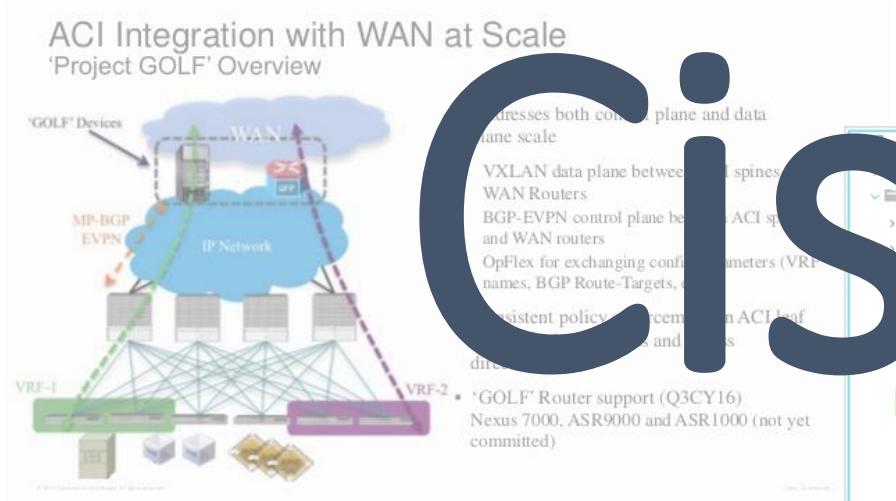
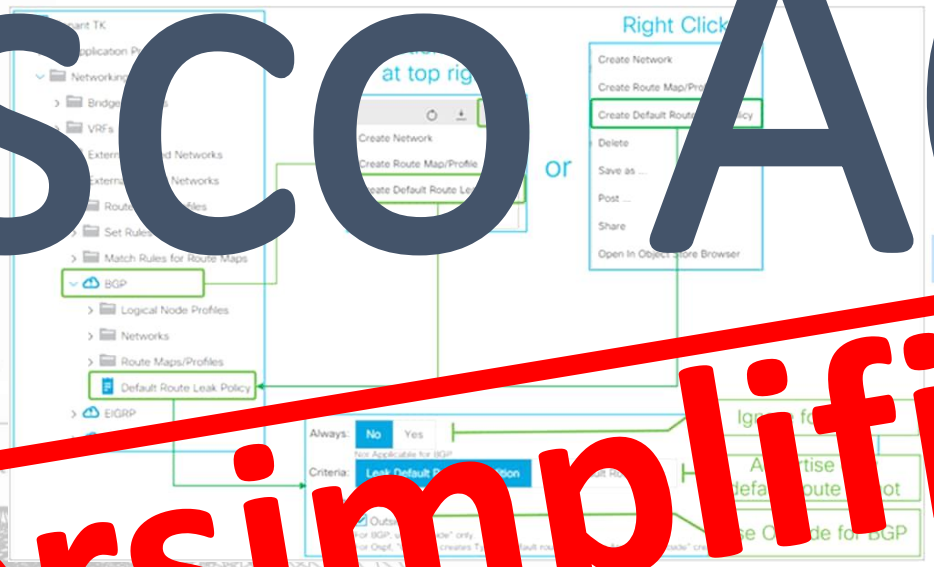
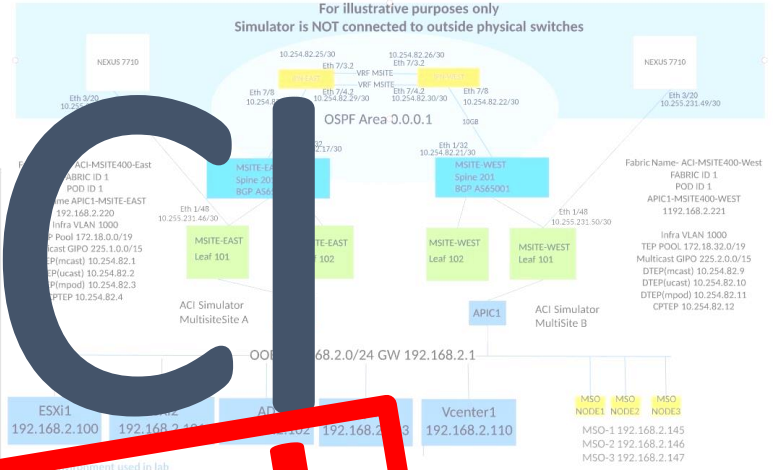
Cisco ACI

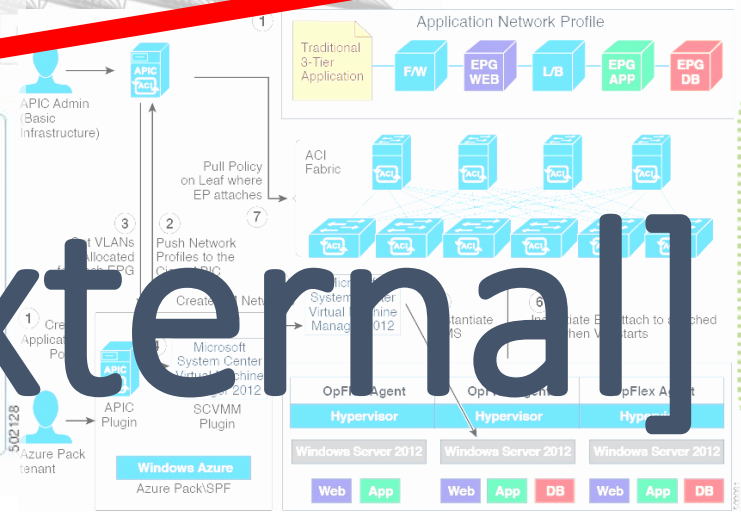
Oversimplified

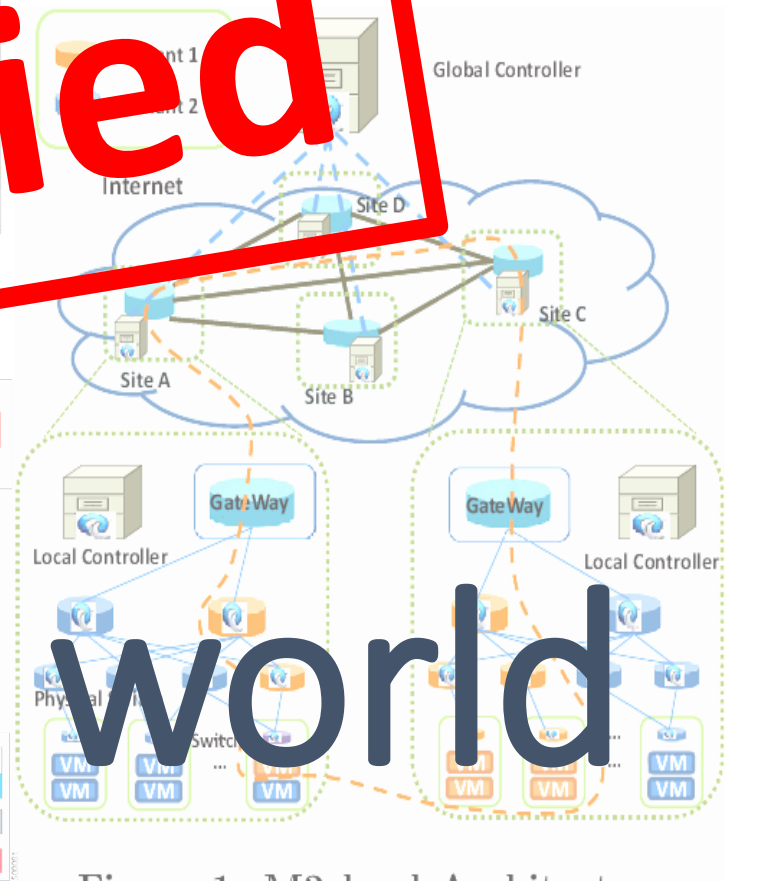
Hello [external] world





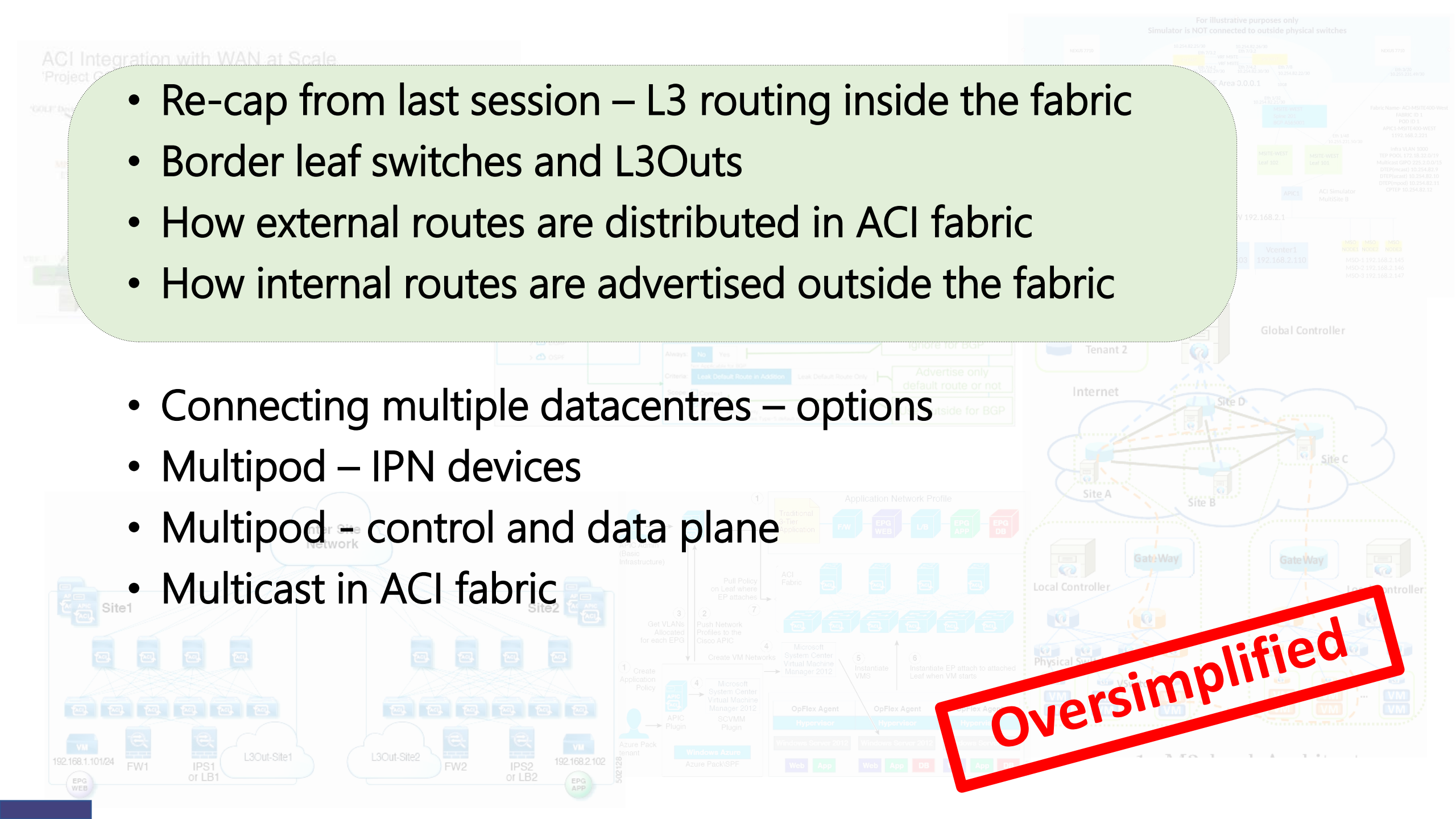






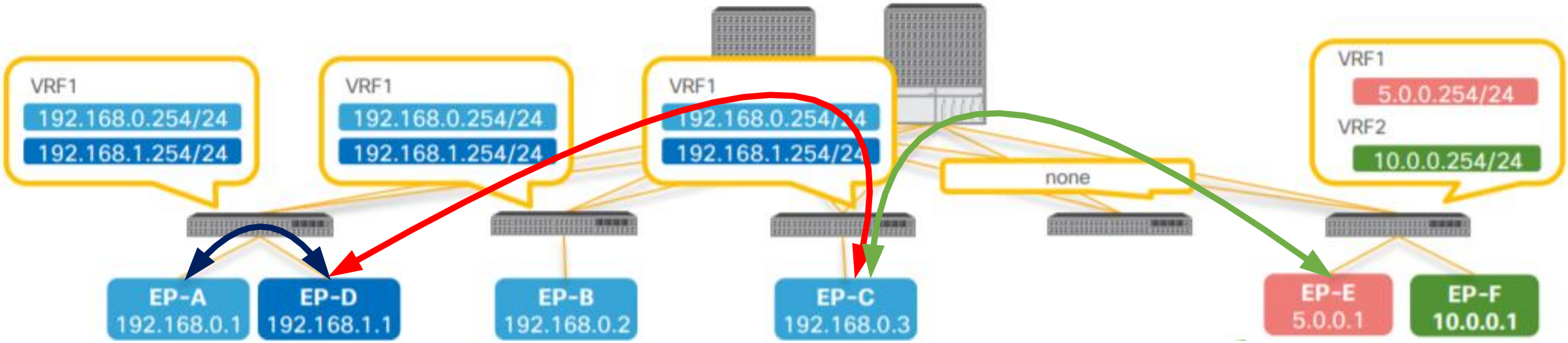
- Re-cap from last session – L3 routing inside the fabric
- Border leaf switches and L3Outs
- How external routes are distributed in ACI fabric
- How internal routes are advertised outside the fabric

- Connecting multiple datacentres – options
- Multipod – IPN devices
- Multipod – control and data plane
- Multicast in ACI fabric



Recap from last session - Anycast (Pervasive) Gateway

- **Every leaf switch** is configured as a default gateway for all connected L3 endpoint subnets
- SVI with same IP and MAC address on all leaf switches
- No concept as active/standby – all leaf switches are 'active' default gateway for **their** connected endpoints
- No central default gateway
- Endpoint send traffic to the local leaf - default GW, the leaf then sends traffic to remote leaf directly using VXLAN
- Traffic goes directly between every leaf (via Spines obviously, as they are physically connected via Spines)
- In ACI it is called Pervasive Gateway, in all other vendors implementations it's called Distributed Anycast Gateway

