ACI Multipod Configuration and Common Issues

Presenter: Edi Wibowo
Panelist: Linda Wang, John Meng and Stephanie Souvleris
Feb 2018

‘Wisdom is not a product of schooling but of the lifelong attempt to acquire it.’

Albert Einstein

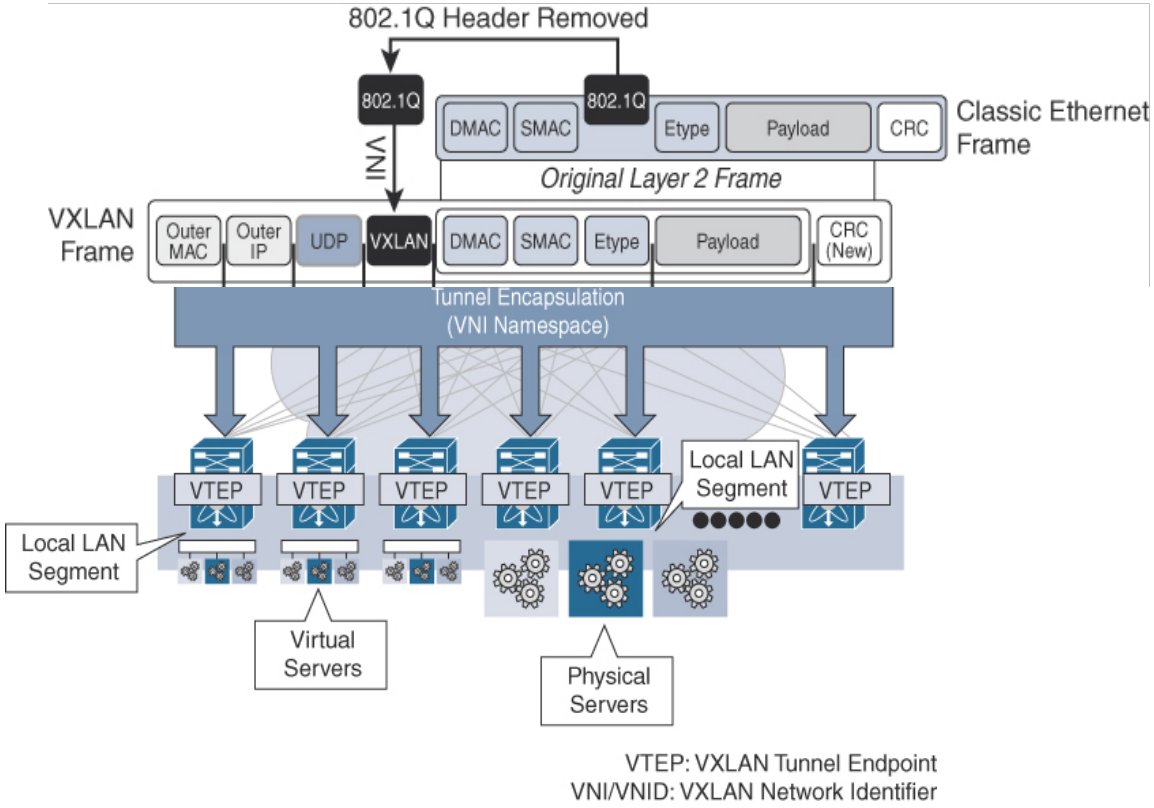
Agenda

- ACI Multipod Solution
- Configuration Overview
- Common Issues

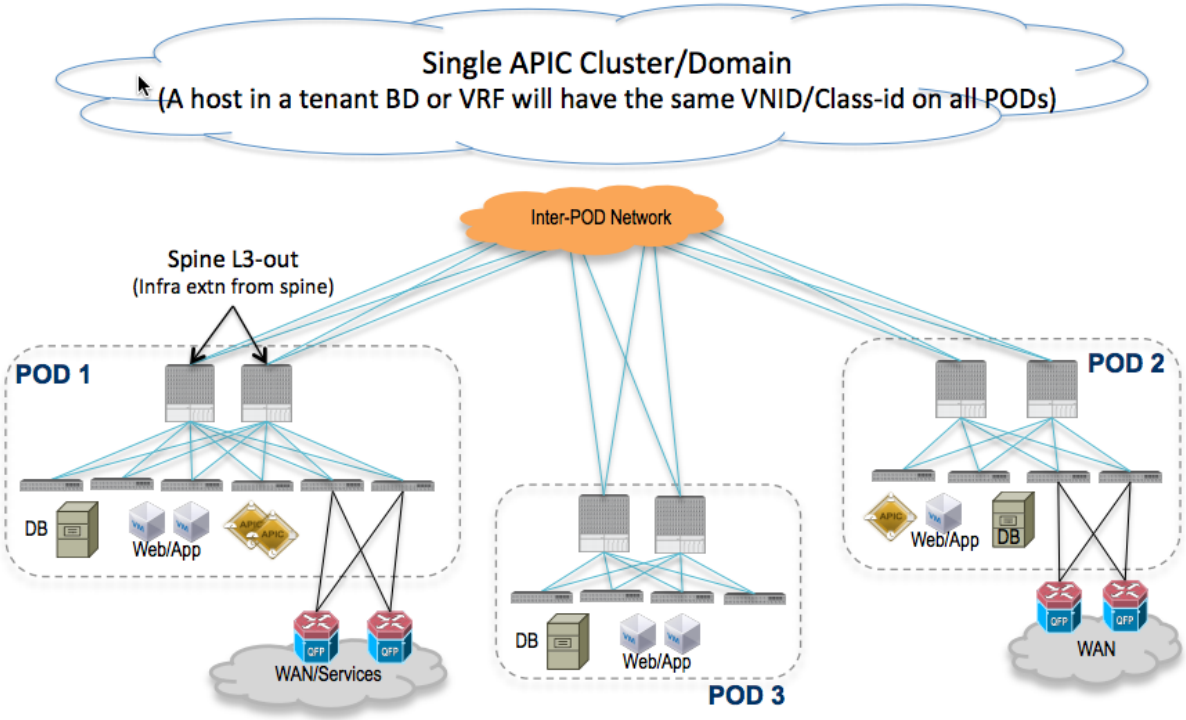
Acronyms

VXLAN	Virtual Extensible LAN
BUM	Layer 2 Broadcast, Unknown unicast, and Multicast
APIC	Application Policy Infrastructure Controller
PIM	Protocol-Independent Multicast is a multicast routing protocols for Internet Protocol networks
MP-BGP	Multi Protocol Border Gateway Protocol
EVPN	Ethernet VPN - MAC addresses to be treated as routes in the BGP table.
ETEP	External Tunnel End Point (Data Plane Forwarding)
CP-ETEP	Control Plane Tunnel End Point (MP-BGP Neighbors)
MTU	Maximum Transmission Unit

VXLAN



Multipod Topology



cisco

Multipod Benefits

- Single APIC Cluster/Single Domain
  A single APIC controller cluster representing the single point of management and policy definition for the entire network, independently from the number of separate ACI fabrics (Pods) compounding it.
- Active/Active
