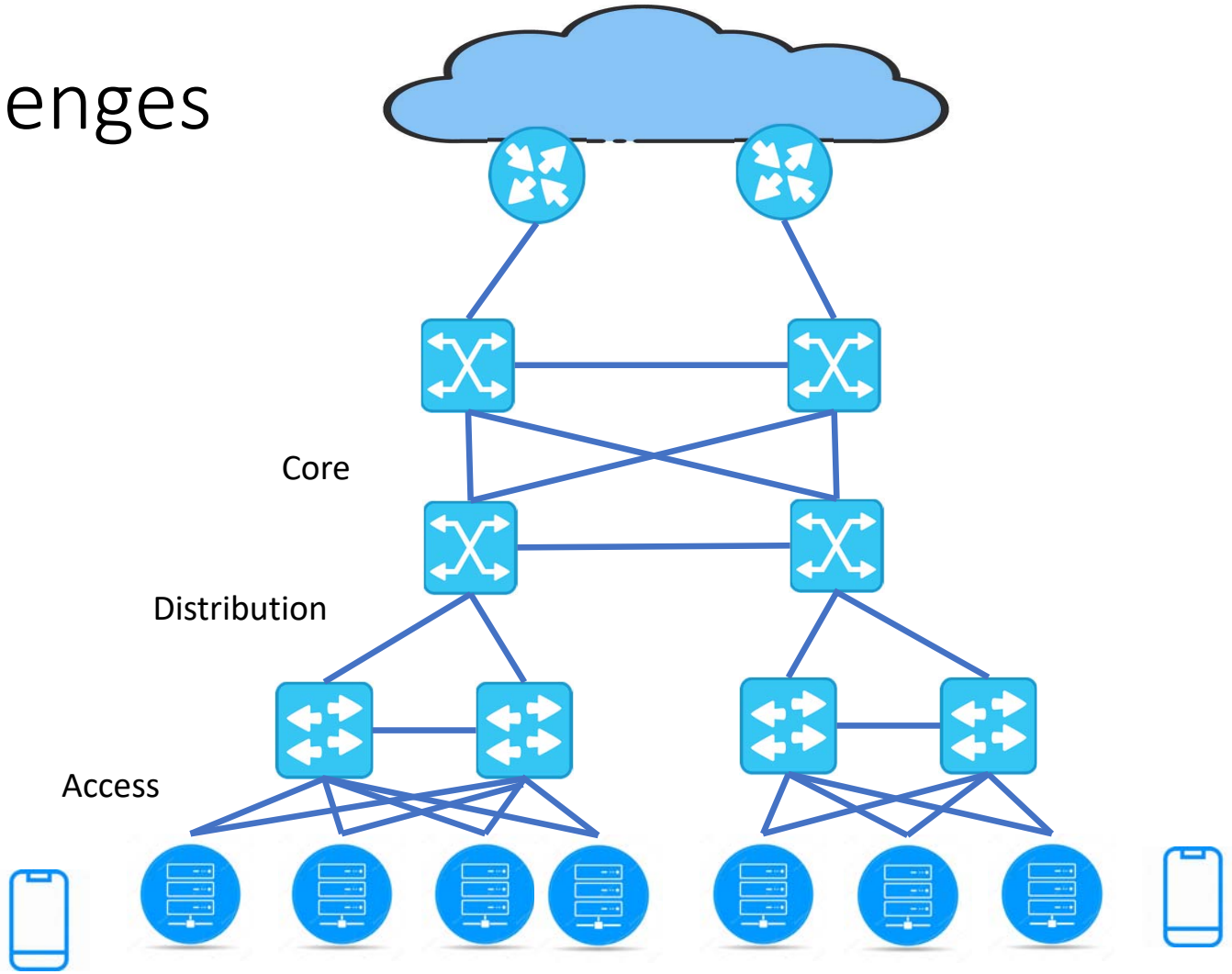
Current Challenges

Traditional Networking



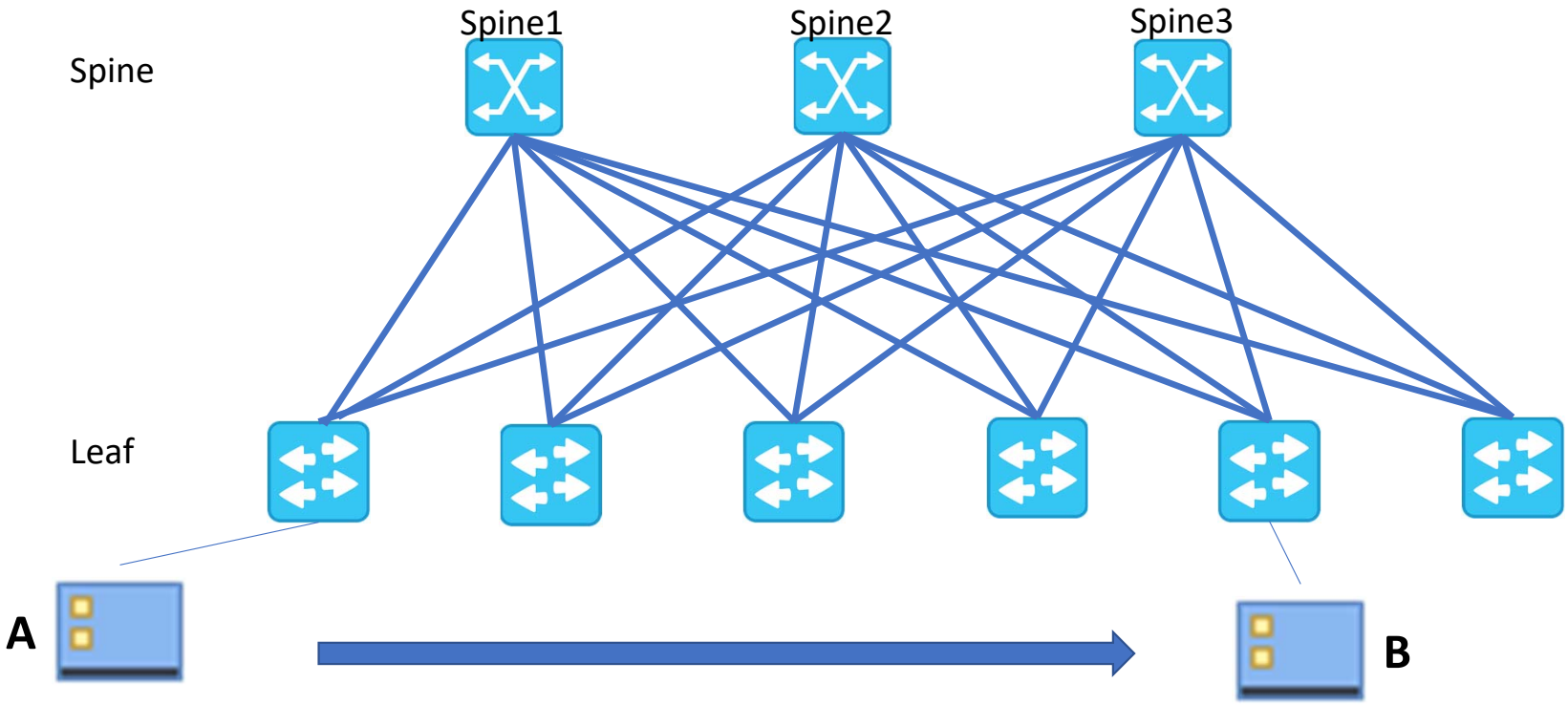
Current Challenges

Hierarchical Architecture