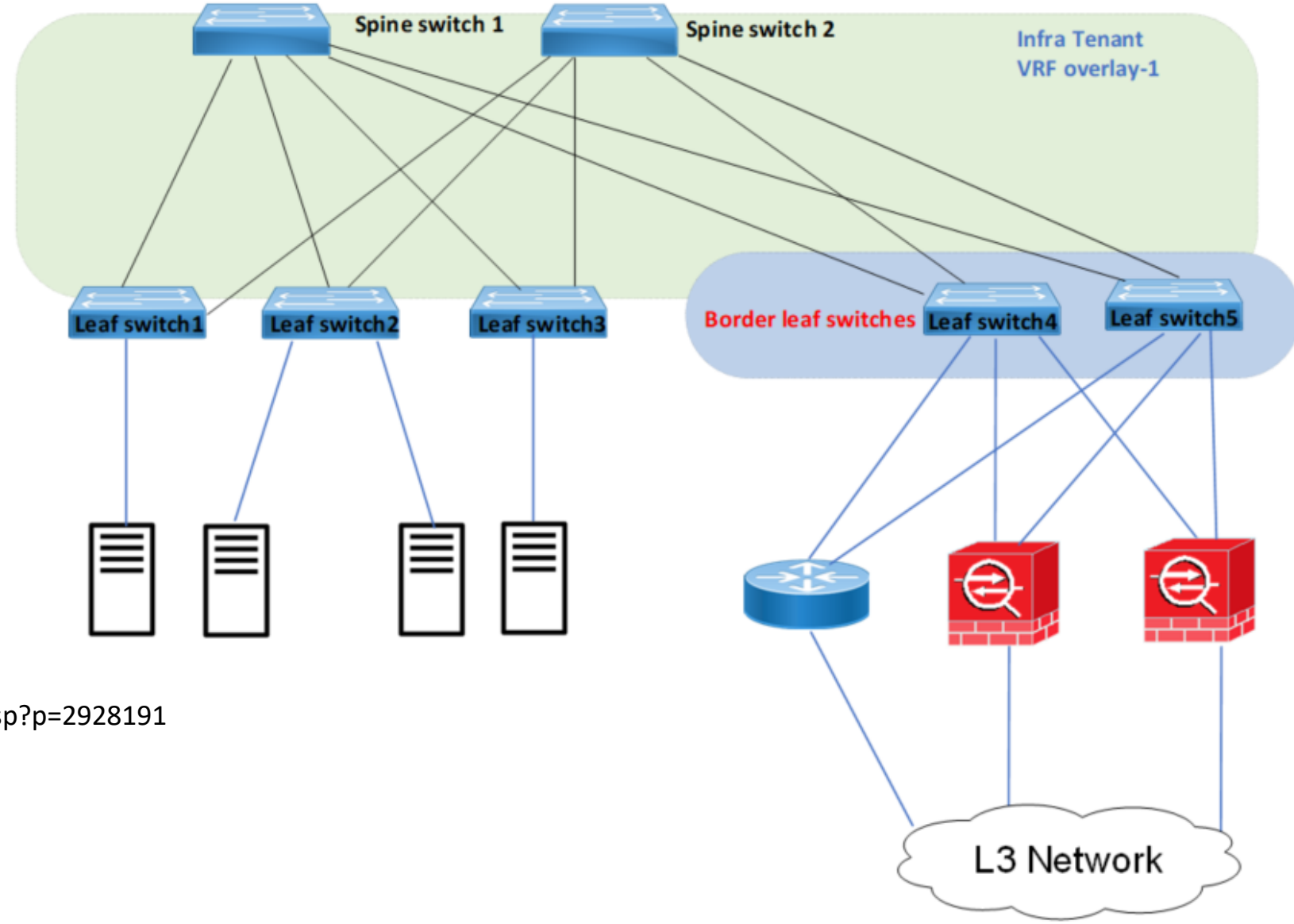


External connectivity via Border Leaf switches – L3Out

- External L3 connections are called L3Out
- External L3 connectivity provided by Border Leaf switches (BL)
- Any Leaf can be border
- Can be both border and compute
- Recommended to have dedicated BLs
- There are external L2 connections – L2Out (not discussed here)
- Various options how to connect to external L3 devices – the same as we connect L3 devices to Nexuses today, see examples here:

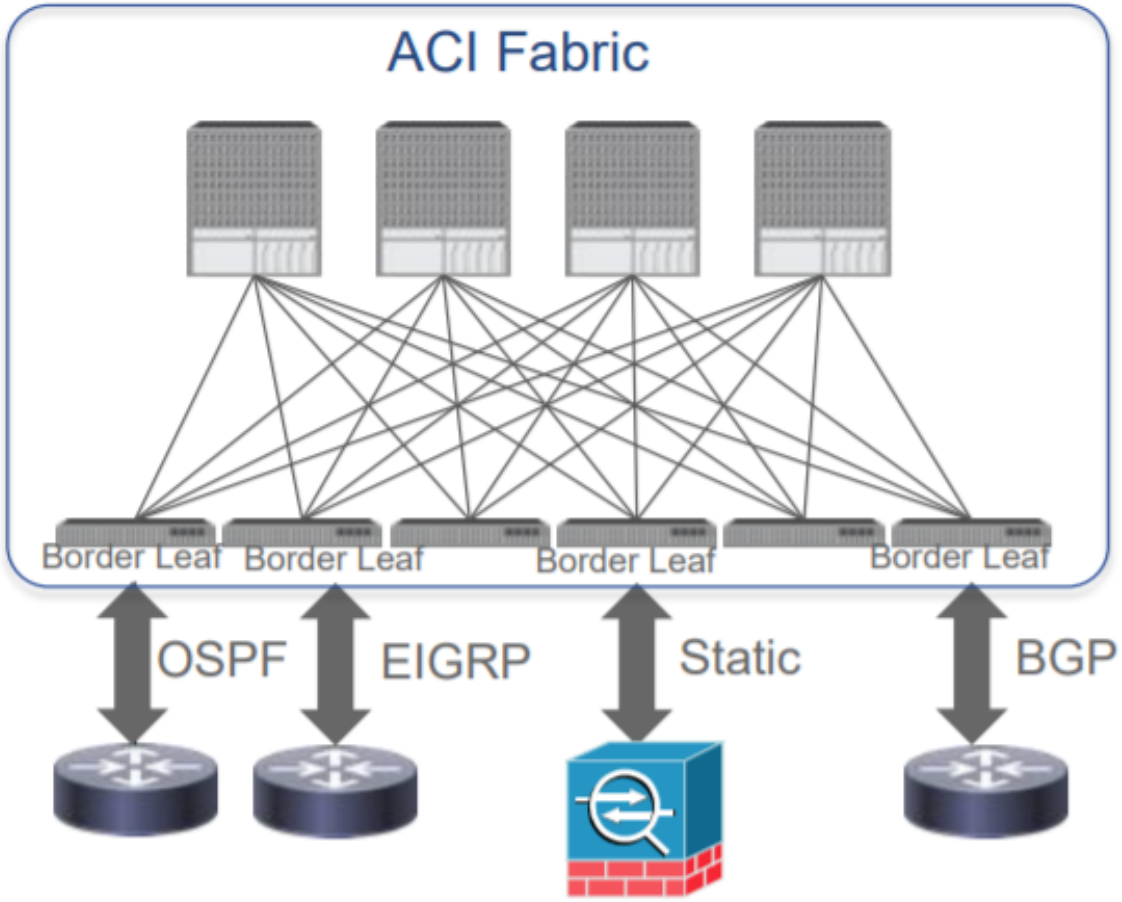
https://www.ciscopress.com/articles/article.asp?p=2928191

- There is an option to use GOLF devices (N7Ks, ASRs), to extend VXLANs up to L3 device (not discussed here)



ACI to external network deployment considerations

- External L3 devices are connected to ACI leaf switches.
- External connections are referred to as “L3 Outside” connections or “L3Out”
- An ACI leaf switch that provides L3 connectivity to outside networks may be referred to as a border leaf.
- Any leaf switch can be a border leaf.
- In large environments it may be preferred to have designated border leaves for scalability reasons
- ACI supports standard L3 protocols (OSPF, BGP), EIGRP, or static routes.
- Supports both IPv4 and IPv6



https://www.ciscolive.com/c/dam/r/ciscolive/emea/docs/2016/pdf/LTRACI-3000.pdf

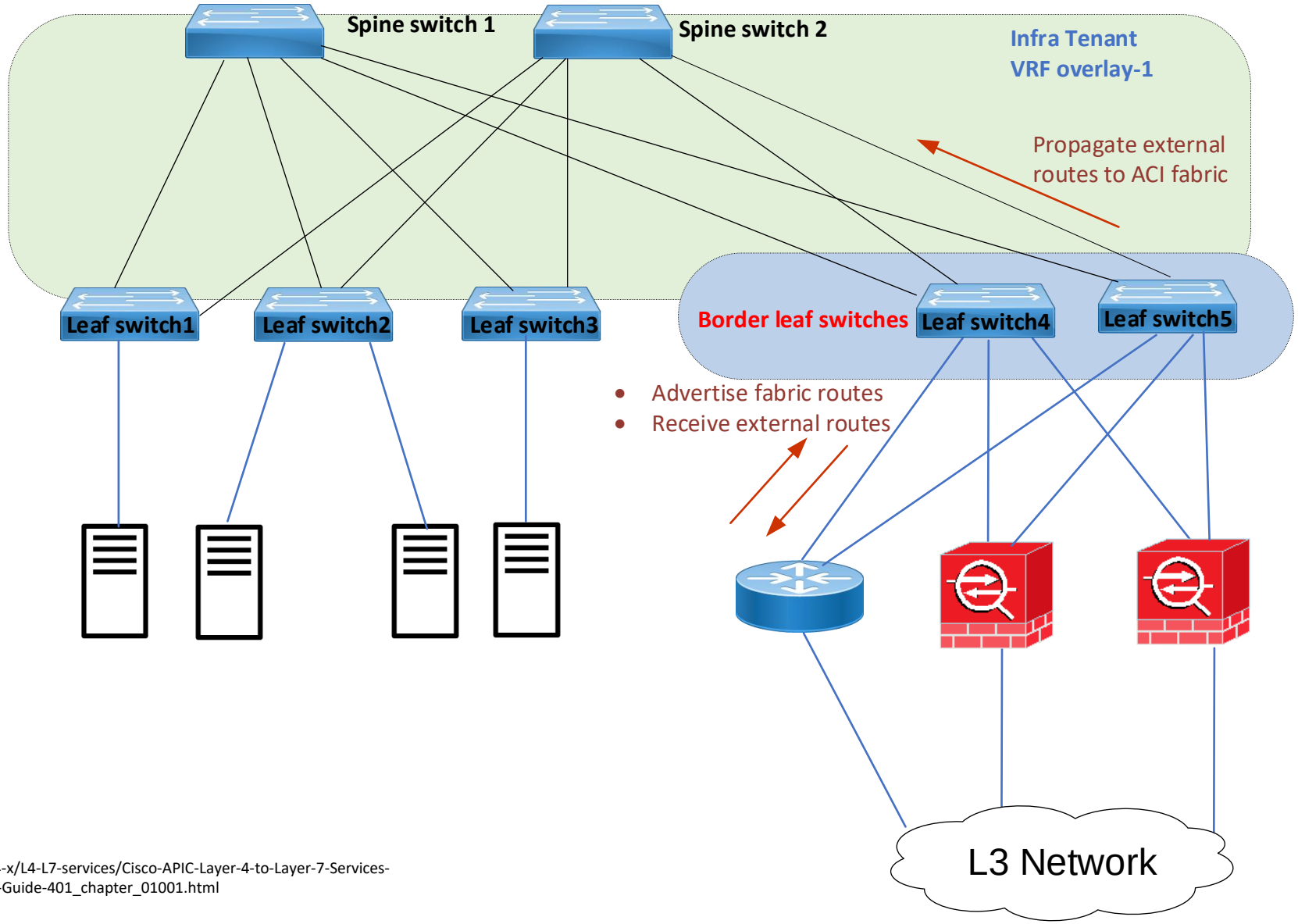
Border Leaf switches - purpose and functions

Purpose of Border Leaves:

- Learn external routes via routing protocols (or static routes)
- Distribute learned external routes (or static routes) to other leaf switches
- Advertise ACI internal routes (BD subnets) to outside ACI
- Advertise learned external routes to other L3Outs (Transit Routing)