  Data Centers are deployed in multiple Pods, so to offer the freedom of deploying the various application components across separate Pods.

Endpoint Learning Models of VXLAN

- **Flood-and-Learn** (Local Leaf to Local Leaf):
  In this model, end-host information learning and VTEP discovery are both data-plane based, with no control protocol to distribute end-host reachability information among VTEPs.

Inside ACI Fabric POD

- **MP-BGP EVPN** (Local Spine to Remote Spine):
  It provides control-plane learning for end hosts behind remote VTEPs. It uses a unified control plane for both Layer 2 and Layer 3 forwarding in a VXLAN overlay network. Each route carries the BD-VNID and/or VRF-VNID in the label field of EVPN route.

Between PODs via IPN

MP-BGP EVPN Benefits

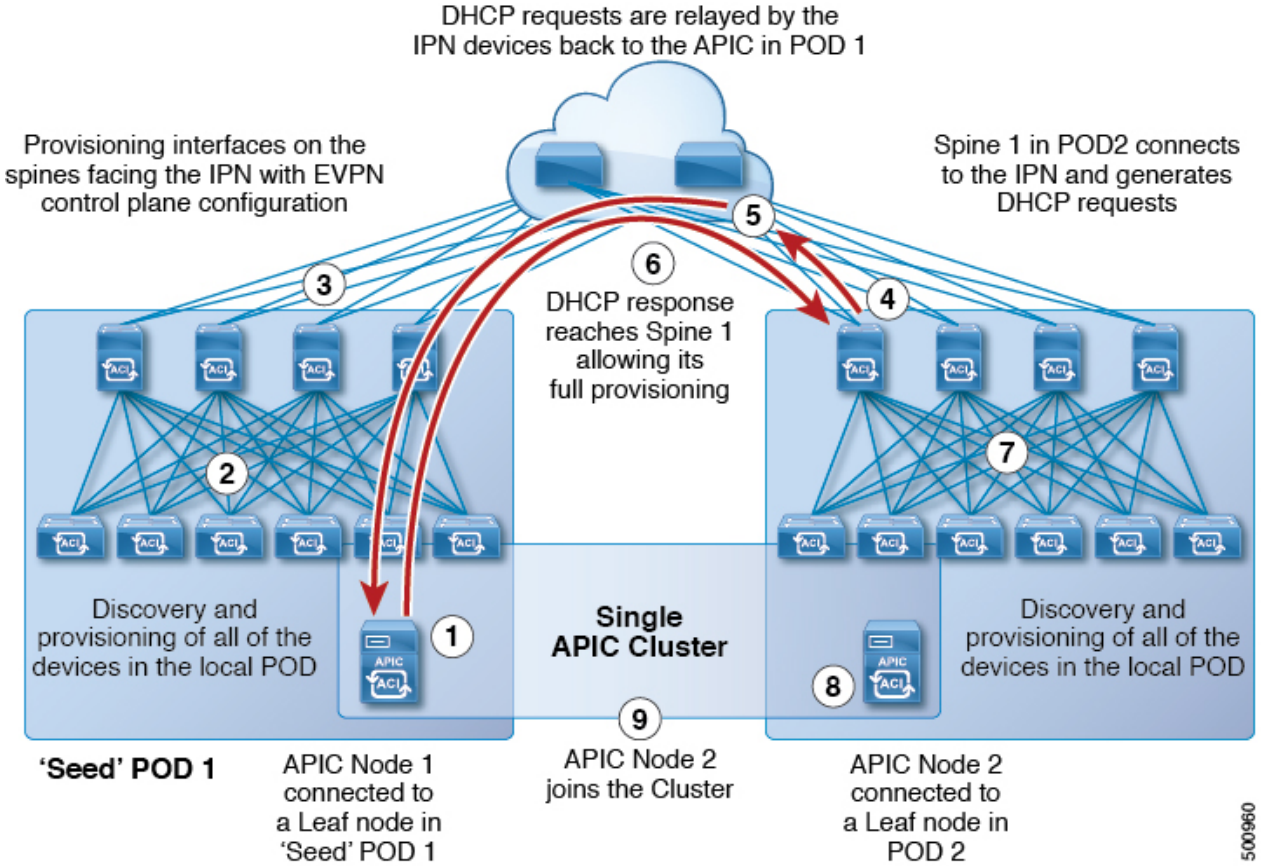
- It uses the very well known MP-BGP VPN technology to support scalable multitenant VXLAN overlay networks.
- The EVPN address family carries both Layer 2 and Layer 3 reachability information, thus providing integrated bridging and routing in VXLAN overlay networks.
- It reduces network flooding through protocol-based host MAC/IP route distribution
- It provides optimal forwarding for east-west and north-south traffic and supports workload mobility with the distributed anycast gateway function.