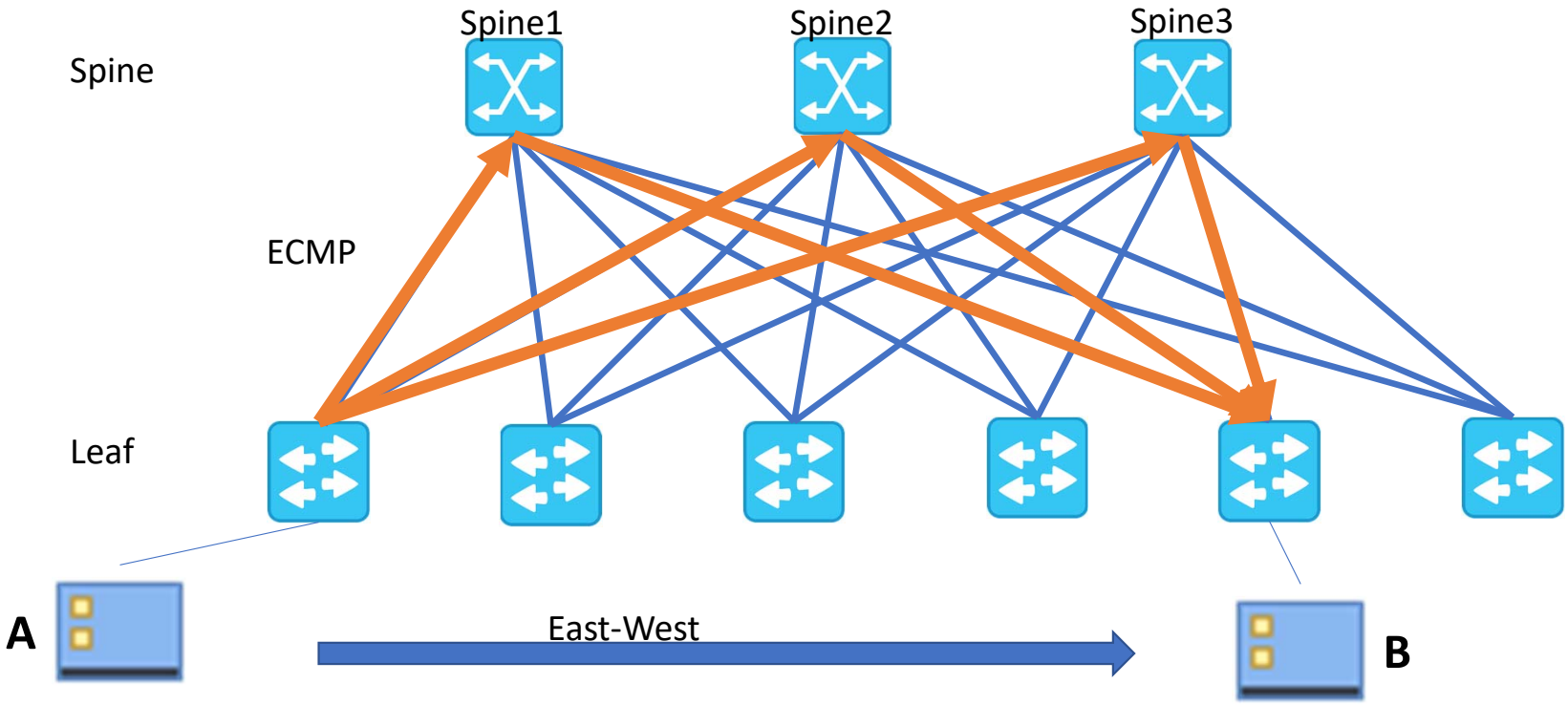
Current Challenges

Spines and Leaves

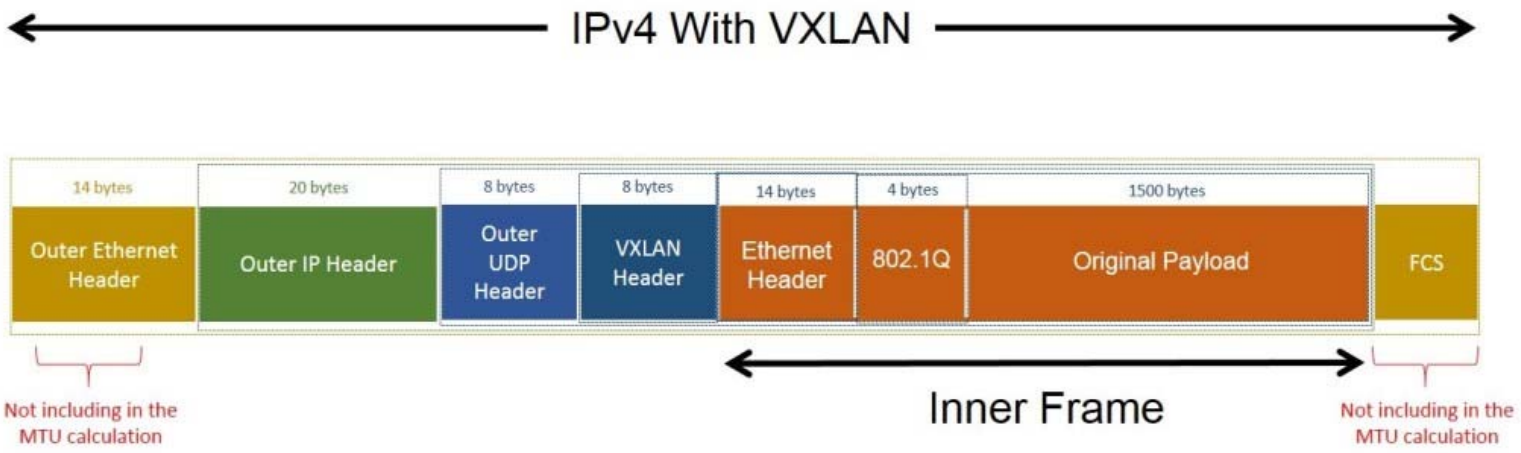


Current Challenges

Spines and Leaves

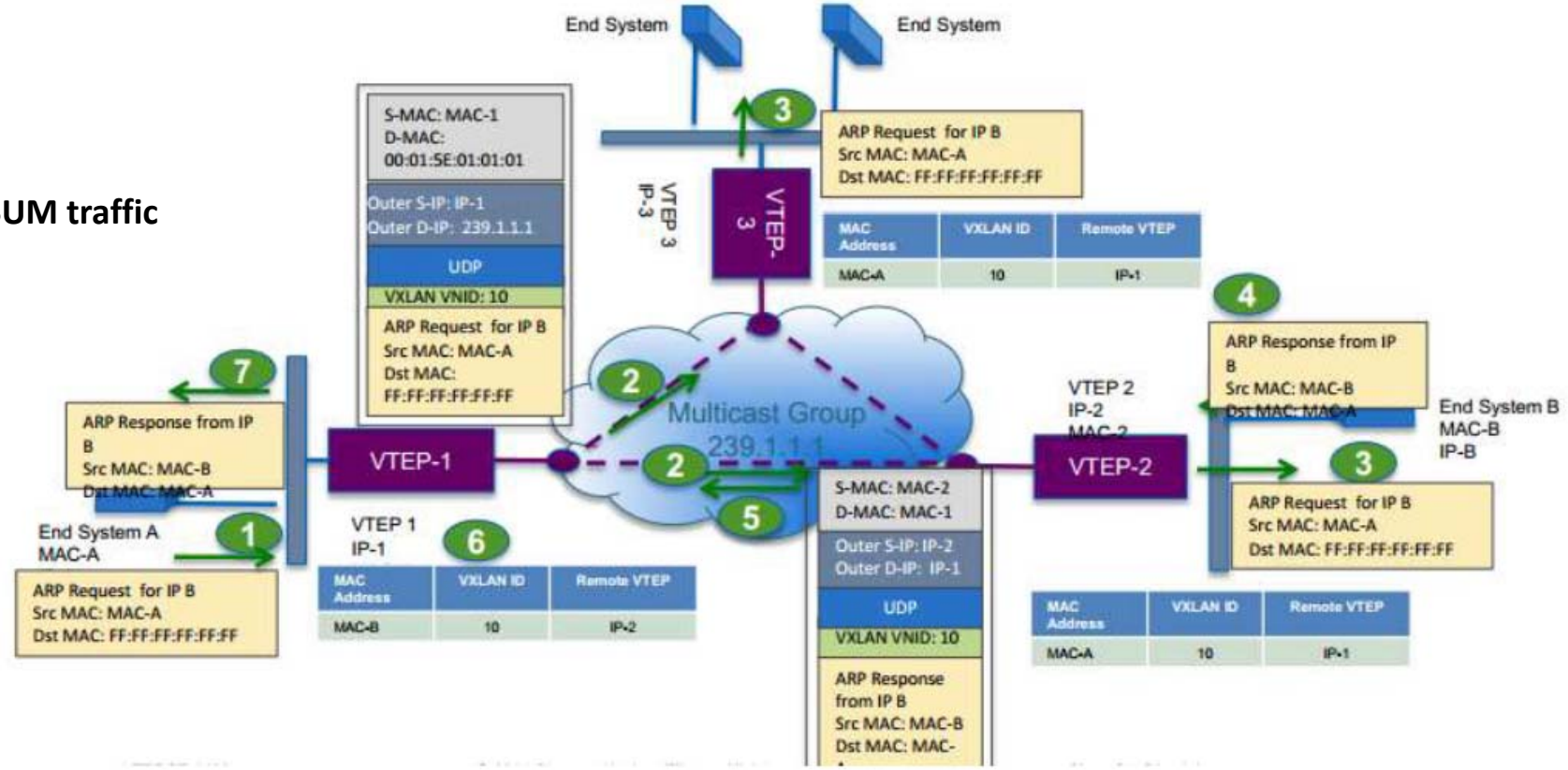