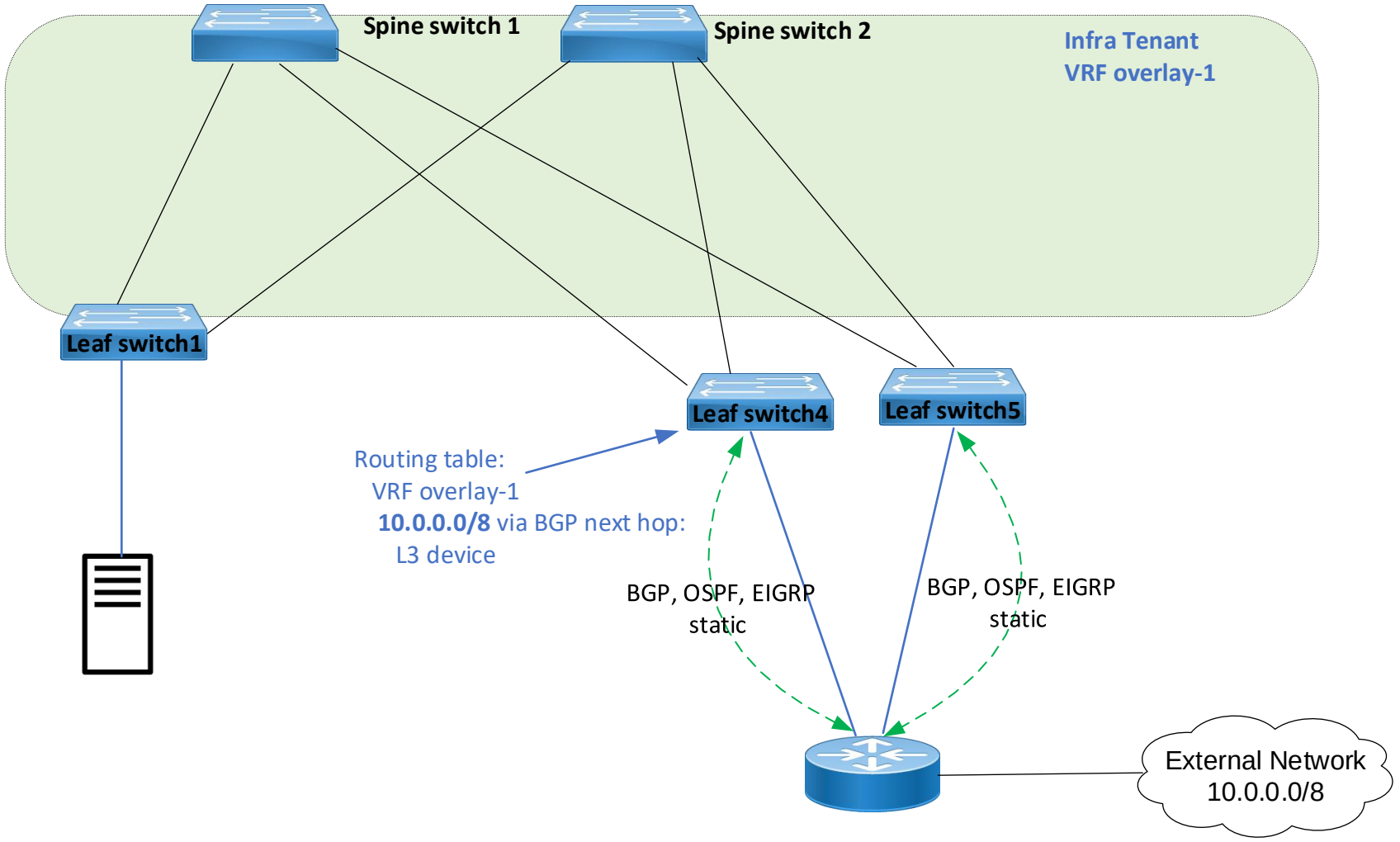
- Note – to divert traffic to **firewalls** and load-balancers **within ACI** it is possible to use **policy-based redirect** or service insertion. These FW and LB services are called L4-L7 services, more 'native' to ACI, only for internal fabric traffic, and different from L3Out

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/4-x/L4-L7-services/Cisco-APIC-Layer-4-to-Layer-7-Services-Deployment-Guide-401/Cisco-APIC-Layer-4-to-Layer-7-Services-Deployment-Guide-401_chapter_01001.html



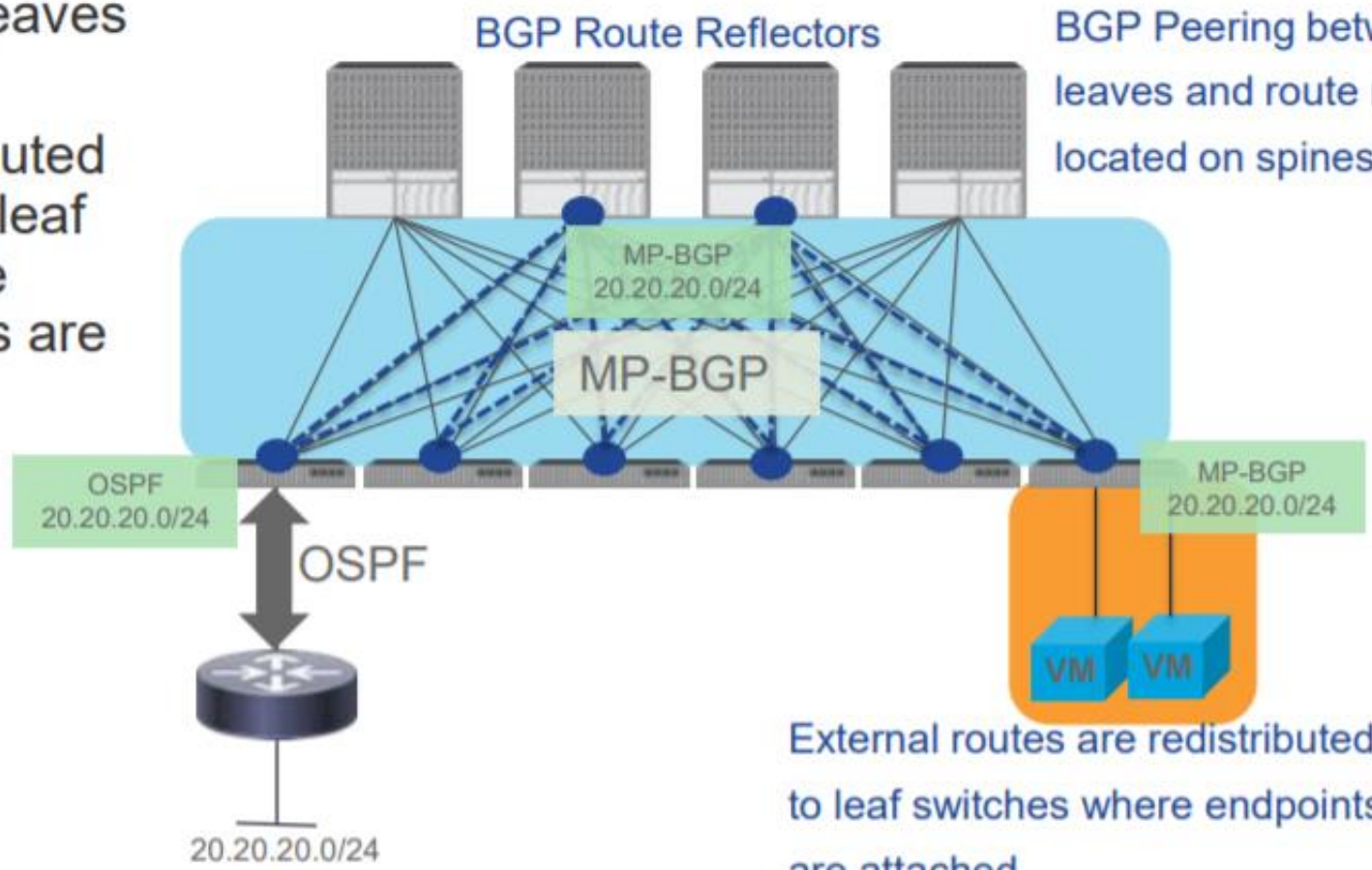
Border Leaves - Learning external routes

- Border Leaves can receive external routes via BGP, OSFP, EIGRP or configure static to L3 device
- External prefixes are received and placed into VRF overlay-1 (infrastructure VRF)
- L3Out is configured per tenant and **associated** to Bridge domain (subnet)
- Possible to change attributes of received routes using route-maps, such as BGP AS-prepend



Redistribution of external Routes in ACI Fabric

- The ACI fabric is not a big router
- ACI runs MP-BGP between leaves and spines
- External prefixes are redistributed across the MP-BGP fabric to leaf switches where endpoints are attached (Where tenant VRFs are deployed).



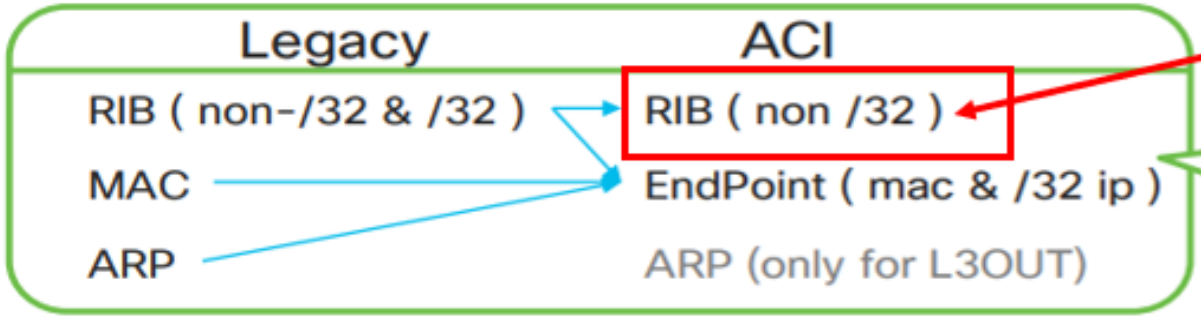
Good old days without magic - routing tables on Leaf switches

- External route are received via iBGP from Spines (Route Reflectors) in **Infra VRF** as VPNv4/v6
- Routes are labelled with route-targets (RT) showing which tenant/VRF they belong to
- External routes imported from the infra VRF to the user VRF based on received RT
- Leaf switches use normal VRF-lite route tables (**RIB**) to look up **external routes**

What Forwarding Table is used?

- End Point Table
  - ➢ host information (MAC and /32 IP address)
- LPM(Longest Prefix Match) Table
  - ➢ non /32 IP route information (exception: /32 for SVI or L3OUT route)

External routes imported by a border leaf and advertised via iBGP are placed in routing table on a leaf

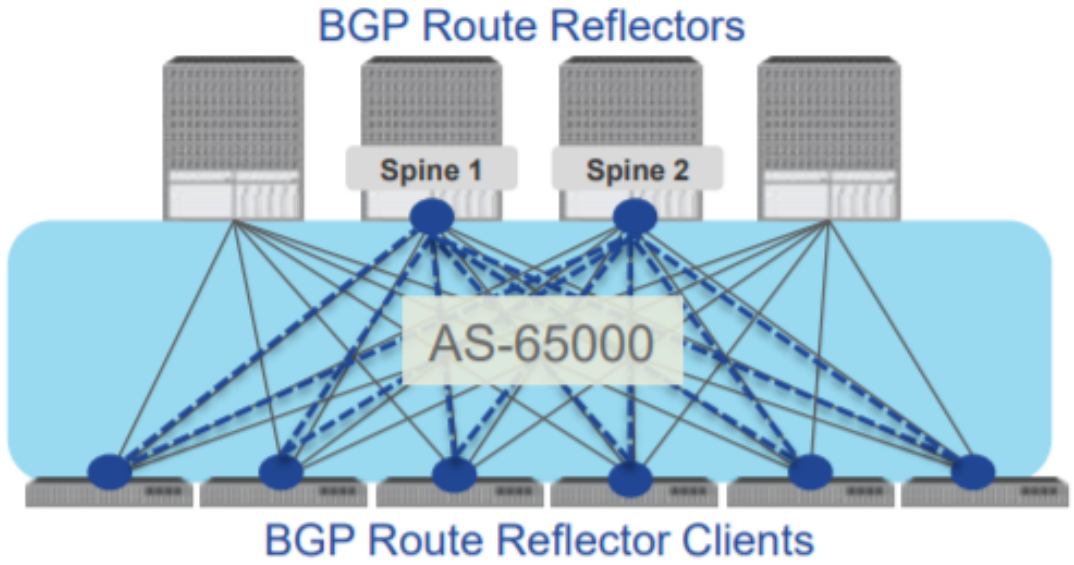


Forwarding table lookup order
1. EndPoint Table (show endpoint)
2. RIB (show ip route)

RIB : Routing Information Base

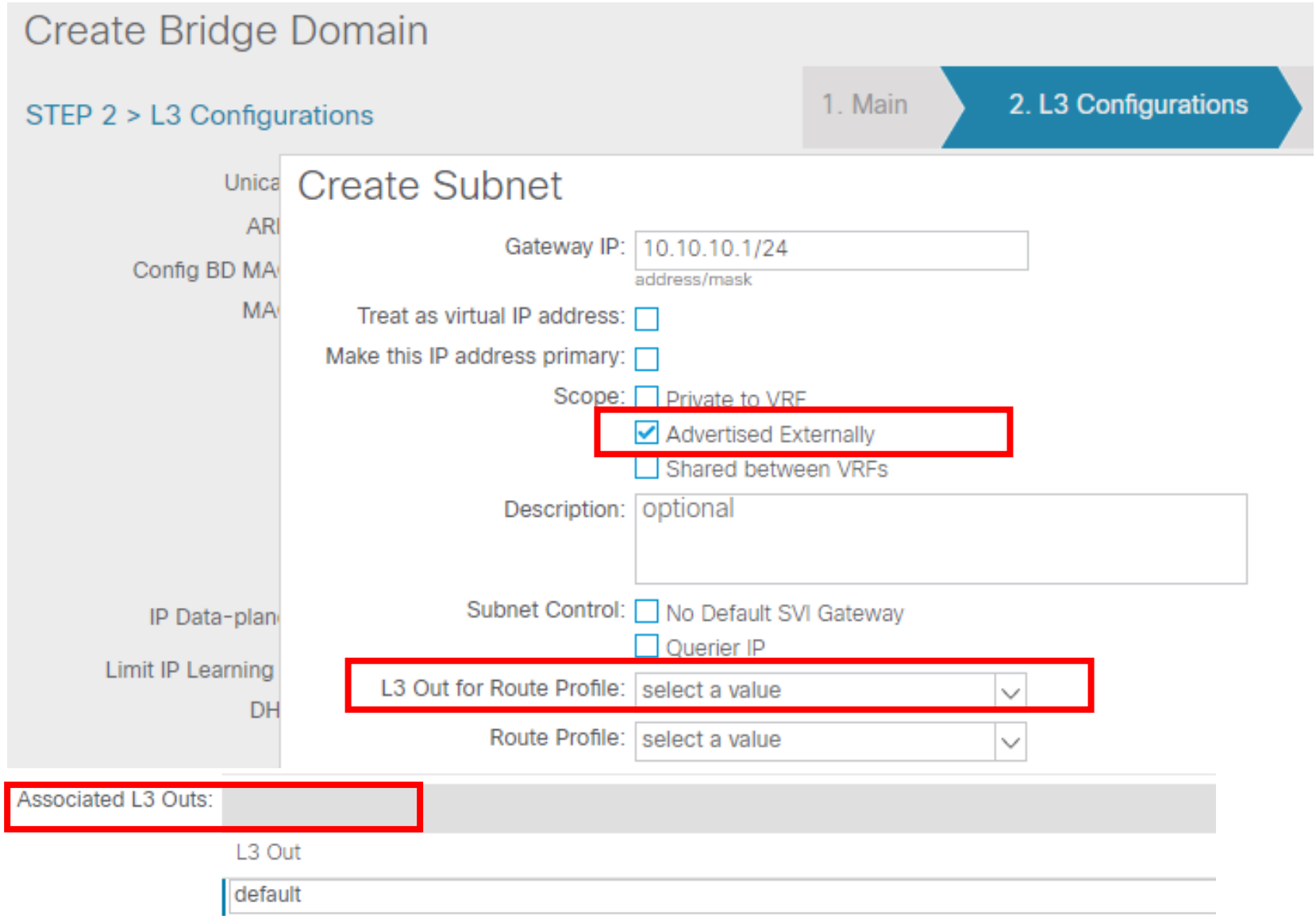
BGP Route Reflector Policy

- MP-BGP is not turned on by default in ACI
- Enabling MP-BGP is as simple as selecting the spines that will operate as BGP Route Reflectors and configuring the Autonomous System Number (ASN)
- The ASN selected for the MP-BGP policy will be the ASN used when connecting the ACI fabric to iBGP neighbors.
- External eBGP neighbors can peer to the MP-BGP ASN or a different ASN using the local-as feature
- iBGP neighbors configured on all leaf switches and route reflectors
- BGP multipath enabled