Agenda

- ACI Multipod Solution
- Configuration Overview
- Common Issues

Configuration Steps

1. On IPN routers: Configure DHCP relay addresses pointing to APICs on POD-1 (Seed POD)
2. On IPN routers: Configure PIM sparse-mode bidir to build multicast groups for overlay/tenant's BUM data traffic
3. On IPN routers and Spines: Configure OSPF
4. On Spines: Configure OSPF and ISIS redistribution
5. On Spines: Configure MP-BGP EVPN



Inter-Pod Network (IPN) Routers

- Fully connected IP network
- 40G/100G ports to connect to spines in all PODs
- IP Multicast – PIM Bidir support – Standalone N9K, N3548 etc
- OSPF protocol for inter-POD reachability
- DHCP Relay to APICs
- Use infra-vlan 4 as sub-interface encap vlan between IPN and spine

IPN Configuration

dp2-p1-ipn-1

```
interface Ethernet1/7
  mtu 9150
  no shutdown

interface Ethernet1/7.4
  description dp2-p1-s1
  mtu 9150
  encapsulation dot1q 4
  vrf member IPN-1
  ip address 192.168.1.1/31
  ip ospf network point-to-point
  ip router ospf IPN area 0.0.0.0
  ip pim sparse-mode
  no shutdown
```

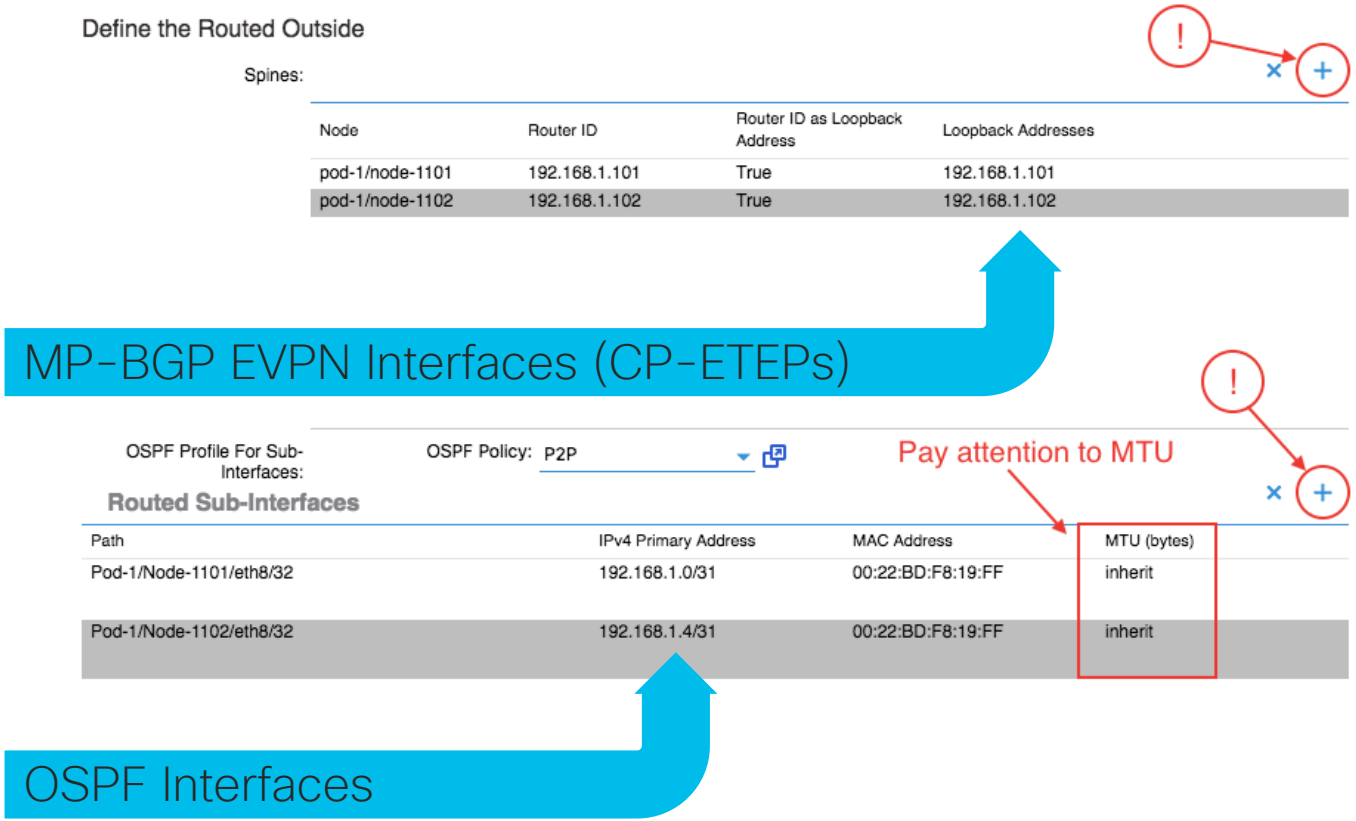
dp2-p2-ipn-1

```
interface Ethernet2/7
  mtu 9150
  no shutdown

interface Ethernet2/7.4
  description dp2-p2-s3
  mtu 9150
  encapsulation dot1q 4
  vrf member IPN-1
  ip address 192.168.2.1/31
  ip ospf network point-to-point
  ip router ospf IPN area 0.0.0.0
  ip pim sparse-mode
  no shutdown
```

```
interface Ethernet 1/7.4
  ip dhcp relay address 10.111.0.1
  ip dhcp relay address 10.111.0.2
  ip dhcp relay address 10.111.0.3
  ip dhcp relay address 10.111.0.4
  ip dhcp relay address 10.111.0.5
```

```
interface Ethernet 2/7.4
  ip dhcp relay address 10.111.0.1
  ip dhcp relay address 10.111.0.2
  ip dhcp relay address 10.111.0.3
  ip dhcp relay address 10.111.0.4
  ip dhcp relay address 10.111.0.5
```