VXLAN Frame Format

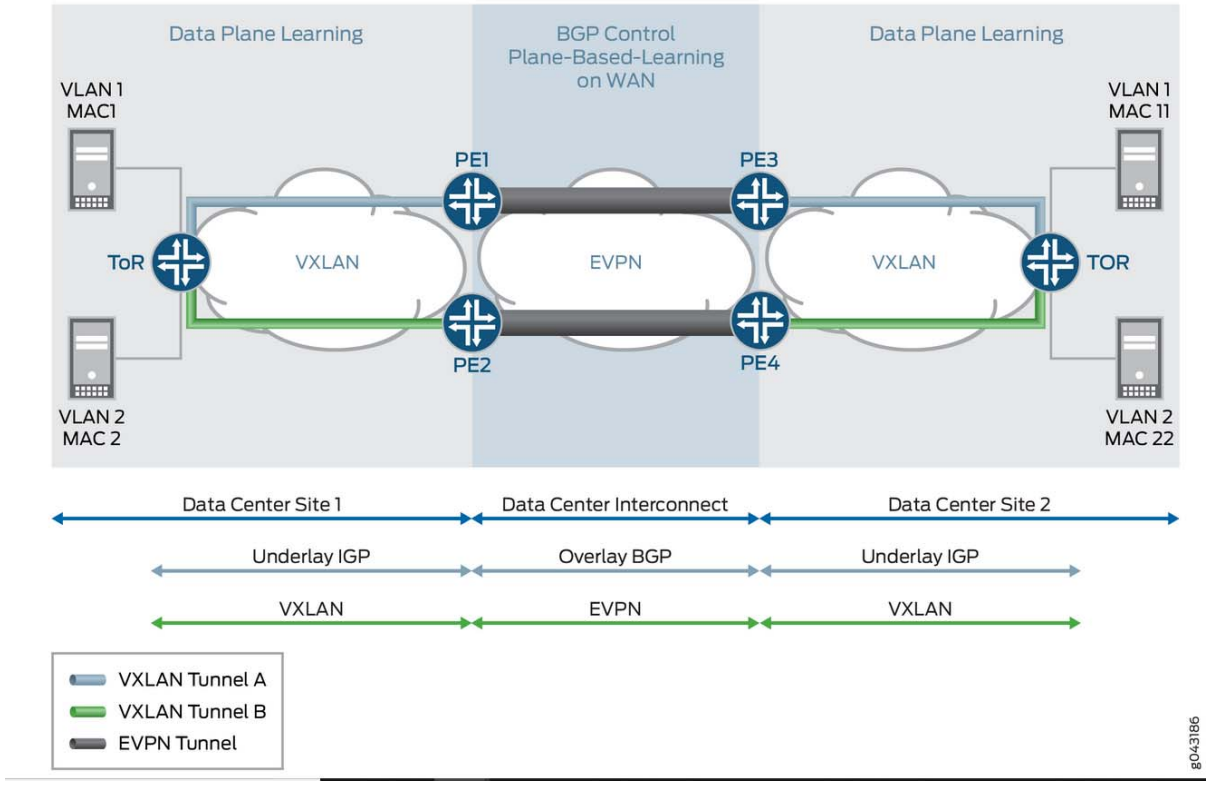


VXLAN VTEP
Peer Discovery & Address Learning

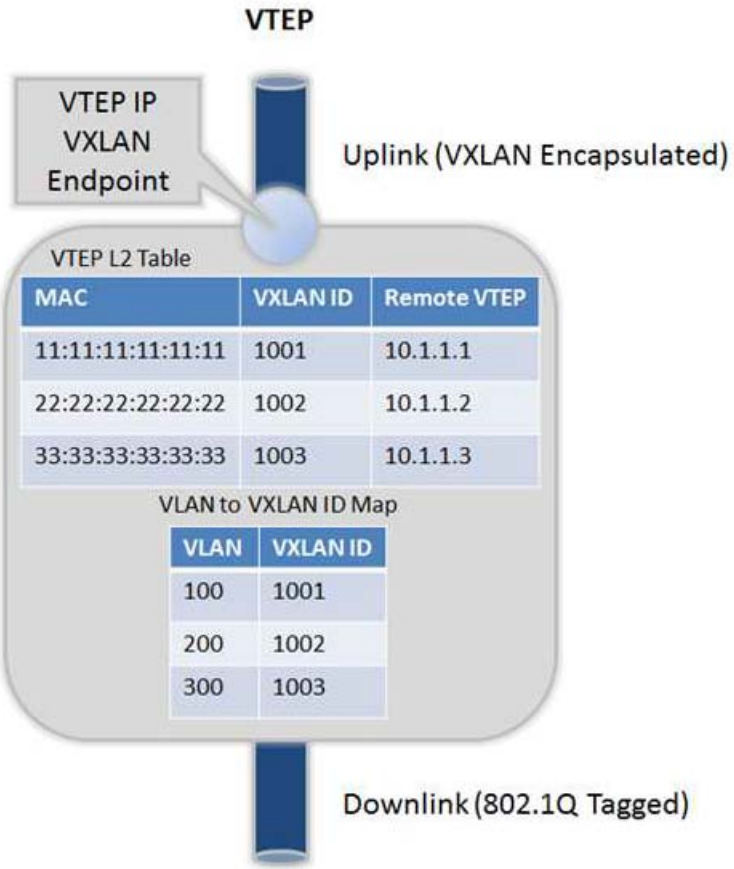
BUM traffic

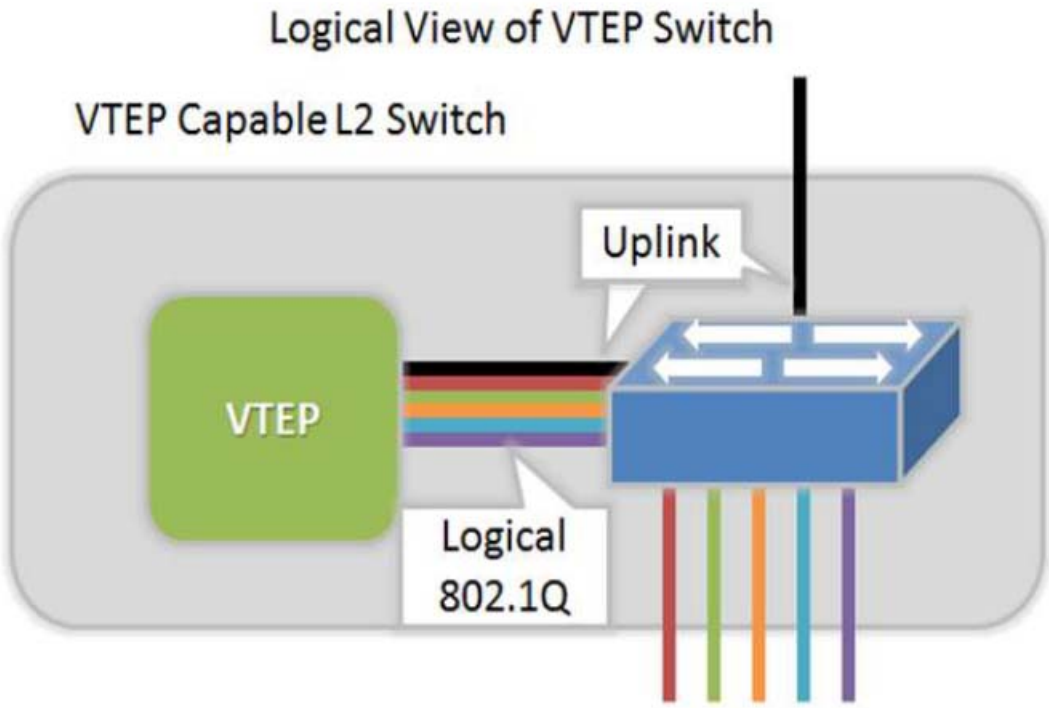