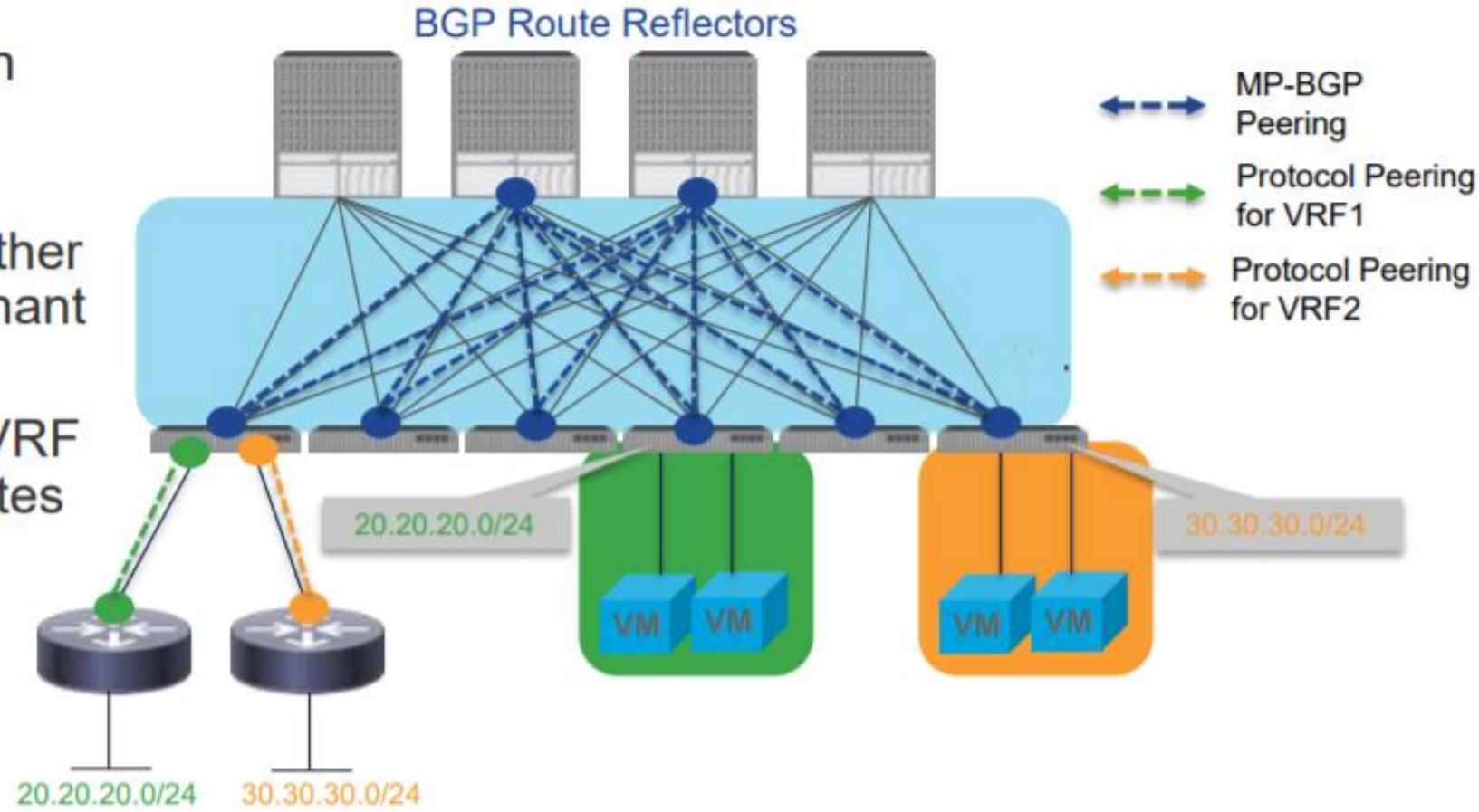
Cisco live!

ACI Tenant subnets advertised externally



Multi-Tenancy Support

- External L3 connectivity is per ACI tenant (VRF)
- Each tenant can have their own external L3 connections
- External prefixes learned are automatically redistributed to other leaves in the fabric on a per tenant (VRF) basis
- Only leaves where the tenant VRF is deployed will receive the routes



Show me ip route!

show ip route vrf <Tenant name>:<VRF name>

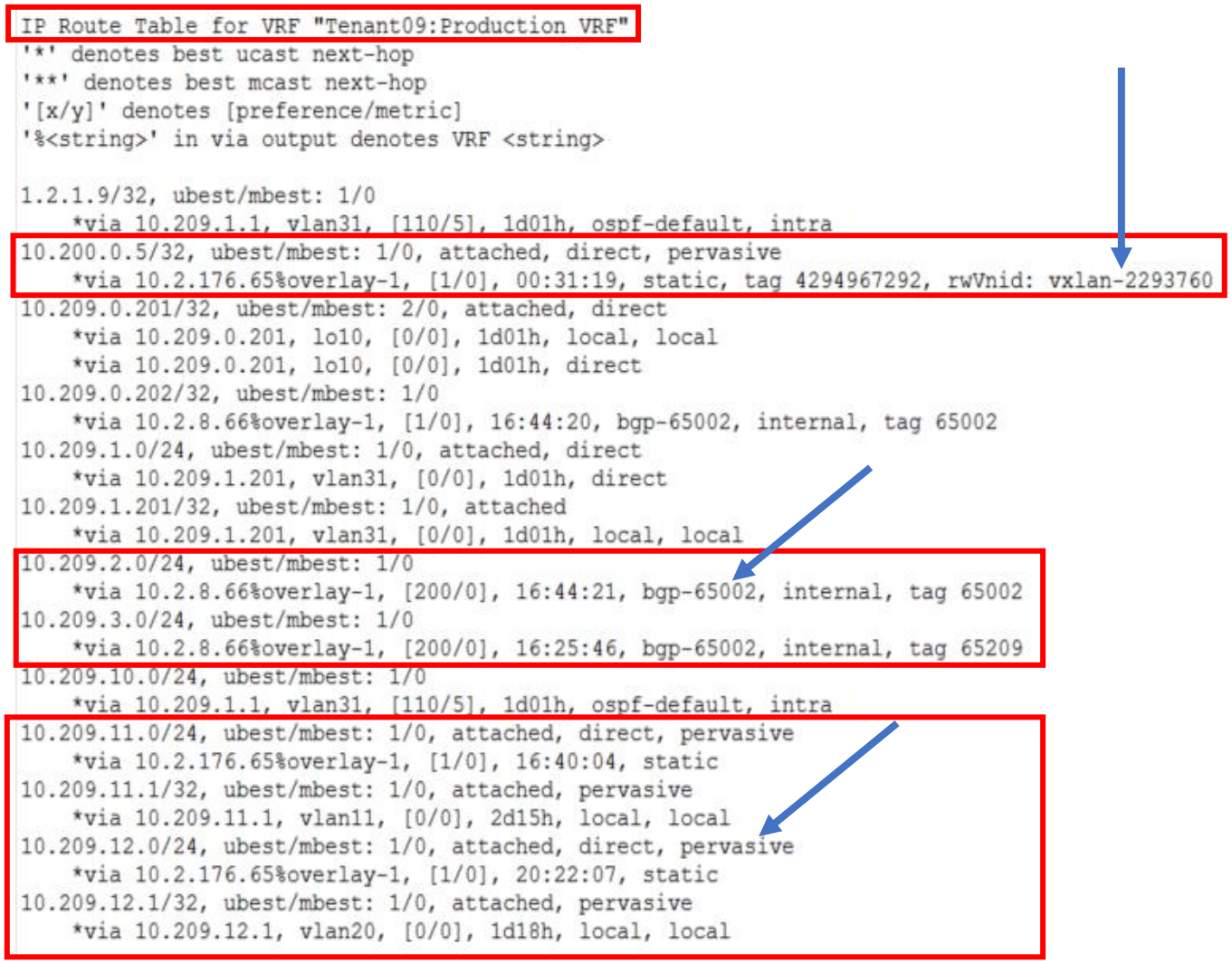
Endpoints via VXLAN tunnel ---------->

External routes receives via BGP ---------->

Note this switch is also border leaf, routes received via OSFP from external router ---------->

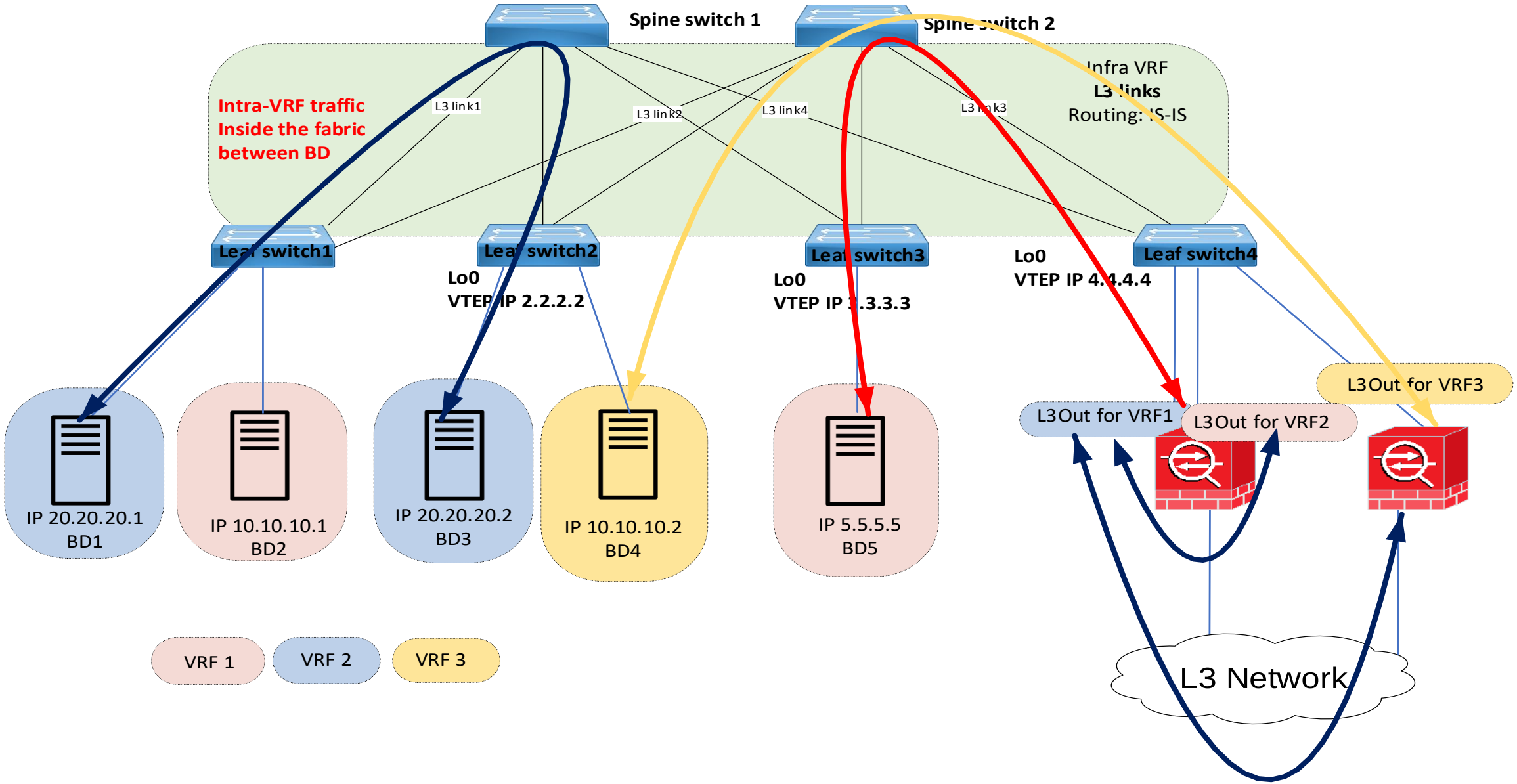
Note pervasive keyword, local SVIs ---------->