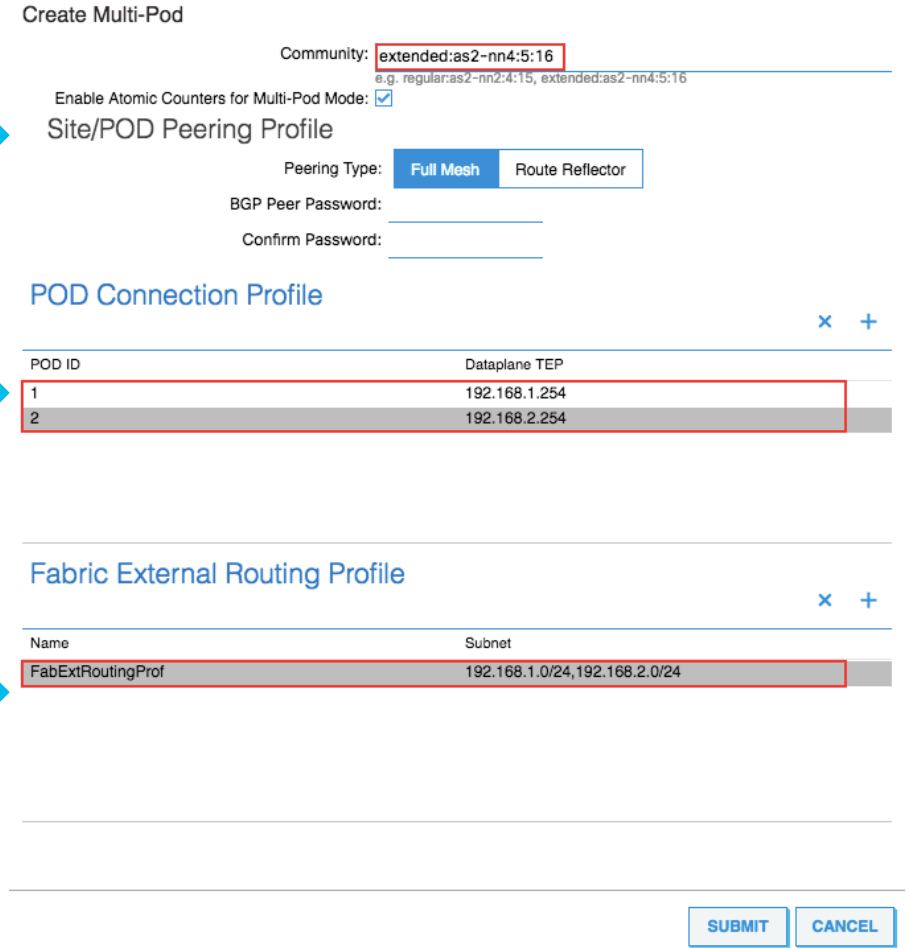
OSPF and MP-BGP Configuration on Spines



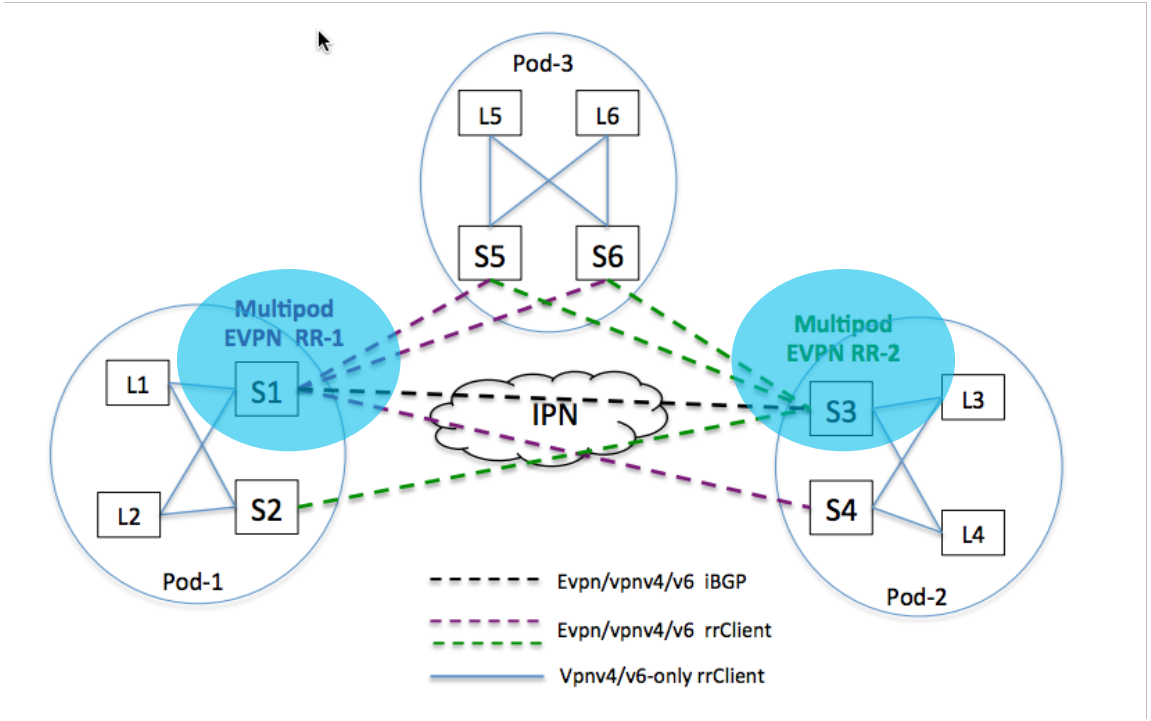
MP-BGP EVPN Full Mesh or RR

External Tunnel End Point (ETEP)