BGP-EVPN with VXLAN



VXLAN Interface (VTEP)





*http://www.definethecloud.net/vxlan-deep-dive/

Configuration Sample

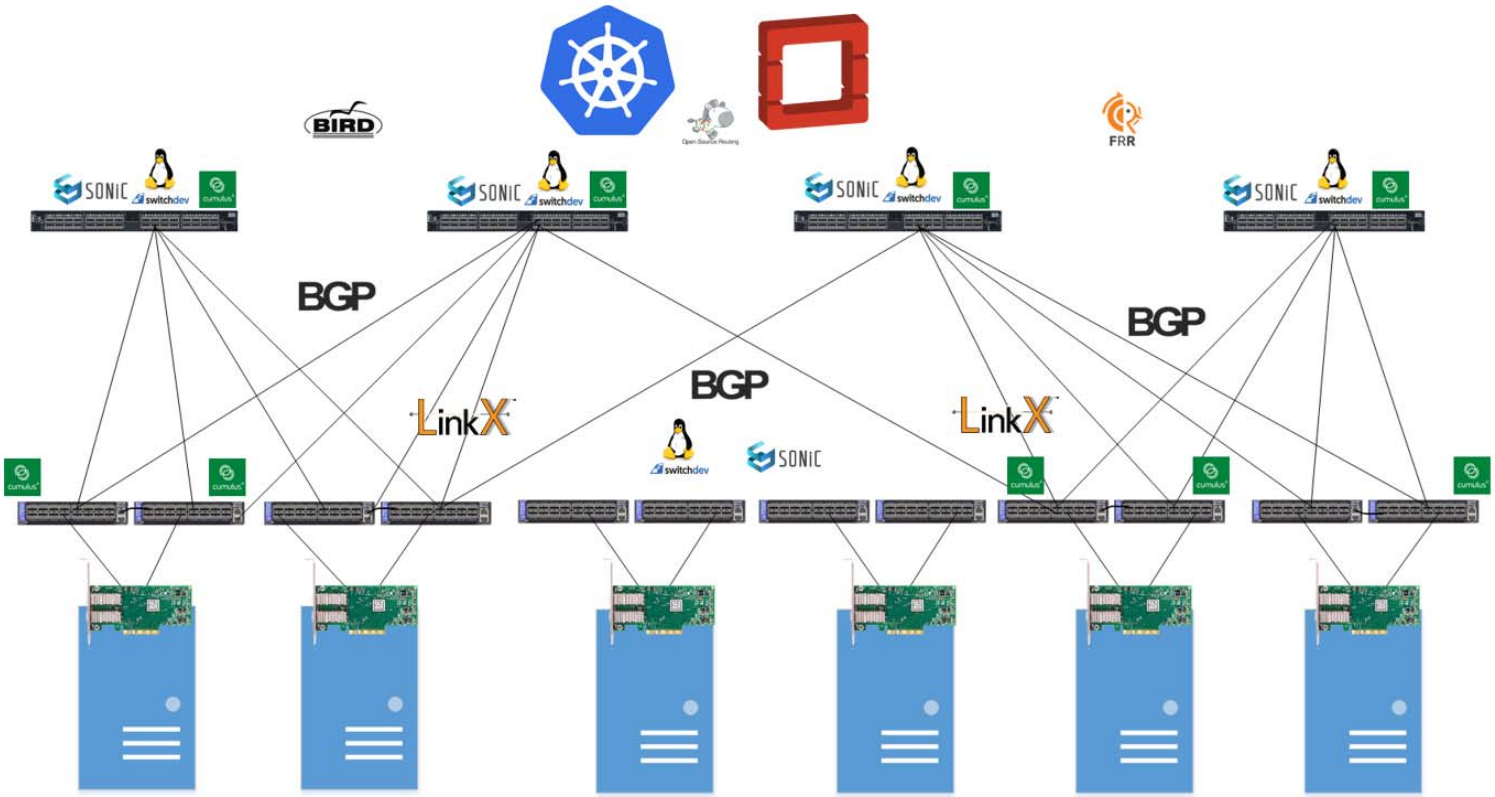
Cisco NX-OS N9K

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

int e4/1
  switchport
  switchport access vlan 100
  no shut

int nve 1
  no shut
  source-interface lo0
  member vni 1010 mcast-group 239.1.1.1