No IS-IS routes in tenant networks, only in underlay



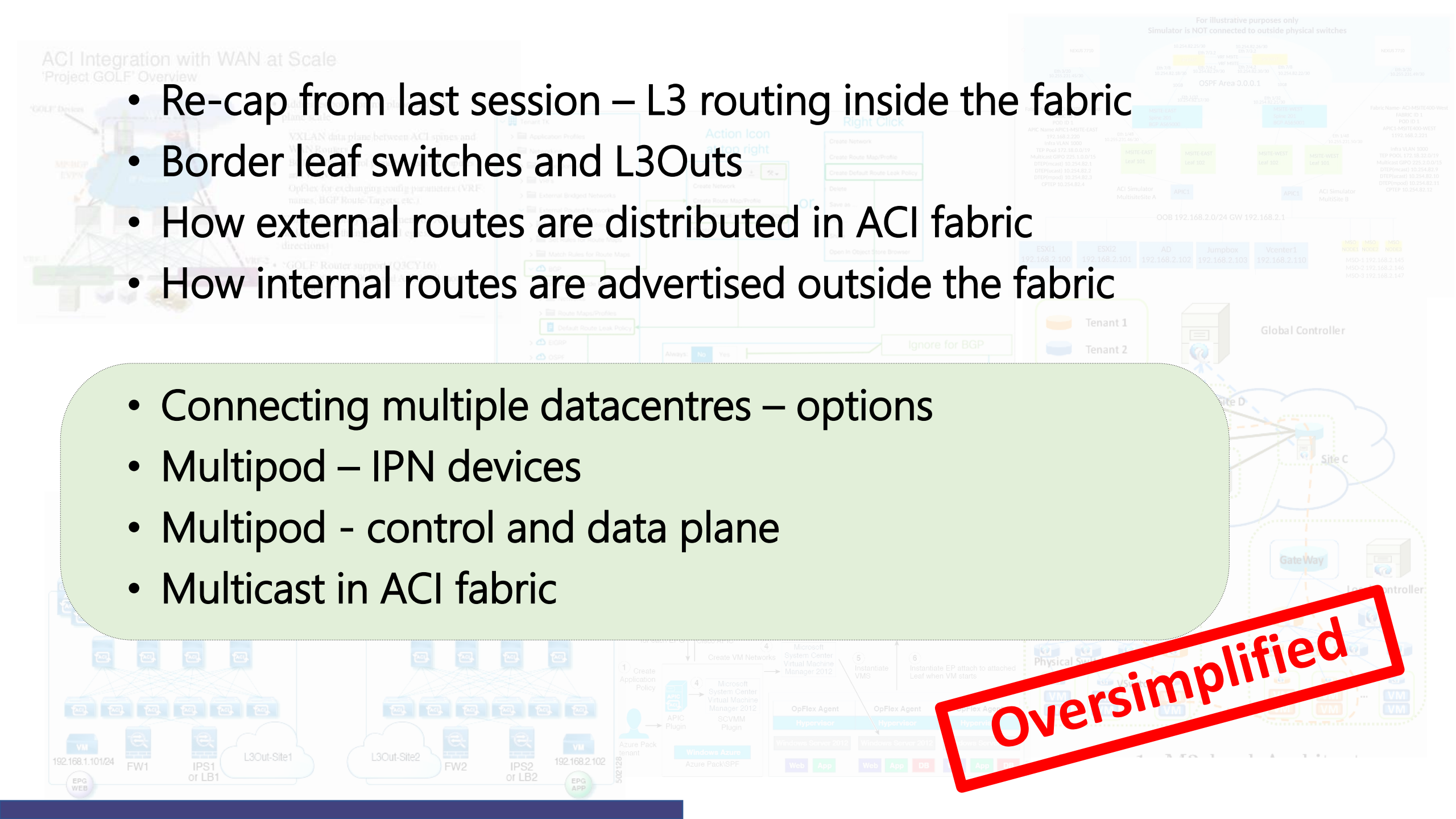
```
IP Route Table for VRF "Tenant09:Production VRF"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

1.2.1.9/32, ubest/mbest: 1/0
    *via 10.209.1.1, vlan31, [110/5], 1d01h, ospf-default, intra
10.200.0.5/32, ubest/mbest: 1/0, attached, direct, pervasive
    *via 10.2.176.65%overlay-1, [1/0], 00:31:19, static, tag 4294967292, rwVnid: vxlan-2293760
10.209.0.201/32, ubest/mbest: 2/0, attached, direct
    *via 10.209.0.201, lo10, [0/0], 1d01h, local, local
    *via 10.209.0.201, lo10, [0/0], 1d01h, direct
10.209.0.202/32, ubest/mbest: 1/0
    *via 10.2.8.66%overlay-1, [1/0], 16:44:20, bgp-65002, internal, tag 65002
10.209.1.0/24, ubest/mbest: 1/0, attached, direct
    *via 10.209.1.201, vlan31, [0/0], 1d01h, direct
10.209.1.201/32, ubest/mbest: 1/0, attached
    *via 10.209.1.201, vlan31, [0/0], 1d01h, local, local
10.209.2.0/24, ubest/mbest: 1/0
    *via 10.2.8.66%overlay-1, [200/0], 16:44:21, bgp-65002, internal, tag 65002
10.209.3.0/24, ubest/mbest: 1/0
    *via 10.2.8.66%overlay-1, [200/0], 16:25:46, bgp-65002, internal, tag 65209
10.209.10.0/24, ubest/mbest: 1/0
    *via 10.209.1.1, vlan31, [110/5], 1d01h, ospf-default, intra
10.209.11.0/24, ubest/mbest: 1/0, attached, direct, pervasive
    *via 10.2.176.65%overlay-1, [1/0], 16:40:04, static
10.209.11.1/32, ubest/mbest: 1/0, attached, pervasive
    *via 10.209.11.1, vlan11, [0/0], 2d15h, local, local
10.209.12.0/24, ubest/mbest: 1/0, attached, direct, pervasive
    *via 10.2.176.65%overlay-1, [1/0], 20:22:07, static
10.209.12.1/32, ubest/mbest: 1/0, attached, pervasive
    *via 10.209.12.1, vlan20, [0/0], 1d18h, local, local
```

Design option 2 – ACI as a big L3 switch

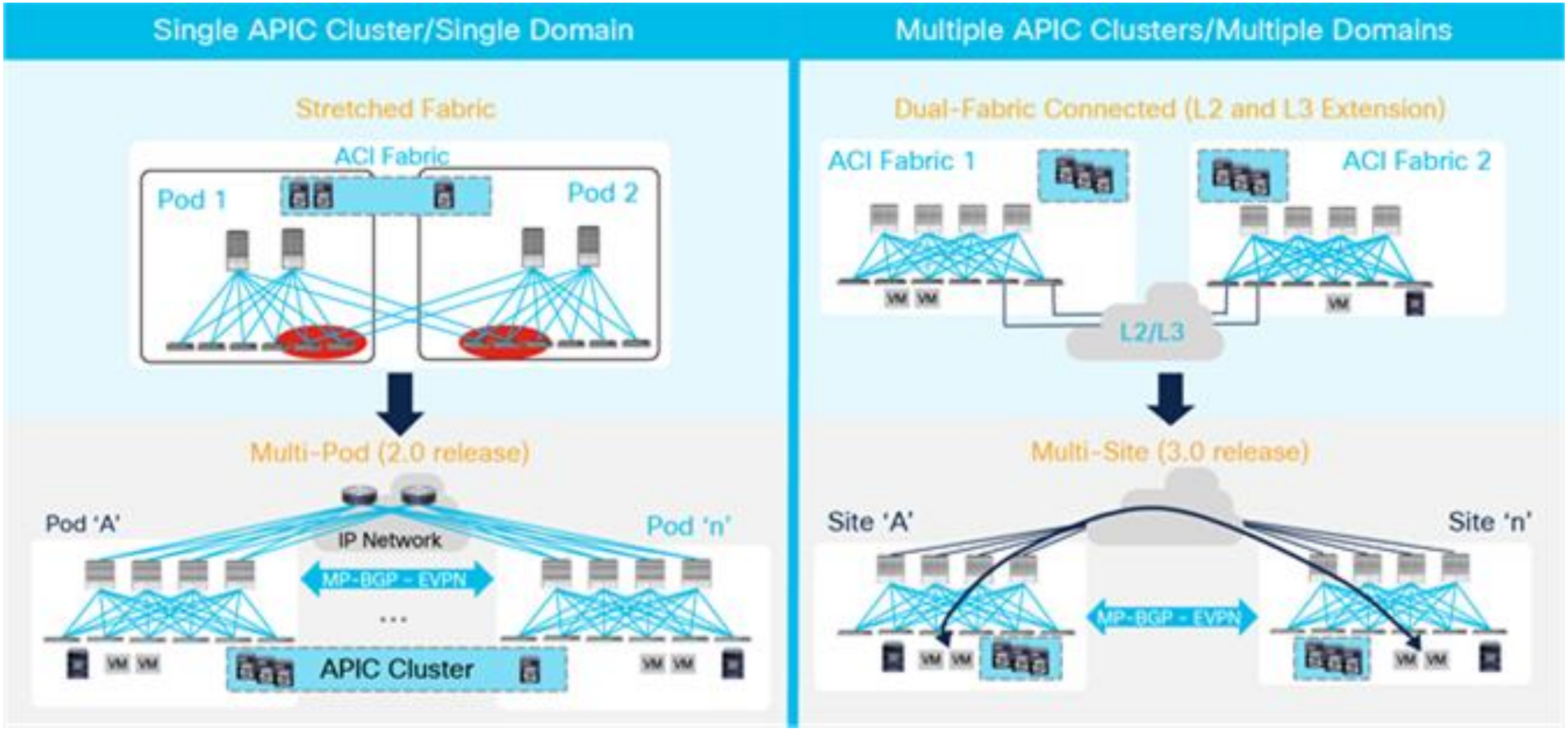


- Re-cap from last session – L3 routing inside the fabric
- Border leaf switches and L3Outs
- How external routes are distributed in ACI fabric
- How internal routes are advertised outside the fabric

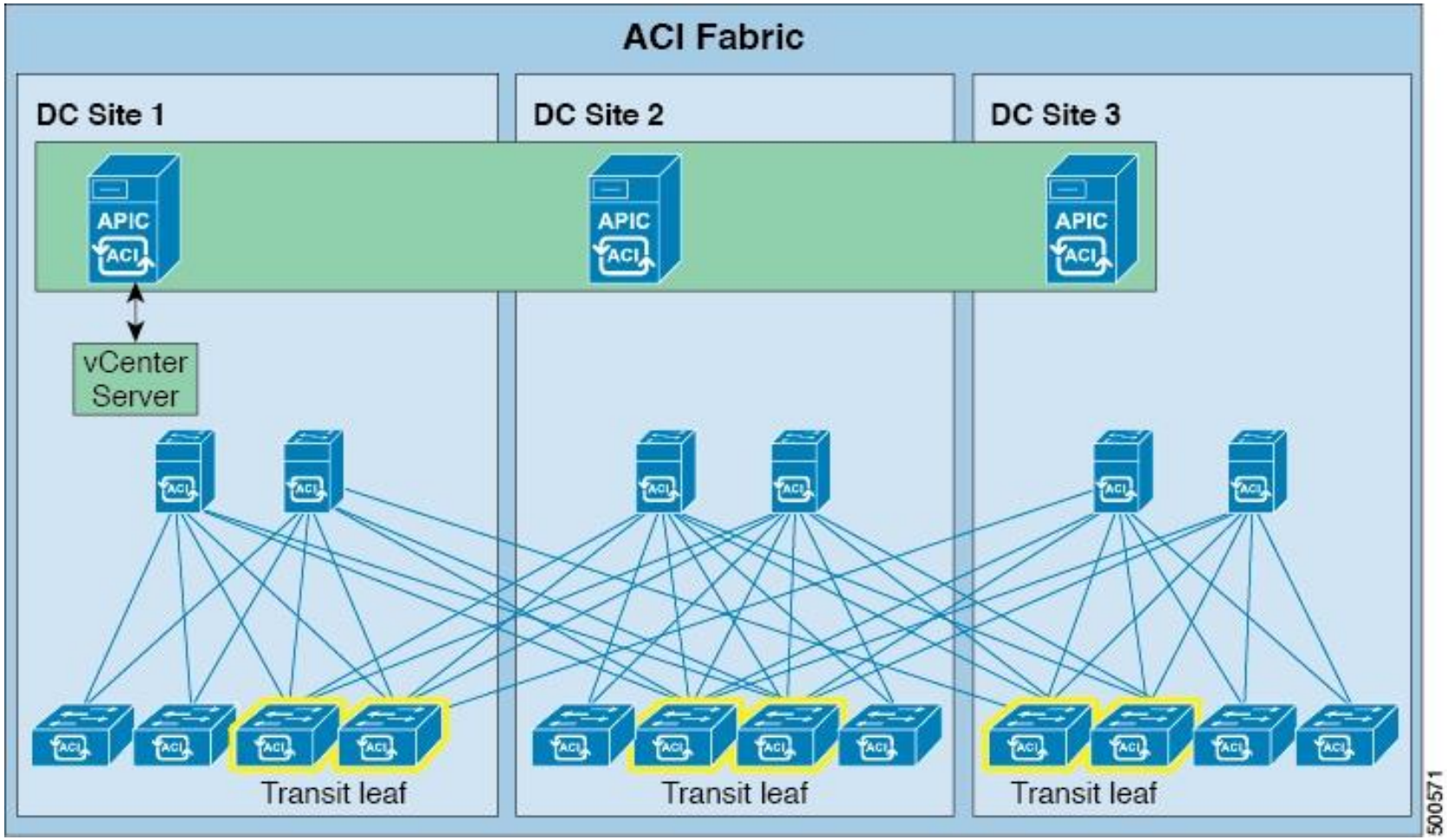
- Connecting multiple datacentres – options
- Multipod – IPN devices
- Multipod - control and data plane
- Multicast in ACI fabric



Connecting multiple DCs



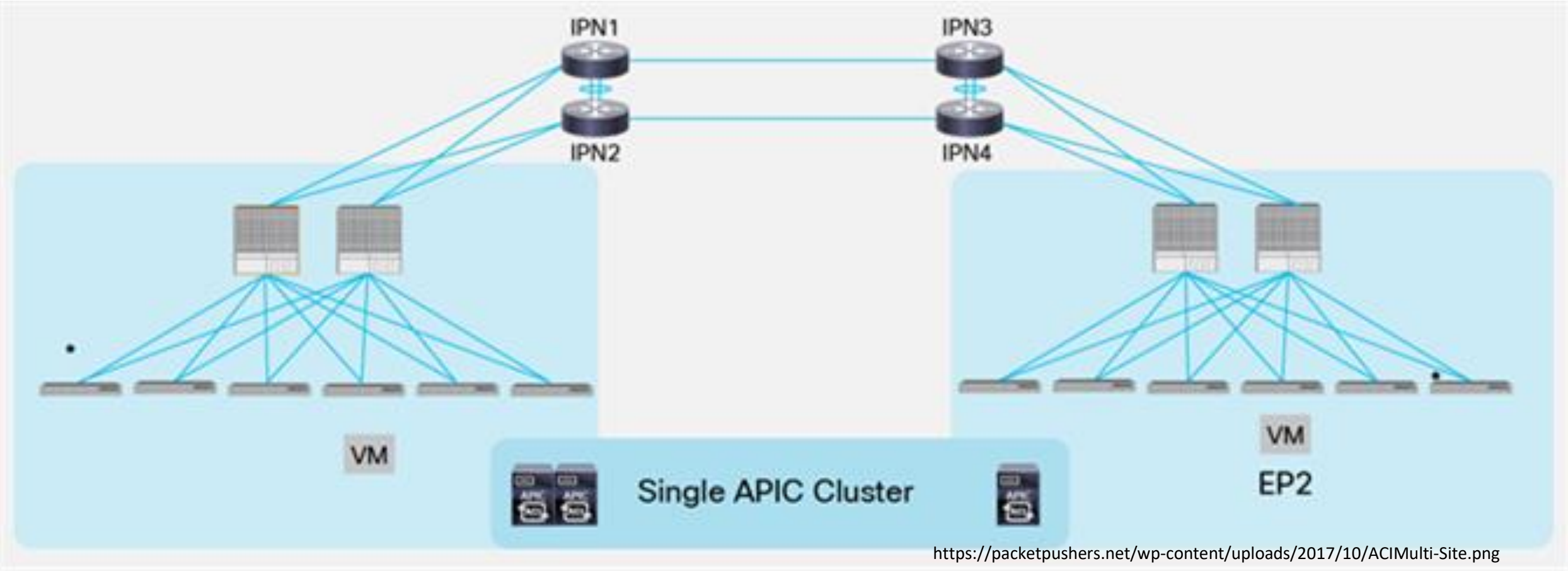
Stretched fabric



Single control plane, single failure domain, requires **full-mesh connectivity between leaf and spines at all sites**