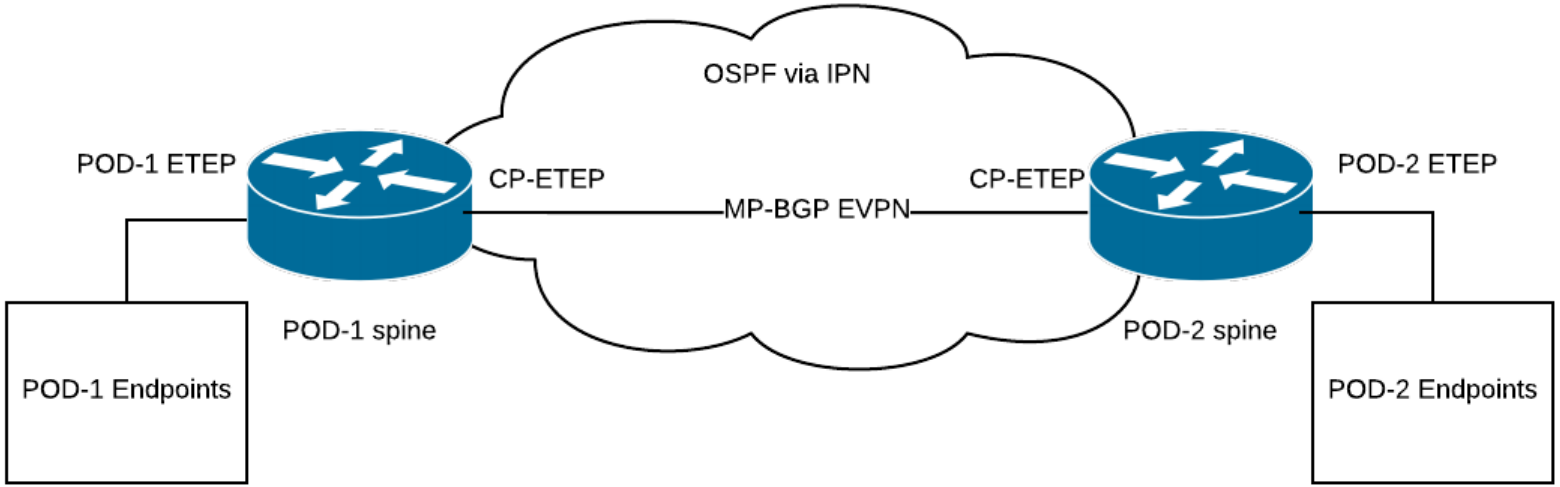
Redistribution into ISIS from OSPF



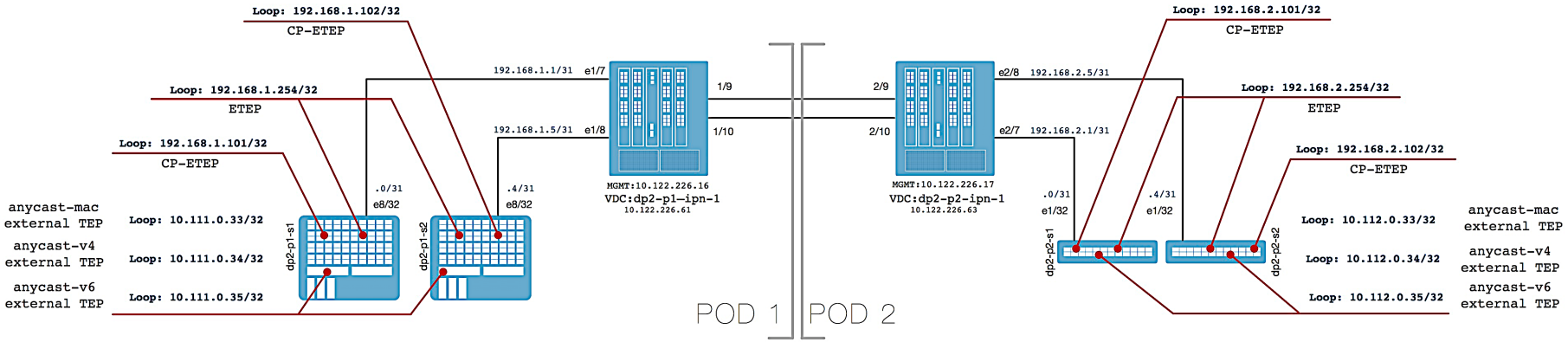
MP-BGP EVPN Route Reflectors (RRs)



Verifying OSPF and BGP (1)



External Tunnel End Point (ETEP)



Verifying OSPF and BGP (2)

```
dp2-p1-ipn1# show ip ospf neighbors vrf IPN-1
 OSPF Process ID IPN VRF IPN-1
 Total number of neighbors: 4
 Neighbor ID     Pri State            Up Time  Address         Interface
 192.168.1.101     1 FULL/ -          02:46:04 192.168.1.0     Eth1/7.4
 192.168.1.102     1 FULL/ -          02:46:02 192.168.1.4     Eth1/8.4
 2.2.2.1           1 FULL/ -          1w6d     192.168.12.1    Po910
```