vlan 10
  vn-segment 1010
```

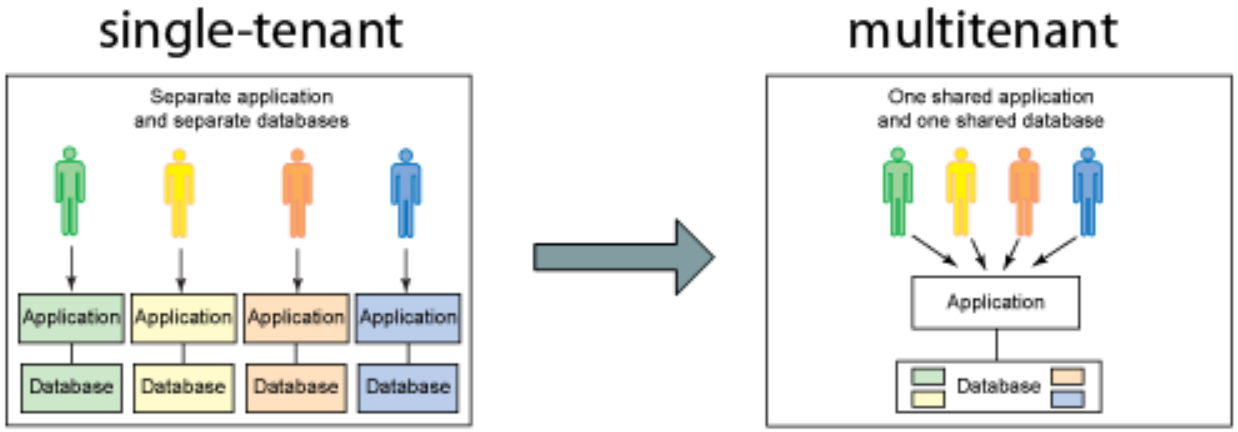
Benefits of using VXLAN

- Theoretically create as many as 16 million VXLANs in an administrative domain
- Enable migration of virtual machines between servers in separate Layer 2 domains by tunneling over Layer 3 networks
- No need to use STP to converge the topology
  - All links can be used
  - Traffic can be load balanced
  - Maximizes performance

Multi-Tenancy



Multi-Tenancy



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.

Multi-Tenancy at Layer-2

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

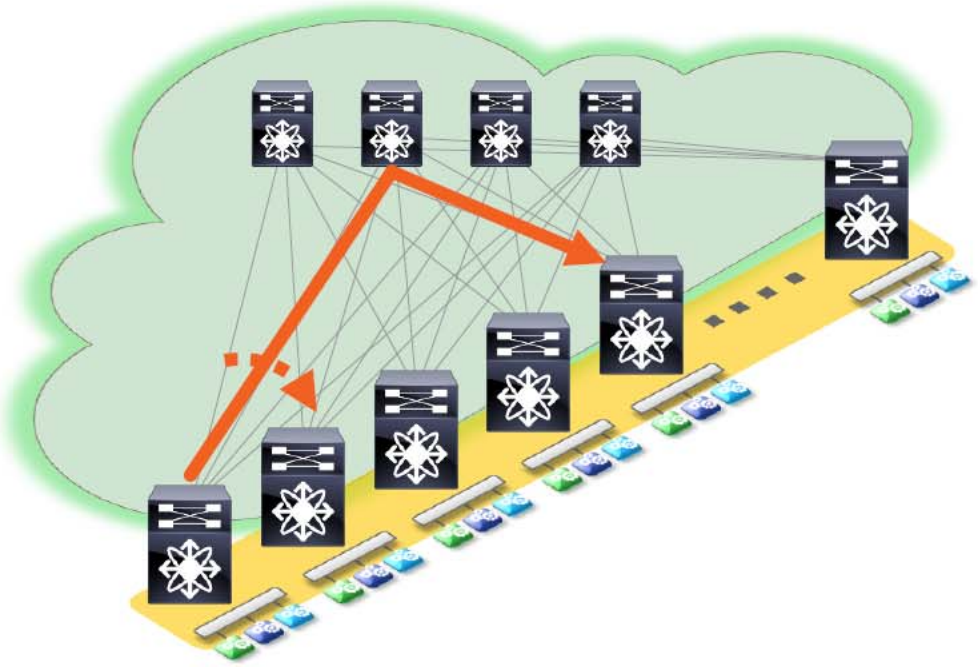
Multi-Tenancy at Layer-3

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

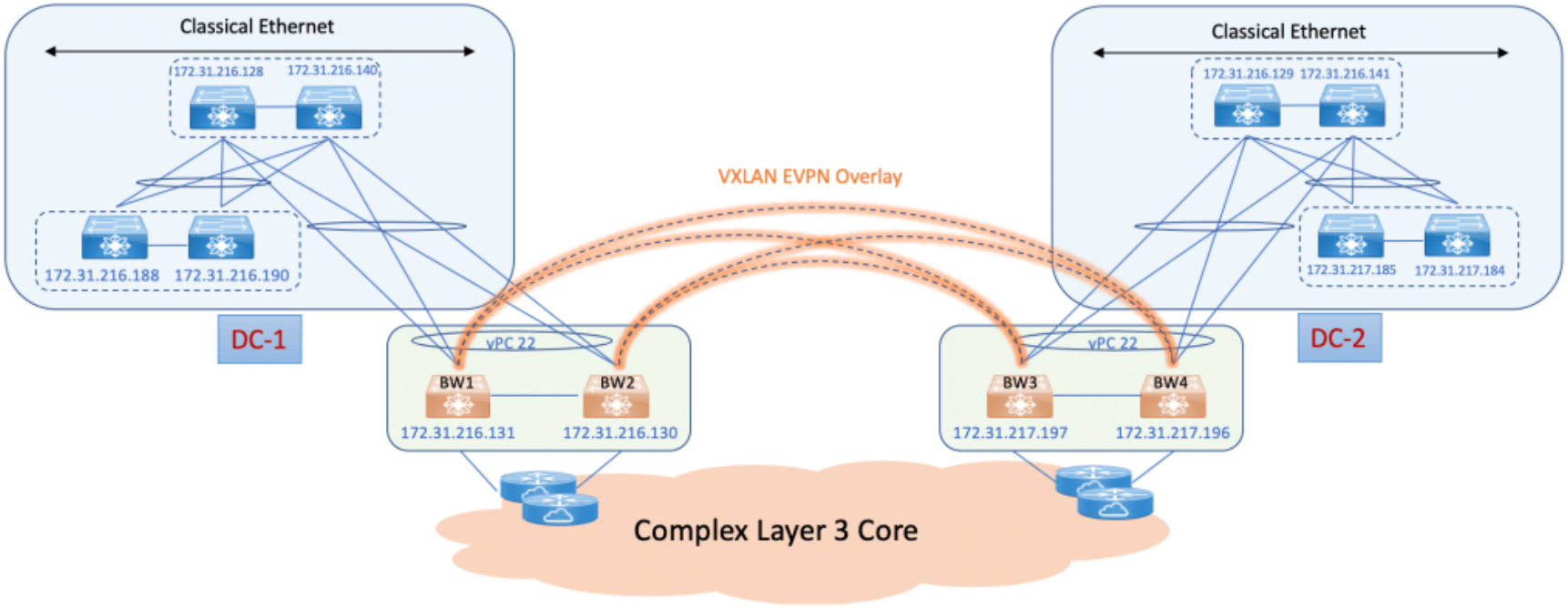
Fabric with Overlays Management

Spine/Leaf Topologies

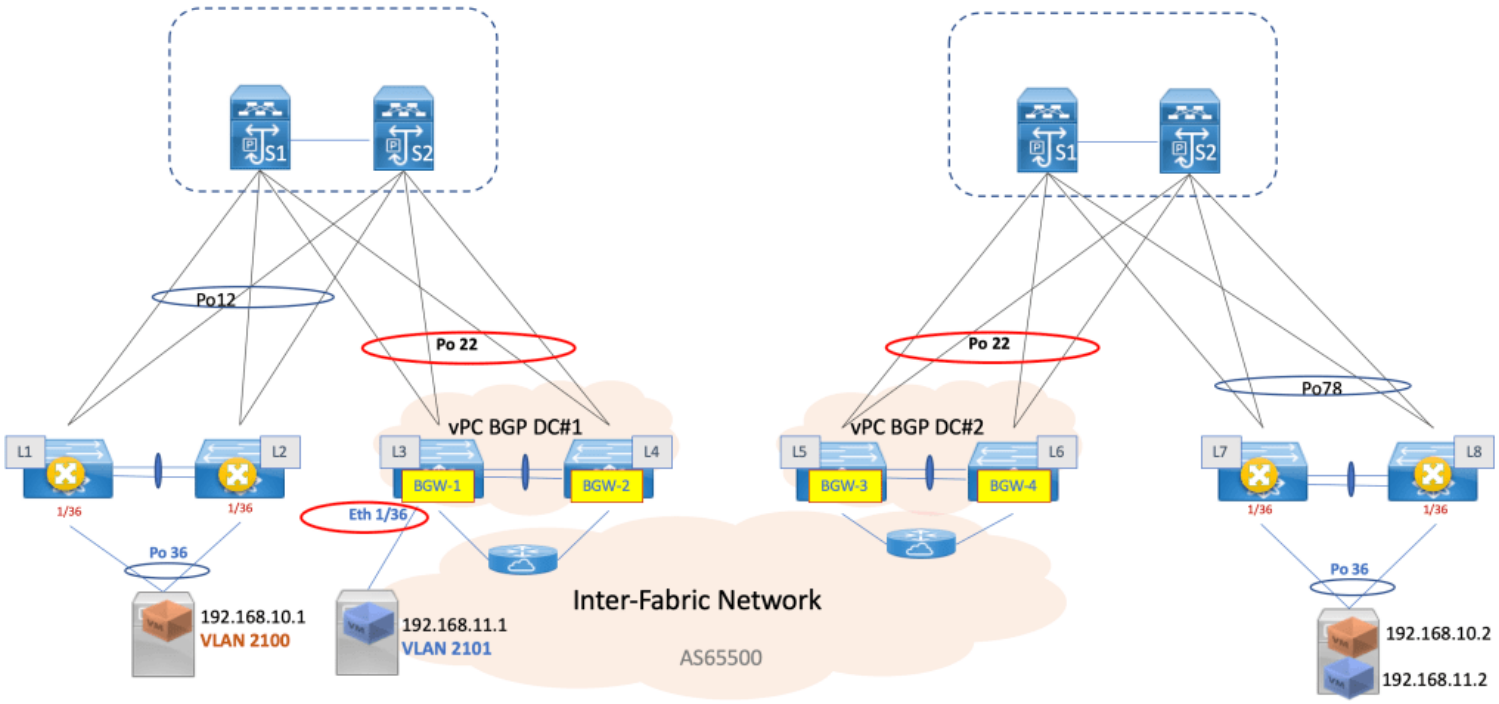
- High Bi-Sectional Bandwidth
- Wide ECMP: Unicast or Multicast
- Uniform Reachability, Deterministic Latency
- High Redundancy: Node/Link Failure
- Line rate, low latency, for all traffic



Use-Cases


Classical Ethernet
172.31.216.128
172.31.216.140
172.31.216.188
172.31.216.190
DC-1
VXLAN EVPN Overlay
vPC 22
BW1
BW2
172.31.216.131
172.31.216.130
Classical Ethernet
172.31.216.129
172.31.216.141
172.31.217.185
172.31.217.184
DC-2
vPC 22
BW3
BW4
172.31.217.197
172.31.217.196
Complex Layer 3 Core

Use-Cases


S1
S2
Po12
Po 22
vPC BGP DC#1
L1
L2
L3
L4
BGW-1
BGW-2
1/36
1/36
Eth 1/36
Po 36
192.168.10.1
VLAN 2100
192.168.11.1
VLAN 2101
Inter-Fabric Network
AS65500
S1
S2
Po 22
Po78
vPC BGP DC#2
L5
L6
L7
L8
BGW-3
BGW-4
1/36
1/36
Po 36
192.168.10.2
192.168.11.2

Network Automation with VXLAN

Examples;

- Cisco ACI Fabric
- EVPN with VXLAN
- Cisco Data Center Network Manager
- Apstra

Network Automation with VXLAN

EVPN-VXLAN campus networks provide the following benefits:

- Consistent, scalable architecture
- Multi-vendor deployment
- Reduced flooding and learning
- Location-agnostic connectivity
- Underlay agnostic
- Consistent network segmentation
- Simplified management

Network Automation with VXLAN

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 Virtualization 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

Overlay Comparisons

VXLAN / STT

Stateless Transport Tunneling Protocol

Similarities

- IP Transport
- IP Multicast
  - For broadcast and multicast frames
- Port Channel Load Distribution
  - 5 Tuple Hashing (UDP vs TCP)

Differences

- IETF Draft Authors
  - VXLAN: Cisco, VMware, Citrix, Red Hat, Broadcom, Arista
  - STT: Nicira
- Encapsulation
  - VXLAN: UDP with 50 bytes
  - STT: "TCP-like" with 72 to 54 bytes (not uniform) *
- Segment ID Size
  - VXLAN: 24 bit
  - STT: 64 bit
- Firewall ACL can act on VXLAN UDP port
  - Firewalls will likely block STT since it has no TCP state machine handshake
- Forwarding Logic
  - VXLAN: Flooding/Learning
  - STT: Not specified

Overlay Comparisons

VXLAN / NVGRE

Network Virtualization using Generic Routing Encapsulation

Similarities

- IP Transport
- IP Multicast
  - For broadcast and multicast frames
- 24 Bit Segment ID

Differences

- IETF Draft Authors
  - VXLAN: Cisco, VMware, Citrix, Red Hat, Broadcom, Arista
  - STT: Microsoft, Intel, Dell, HP, Broadcom, Emulex, Arista
- Encapsulation
  - VXLAN: UDP with 50 bytes
  - NVGRE: GRE with 42 bytes
- Port Channel Load Distribution
  - VXLAN: UDP 5-tuple hashing
  - Most (if not all) current switches do not hash on the GRE header
- Firewall ACL can act on VXLAN UDP port
  - Difficult for firewall to act on the GRE Protocol Type field
- Forwarding Logic
  - VXLAN: Flooding/Learning
  - NVGRE: Not specified

Overlay Comparisons

VXLAN / OTV

Overlay Transport Virtualization