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/kb/b_kb-aci-stretched-fabric.html

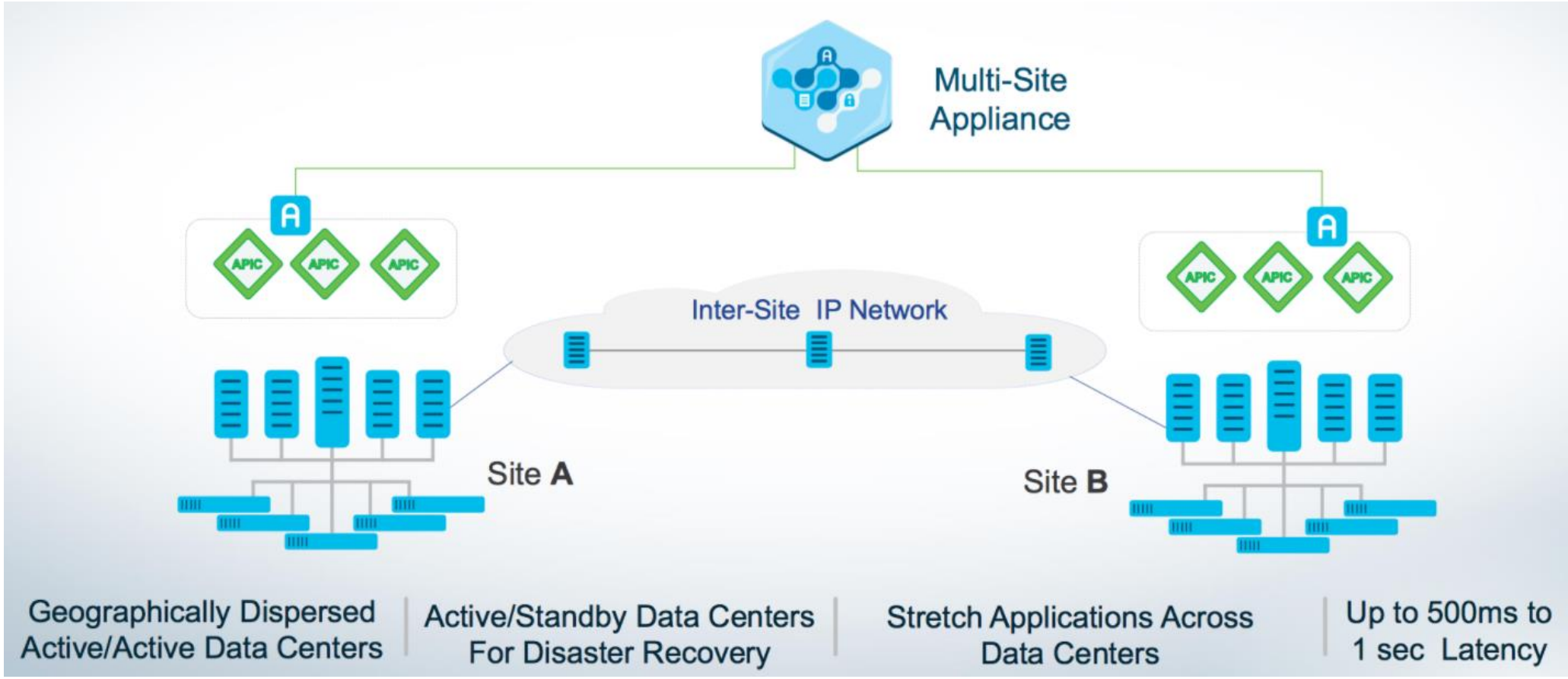
Connecting multiple DCs – Multi-Pod



https://packetpushers.net/wp-content/uploads/2017/10/ACIMulti-Site.png

- Requires additional devices called Inter-Pod Network (**IPN**)
- **Single** APIC cluster
- Pods are separate from control plane perspective, but the same BDs, VNIs, etc.
- Max latency is 50 ms (in old versions 10 ms)

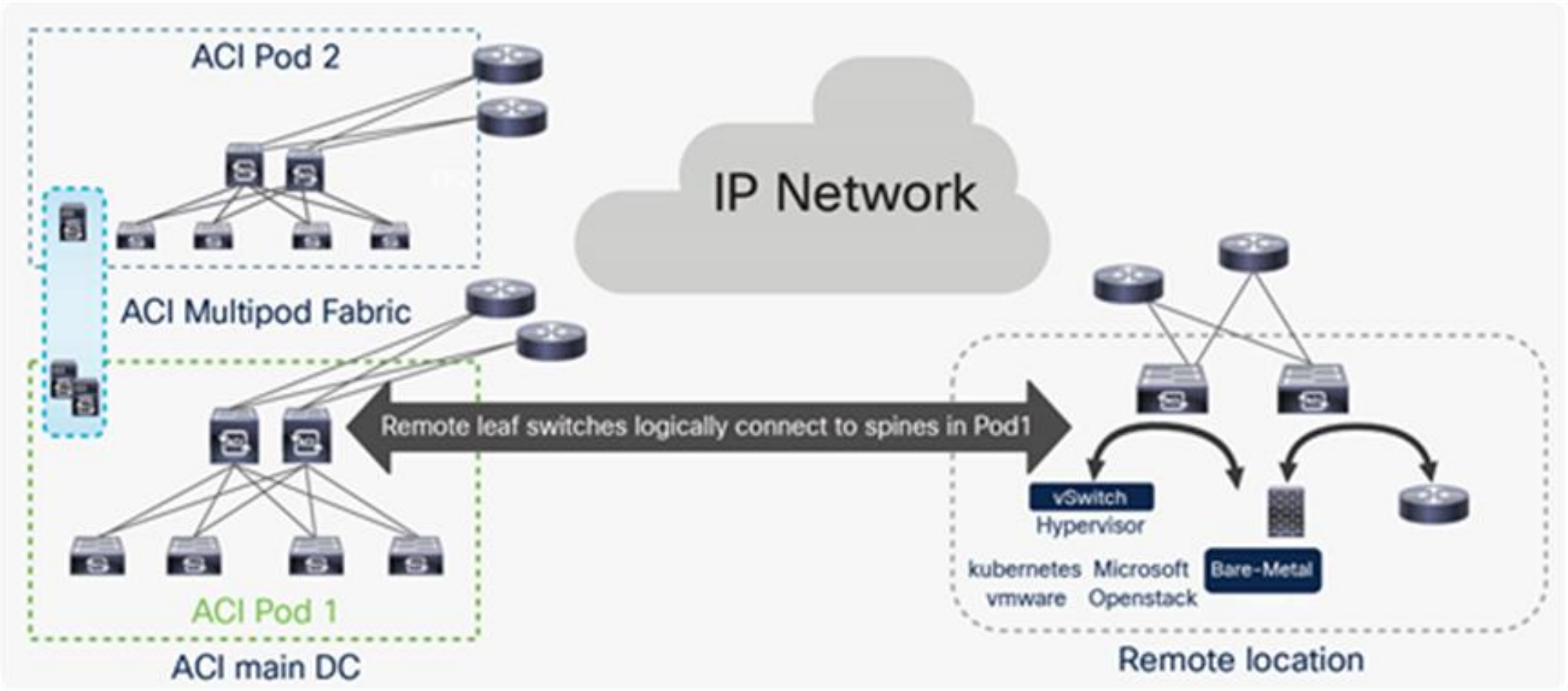
Connecting multiple DCs – Multi-Site



- Requires additional 'controller of controllers' - MSO – Multi-site orchestrator
- **Multiple** APIC clusters
- Pods are separate from control plane perspective, with different BDs, VNIs, but MSO 'converts' different IDs on different sites
- Requires higher-level licenses on switches (Advantage)

https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-739609.html

Remote site under ACI control – Remote Leaf



- Remote site requires physical leaf Nexuses logically connected to Spines
- Up to 300 ms latency to remote site and min 100Mbps, Transport MTU – 9150 bytes
- Remote location becomes a part of DC

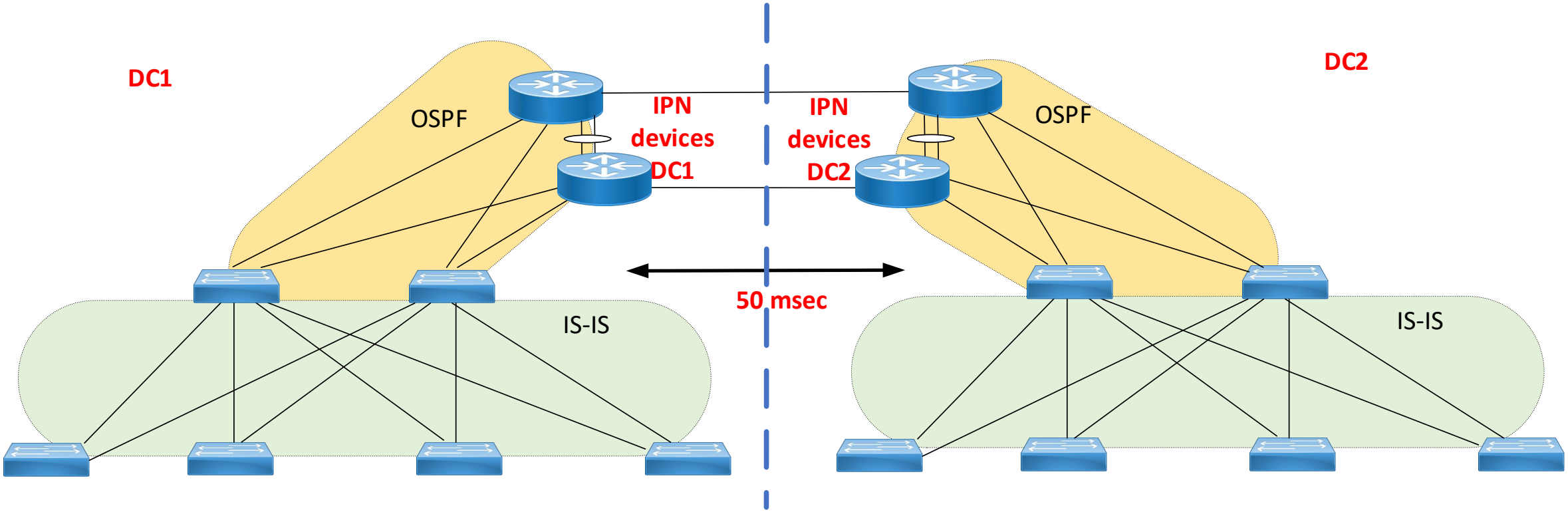
https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-740861.html#IPNetworkIPNrequirementsforRemoteleaf

- Re-cap from last session – L3 routing inside the fabric
- Border leaf switches and L3Outs
- How external routes are distributed in ACI fabric
- How internal routes are advertised outside the fabric

- Connecting multiple datacentres – options
- Multipod – IPN devices
- Multipod - control and data plane
- Why multicast in ACI fabric