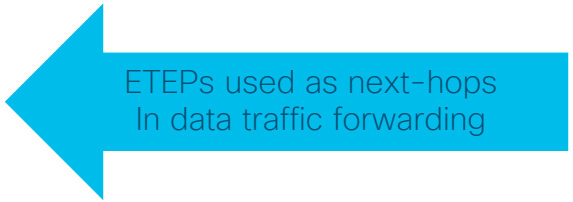
OSPF session used for CP-ETEPs reachability

```
dp2-p1-s1# show bgp l2vpn evpn summary vrf overlay-1
BGP summary information for VRF overlay-1, address family L2VPN EVPN
BGP router identifier 192.168.1.101, local AS number 65000
BGP table version is 806, L2VPN EVPN config peers 2, capable peers 2
57 network entries and 73 paths using 10140 bytes of memory
BGP attribute entries [6/864], BGP AS path entries [0/0]
BGP community entries [0/0], BGP clusterlist entries [0/0]

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
192.168.2.101   4 65000   35723   36745      806    0    0     1w5d 13
192.168.2.102   4 65000   35725   36744      806    0    0     1w5d 13
```

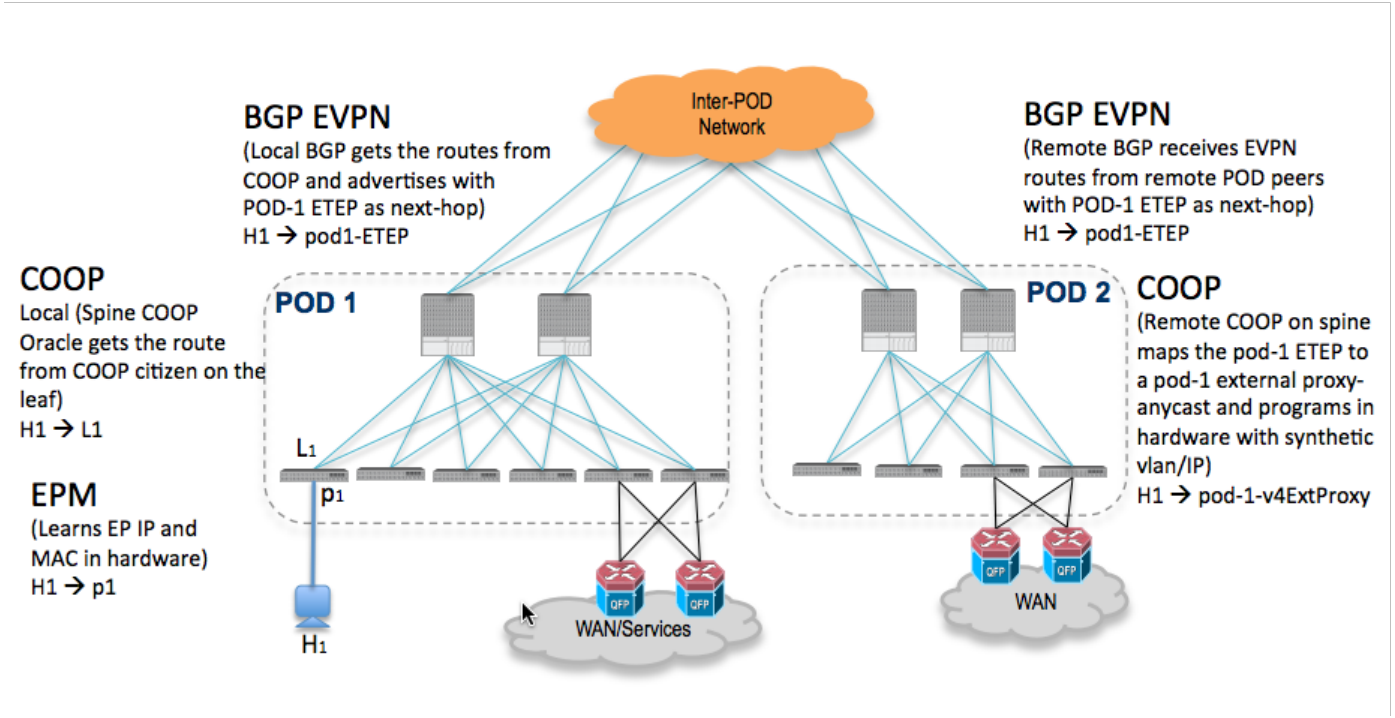
CP-ETEPs used for MP-BGP session

Verifying OSPF and BGP (3)