Similarities

- Same UDP based encapsulation header
  - VXLAN does not use the OTV Overlay ID field
- IP Multicast
  - For broadcast and multicast frames (optional for OTV)
- 24 Bit Segment ID

Differences

- Forwarding Logic
  - VXLAN: Flooding/Learning
  - OTV: Uses the IS-IS protocol to advertise the MAC address to IP bindings
- OTV can locally terminate ARP and doesn't flood unknown MACs
- OTV can use an adjacency server to eliminate the need for IP multicast
- OTV is optimized for Data Center Interconnect to extend VLANs between or across data centers
- VXLAN is optimized for intra-DC and multi-tenancy

Overlay Comparisons

VXLAN / LISP

Locator / ID Separation Protocol

Similarities

- Same UDP based encapsulation header
  - VXLAN does not control flag bits or Nonce/MapVersion field
  - 24 Bit Segment ID

Differences

- LISP carries IP packets, while VXLAN carries Ethernet frames
- Forwarding Logic
  - VXLAN: Flooding/Learning
  - LISP: Uses a mapping system to register/resolve inner IP to outer IP mappings
- IP Multicast is only required to carry host IP multicast traffic
- LISP is designed to give IP address (Identifier) mobility / multi-homing and IP core route scalability
- LISP can provide optimal traffic routing when Identifier IP addresses move to a different location

Introduction to VXLAN

Q & A

Introduction to VXLAN

Thank You

References

- https://www.youtube.com/watch?v=XC62Dqn8S-g
- https://www.youtube.com/watch?v=dpbXjRx3hB8
- https://www.youtube.com/watch?v=QPqVtguOz4w&t=1355s
- https://docplayer.net/21451213-Vxlan-bridging-routing.html
- https://dev.to/jjude/what-is-a-multi-tenant-system-bpd
- https://www.techopedia.com/definition/4804/virtual-local-area-network-vlan
- Cisco Live 2016
- https://www.juniper.net/documentation/en_US/junos/topics/concept/evpn-vxlan-data-plane-encapsulation.html