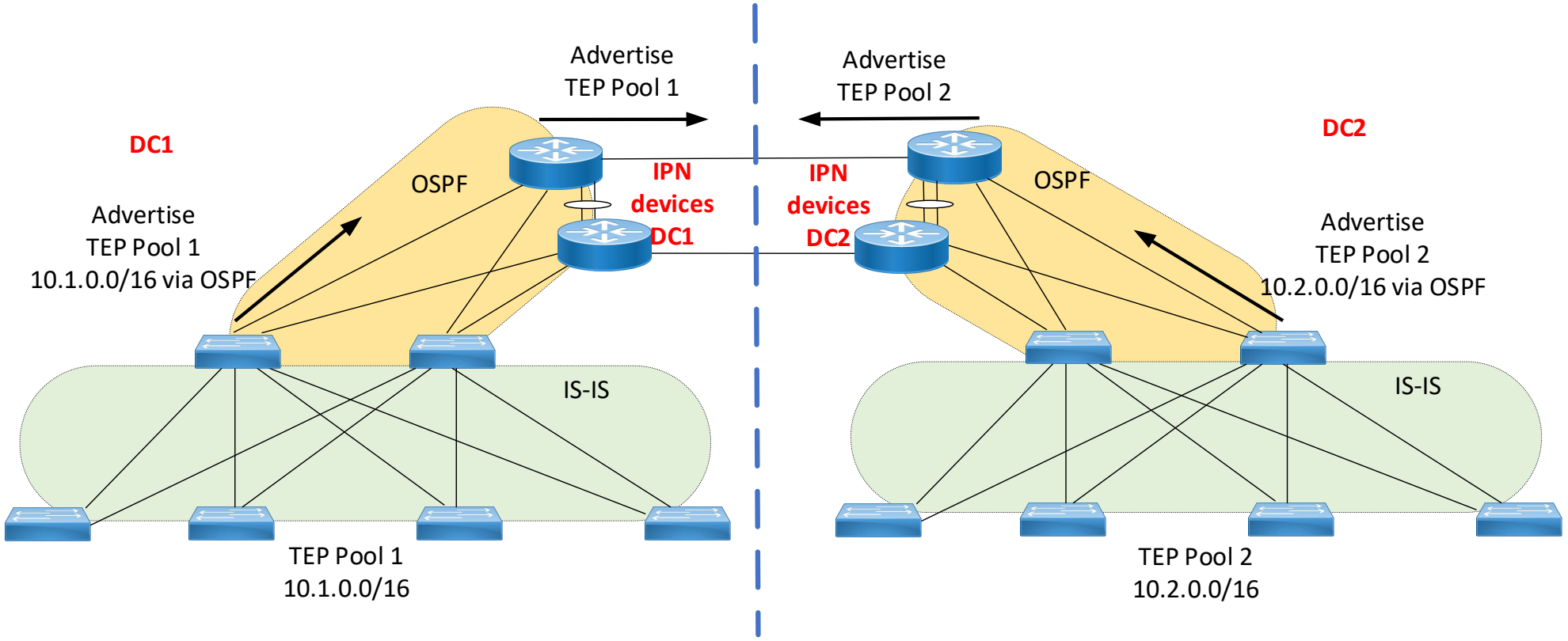
Oversimplified

Multi-Pod – IPN devices



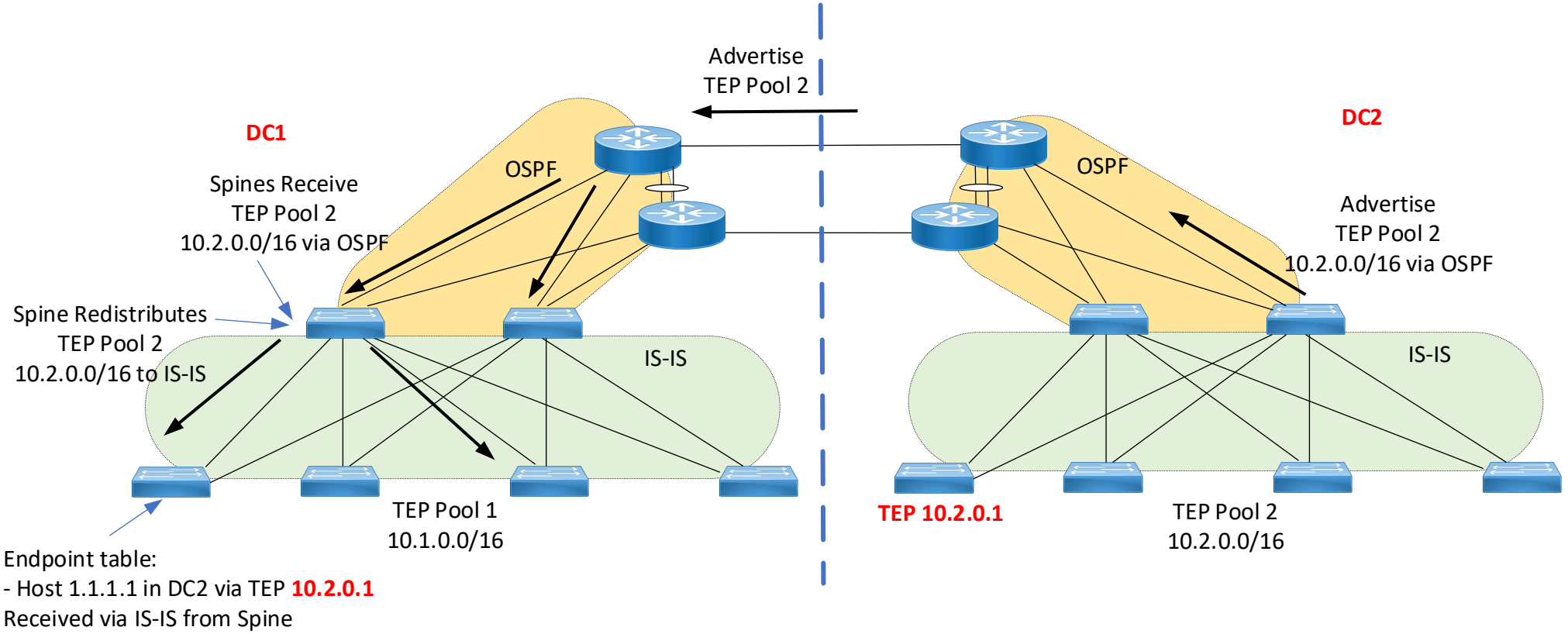
- IPN (Inter-Pod Networking) devices are not a part of the fabric, but external devices
- IPN devices are connected to Spines and form OSPF neighborship with them (OSPF is the only supported protocol)
- IPN to Spines require full-mesh connectivity
- IPNs on the same site will need connectivity between them or full mesh with other site's IPNs
- Max 50 msec latency between the sites

Multi-Pod – Different addresses for TEP pools



- Both sites have different IP address pools for TEP – Leaf switches
- These pools are advertised by Spines via OSPF to IPN devices
- IPN devices advertise them to each other **over any routing protocol**

Multi-Pod – Reachability of TEP addresses via IPN



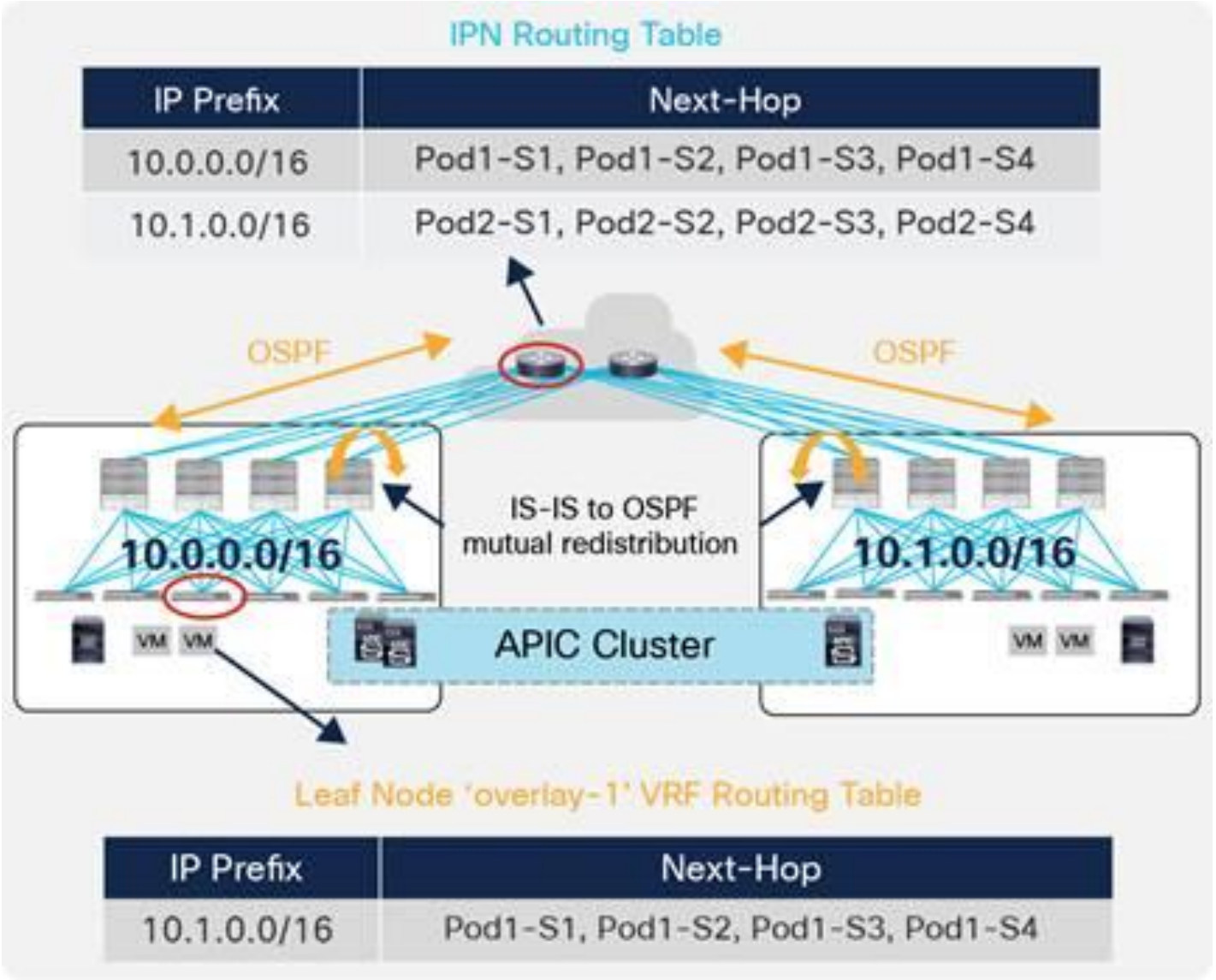
- **IPN devices** receive TEP pool prefix from the other site, and advertise to Spines
- **Spines** redistribute this prefix to IS-IS
- **Leaves** receive this prefix and use for reachability info of leaves hosted in the other DC

OSPF peering with IPN is used to exchange TEP IP addresses (underlay)

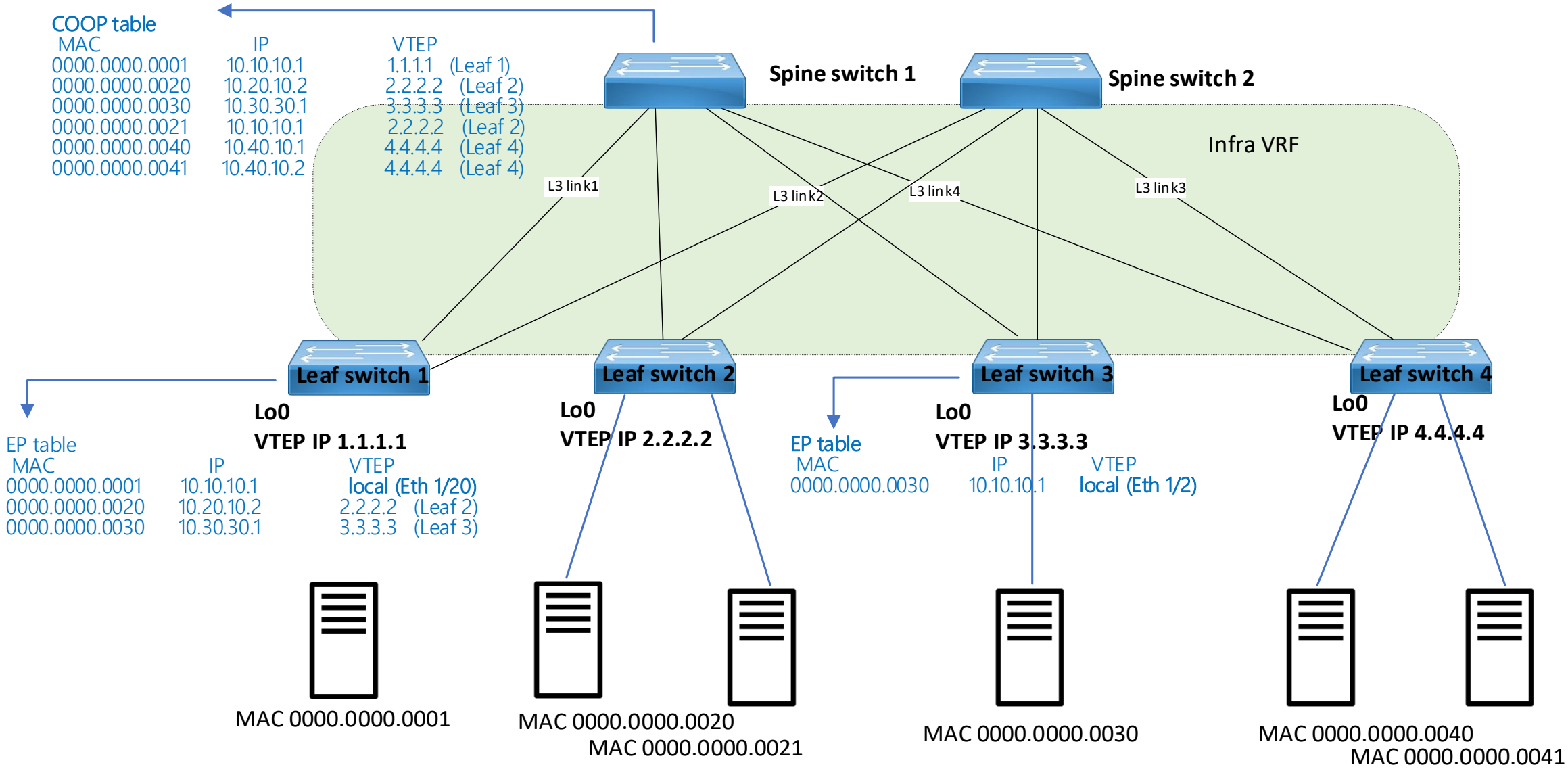
Multi-Pod – Cisco Whitepaper

Excellent document

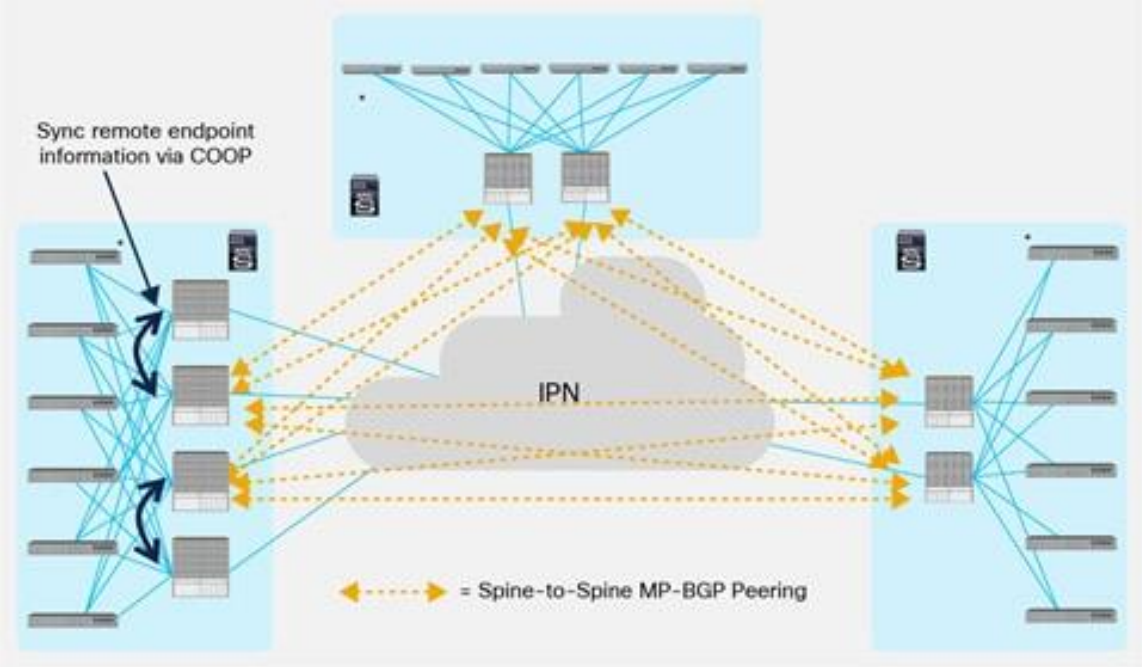
https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-737855.html