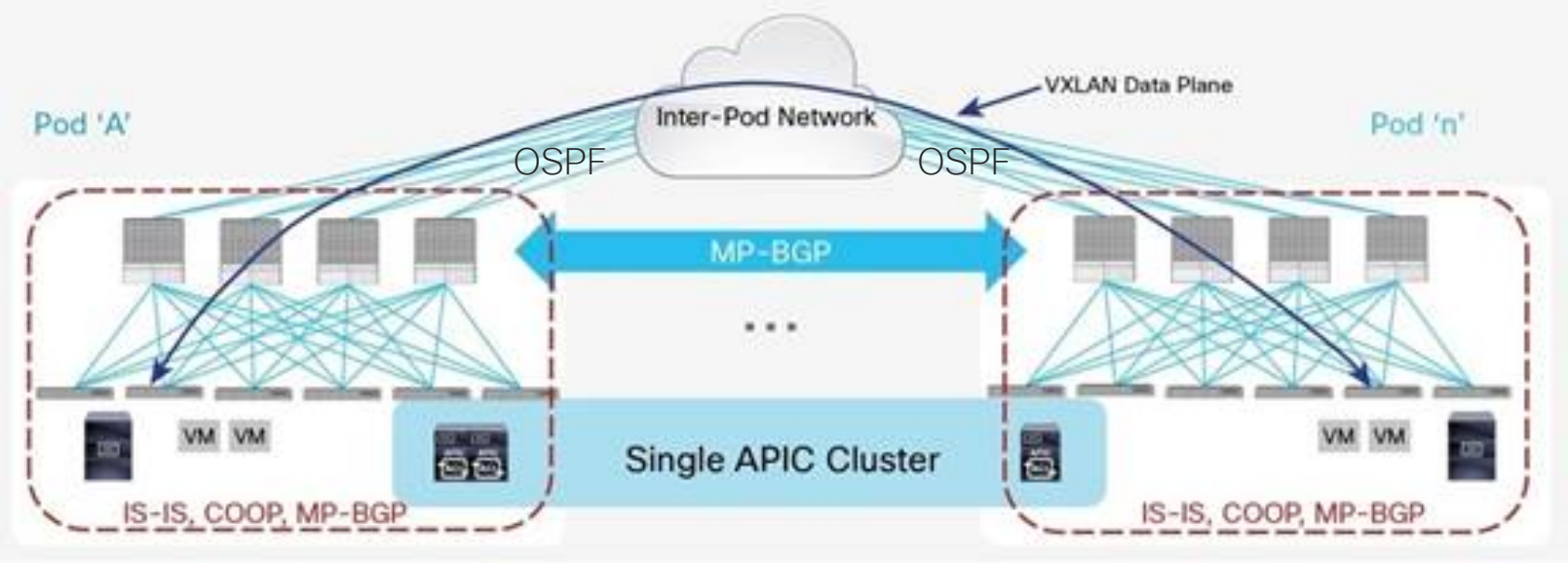
```
dp2-p1-s1# show bgp l2vpn evpn vrf overlay-1
BGP routing table information for VRF overlay-1, address family L2VPN EVPNBGP table version
is 578, local router ID is 192.168.1.101Status: s-suppressed, x-deleted, S-stale, d-
dampened, h-history, *-valid, >-bestPath type: i-internal, e-external, c-confed, l-local,
a-aggregate, r-redist, I-injectedOrigin codes: i - IGP, e - EGP, ? - incomplete, | -
multipath, & - backup

Network            Next Hop            Metric     LocPrf     Weight Path
Route Distinguisher: 192.168.1.254:65003 (L2VNI1)
*>l[2]:[0]:[14680073]:[48]:[dcce.c15b.1e46]:[0]:[0.0.0.0]/216
192.168.1.254                          100      32768
i*>i[2]:[0]:[14712846]:[48]:[dcce.c15b.1e47]:[0]:[0.0.0.0]/216
192.168.2.254                          100          0
i*>l[2]:[0]:[14843893]:[48]:[000c.29dd.0164]:[0]:[0.0.0.0]/216
192.168.1.254                          100      32768
i*>i[2]:[0]:[14843893]:[48]:[003a.7d4e.640c]:[0]:[0.0.0.0]/216
192.168.2.254                          100          0 i
```



Endpoint Route Distribution

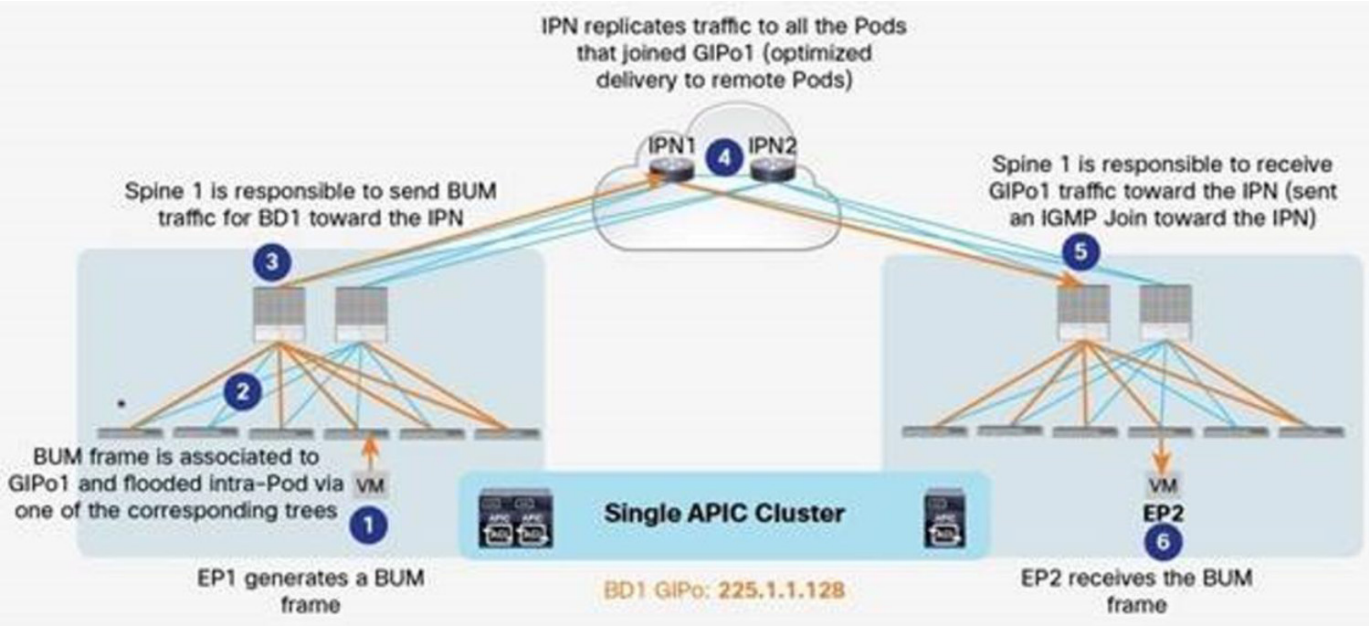


Multipod Building Blocks



Multicast Groups

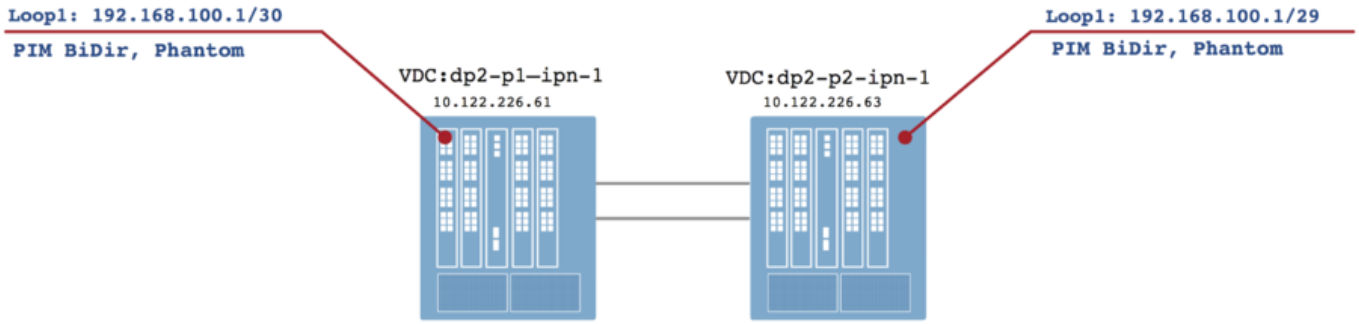
- Tenant's BUM traffic is flooded to a multicast group so that other PODs receive the traffic. The group is advertised to other PODs through BGP EVPN type-6 routes.



Agenda

- ACI Multipod Solution
- Configuration Overview
- Common Issues

Common Issue #1: Phantom RP (1)



dp2-p1-ipn-1

```
interface loopback1
  description BIDIR Phantom RP
  vrf member IPN-1
  ip address 192.168.100.1/30
  ip ospf network point-to-point
  ip router ospf IPN area 0.0.0.0
  ip pim sparse-mode
```

dp2-p2-ipn-1

```
interface loopback1
  description BIDIR Phantom RP
  vrf member IPN-1
  ip address 192.168.100.1/29
  ip ospf network point-to-point
  ip router ospf IPN area 0.0.0.0
  ip pim sparse-mode
```

Common Issue #1: Phantom RP

- Loopback address uses different lengths.
- IGP problem: RP address is not in the routing table of all IPNs.
- Loopback must be “ip ospf network-type point-to-point”
  By default OSPF will advertise this route to loopback as /32 (most specific route to that loopback). To override this we have to change the network type to point-to-point. After this OSPF will advertise the address to loopback as /30 or /29.

Common Issue #2: L3 MTU

Since VXLAN data-plane traffic is MAC-in-IP encapsulation, the IPN must ensure to be able to support an increased MTU on its physical connections, in order to avoid the need for fragmentation and reassembly. The requirement is to increase to 9150 bytes the supported MTU on all the Layer 3 interfaces of the IPN devices.

Common Issue #3: Remote APICs

```
Cluster configuration ...
Enter the fabric name [ACI Fabric1 #1]: dp2-fabric
Enter the fabric ID (1-128) [1]:
Enter the number of controllers in the fabric (1-9) [3]: 5
Enter the POD ID (1-9): 2
Enter the controller ID (1-3) [1]:
Enter the controller name [apic1]: dp2-apic4
Enter address pool for TEP addresses [10.0.0.0/16]: 10.111.0.0/16
Note: The infra VLAN ID should not be used elsewhere in your environment and
should not overlap with any other reserved VLANs on other platforms.
Enter the VLAN ID for infra network (2-4094): 3967
Enter address pool for BD multicast addresses (GIPO) [225.0.0.0/15]:
```

POD ID 2

TEP Pool of POD 1

Common Issue #4: Known Bugs (1)

CSCuz28088

Symptom:
Fabric discovery will fail if ospf auth configured in the spine which is connected to pod-1 via IPN.

Conditions:
Fabric discovery will fail if ospf auth configured in the spine which is connected to pod-1 via IPN.

Workaround:
Disable ospf auth before fabric bringup/spine reboot.

Further Problem Description:
Fabric discovery will fail if ospf auth configured in the spine which is connected to pod-1 via IPN.

Common Issue #4: Known Bugs (2)

CSCvh29461

Symptom:
BGP between spine switches in different pods in a multipod environment goes down and is unable to be re-established. After upgrading to a Cisco APIC 3.1 release, BGP between spine switches in different pods goes down if a QoS CoS translation policy is enabled.

Conditions:
In a multipod environment with spine switches that have generation EX line cards:
1. Upgrade to 3.1(1i)
2. Enable a DSCP CoS translation policy under the infra tenant.
3. Within this policy, set the control plane policy to cs7.

Additionally, if the IPN is marking BGP traffic to DSCP 56, 59, 60, 61, 62, or 63, this condition can be hit as well.

The BGP traffic will not be classified as BGP traffic internally and as a result the spines will not see additional messages received from the peers in the remote pod.

Workaround:
Disable the QoS cos-translation policy or mark / set control-plane traffic to something besides CS7

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/kb/b_Multipod_QoS.html

Thank You

Post your Questions at Community Support

https://supportforums.cisco.com/t5/data-center-documents/tkb-p/4436-docs-data-center


CISCO

Reference

- https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-739714.html
- https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-737855.html
- https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/2-x/L3_config/b_Cisco_APIC_Layer_3_Configuration_Guide/b_Cisco_APIC_Layer_3_Configuration_Guide_chapter_010011.html
- https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/kb/b_Multipod_QoS.html
- Building Data Centers with VXLAN BGP EVPN: A Cisco NX-OS Perspective by David Jansen; Shyam Kapadia; Lukas Krattiger