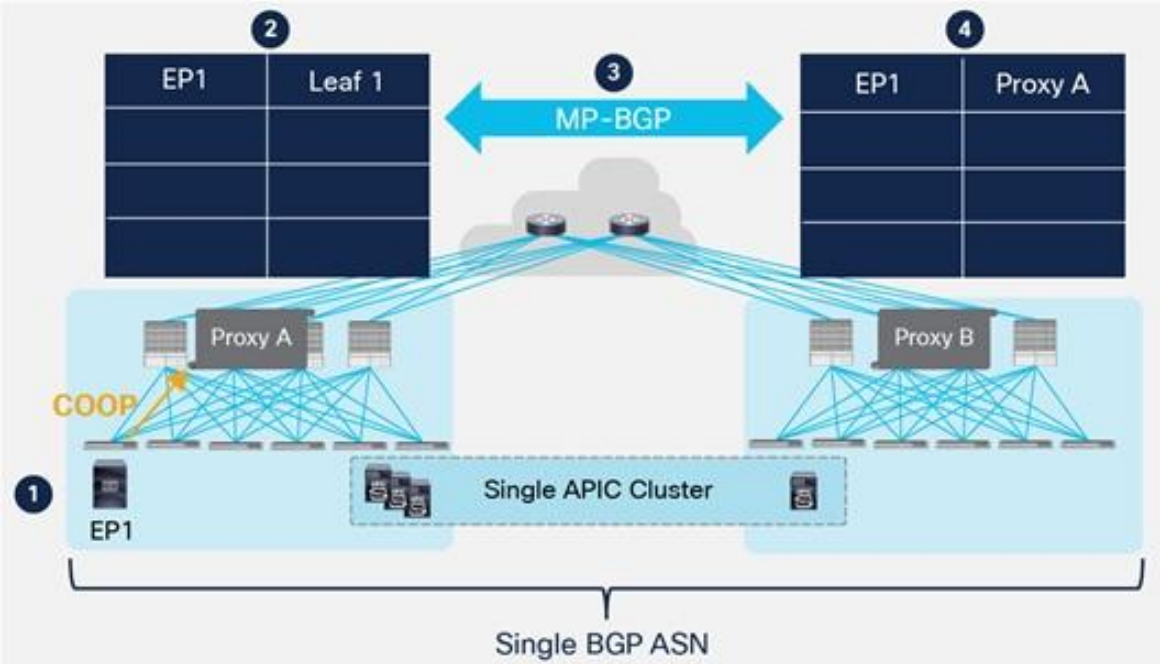


Recap - Endpoint databases in local fabric



Multi-Pod – Reachability of Endpoint addresses



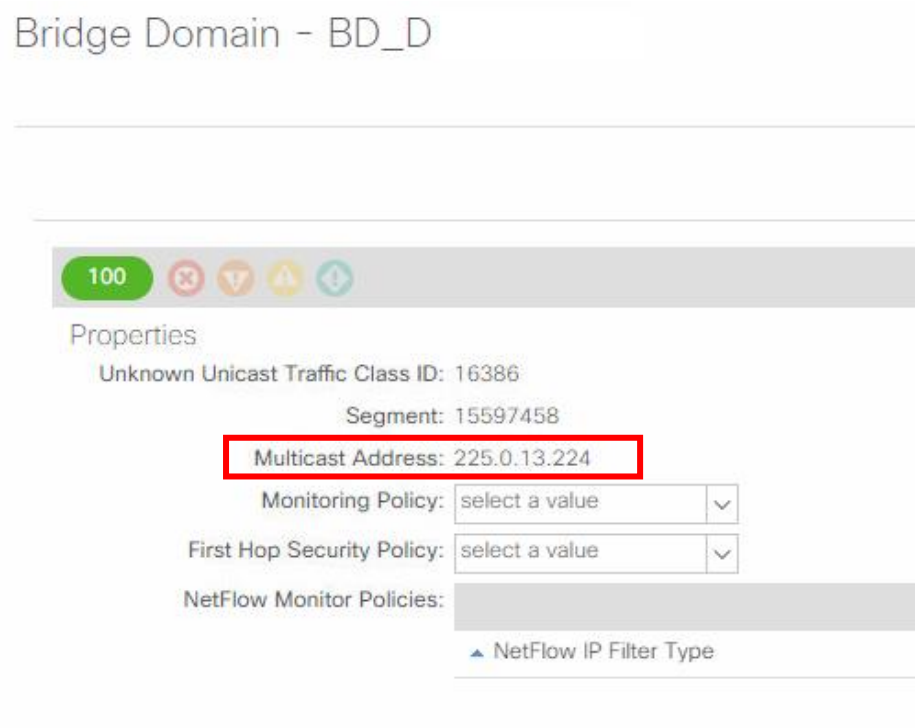


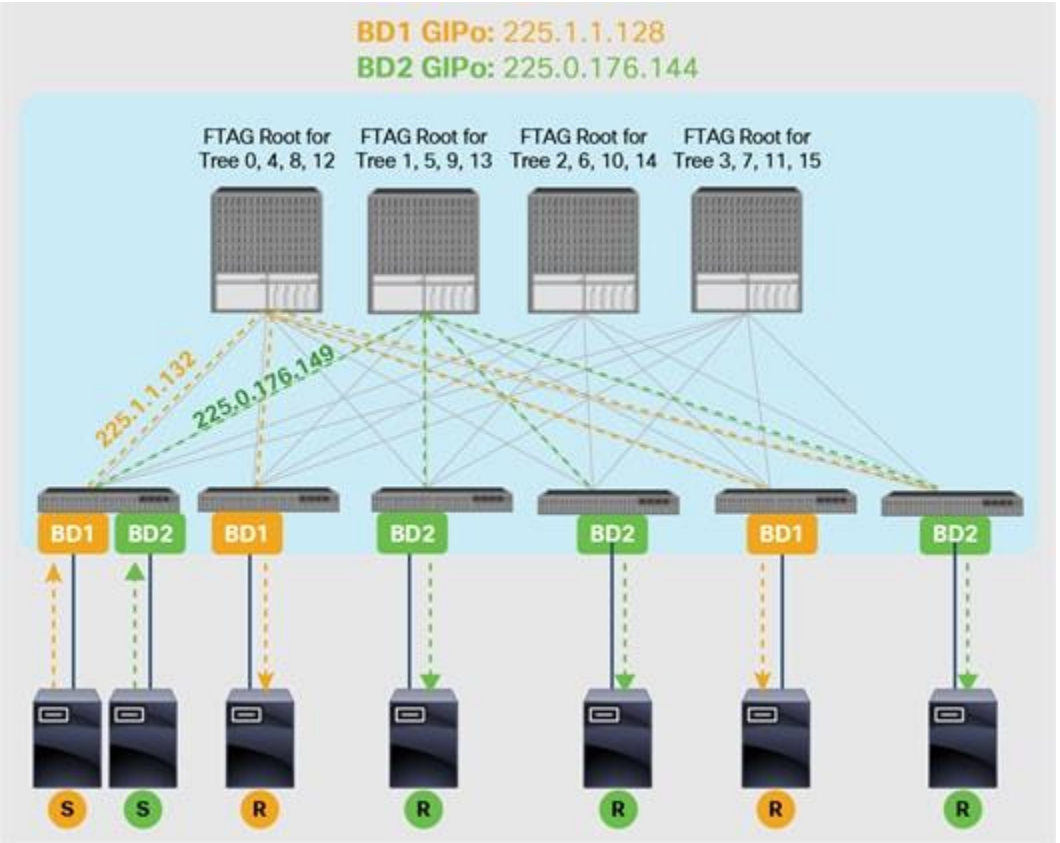
- Spines establish iBGP sessions between other spines on another site
- Leaf switches report their connected endpoints to Spines
- Spines send Endpoint details in MP-BGP EVPN address-family to their BGP peers on the other site
- The receiving Spine adds the information to the COOP
- Result – both sites know each others' Endpoints

- Re-cap from last session – L3 routing inside the fabric
- Border leaf switches and L3Outs
- How external routes are distributed in ACI fabric
- How internal routes are advertised outside the fabric

- Connecting multiple datacentres – options
- Multipod – IPN devices
- Multipod – control and data plane
- Why multicast in ACI fabric

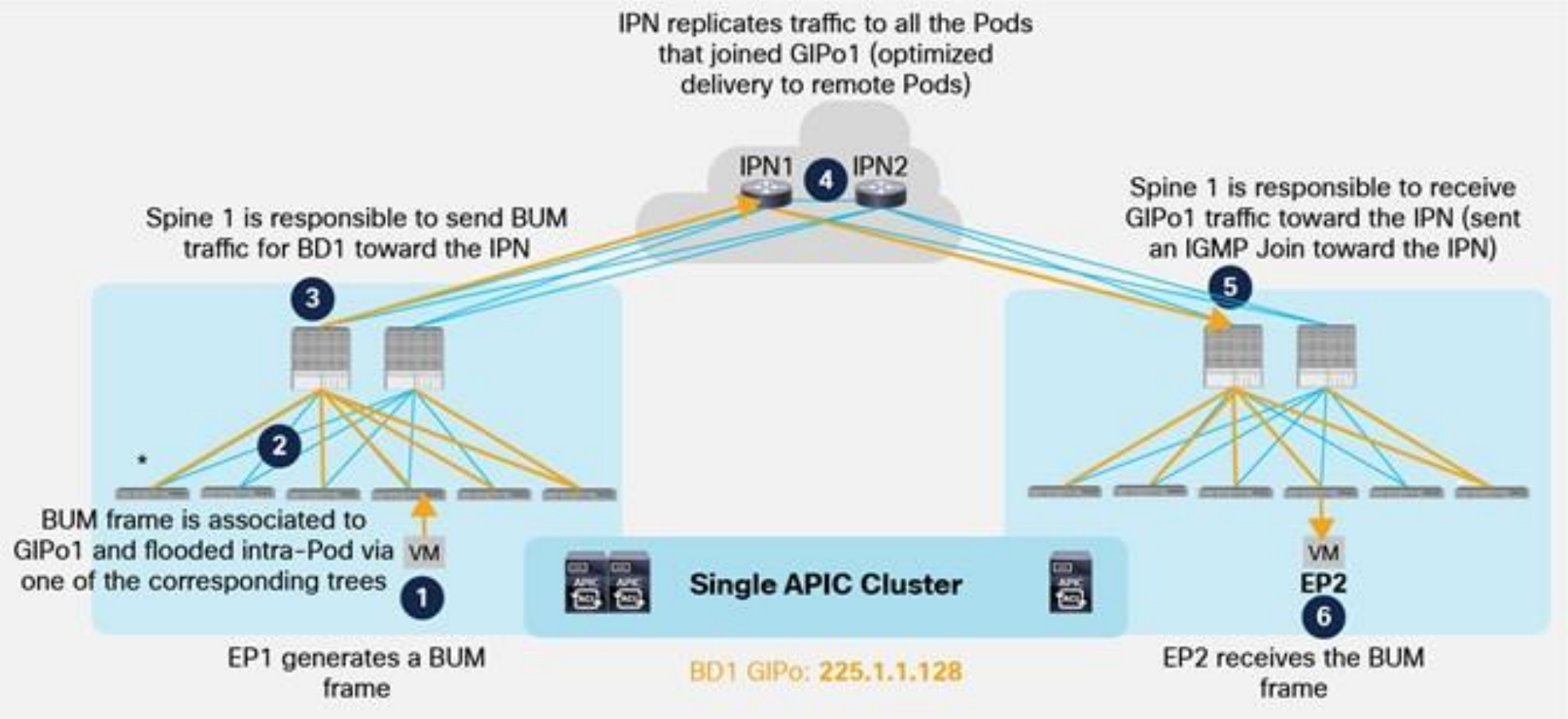
Multicast in ACI - BUM





- Broadcast, Unknown Unicast and Multicast (BUM) tenant traffic can be propagated by encapsulating it into VXLAN packets addressed to **a multicast group** (instead of broadcasting this traffic)
- A unique multicast group is associated to each defined Bridge Domain and takes the name of Bridge Domain Group IP–outer (BD GIPo).
- Each Bridge Domain has associated a separate multicast group (named 'GIPo') to ensure granular delivery of multi-destination frames only to the endpoints that are part of a given Bridge-Domain.
- When BD is deployed on a leaf switch, the leaf send IGMP join to Spine. Spine is multicast RP

Multicast in Multi-Pod – IPN devices are essential



- Each BD shares the same multicast group
- IPN devices need to support multicast PIM BiDir and configured appropriately
- Spines rely on IPN devices to replicate BUM traffic between different DCs

https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-739714.html

Summary

- **Main definitions:** Border Leaf, L3Out, IPN
- **Control plane protocols in ACI**: MP-BGP, Multicast
- **Endpoint learning** in Multi-Pod
- **Traffic forwarding** in Multi-Pod

Possible topic for next sessions

- Endpoint Groups and Contracts, micro segmentation
- Integration with VMware ESXi
- Policy-based routing
- ACI controllers, main UI sections

Good reading

- Connecting Cisco ACI to Layer 3 Routed Domains

https://www.ciscolive.com/c/dam/r/ciscolive/emea/docs/2016/pdf/LTRACI-3000.pdf

https://rednectar.net/2018/03/01/isis_coop_bgp_and_mp-bgp_in_cisco_aci/

- External Routing with ACI

https://www.ciscopress.com/articles/article.asp?p=2928191

- ACI Multi-Pod White Paper

https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-737855.html#MultiPodOverlayControlandDataPlanes

- Cisco ACI Remote Leaf Architecture White Paper

https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-740861.html

Thanks!