The bridge to possible

Cisco ACI Multi-Pod

Design and Deployment

John Weston, Technical Marketing Engineer, Cloud Networking

cisco Live!

BRKDCN-2949

Cisco Webex App

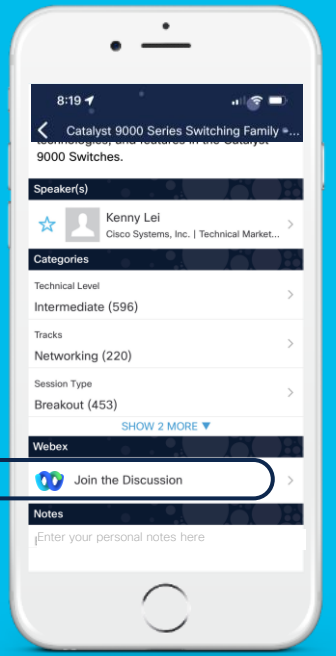
Questions?

Use Cisco Webex App to chat
with the speaker after the session

How

1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install the Webex App or go directly to the Webex space
4. Enter messages/questions in the Webex space

Webex spaces will be moderated until February 24, 2023.



Session Objectives

- At the end of the session, the participants should be able to:
  - Articulate the different deployment options to interconnect Cisco ACI networks (Multi-Pod and Multi-Site) and when to choose one vs. the other
  - Understand the functionalities and specific design considerations associated to the ACI Multi-Pod architecture
- Initial assumption:
  - The audience already has a good knowledge of ACI main concepts (Tenant, BD, EPG, L2Out, L3Out, etc.)

Early Access. Yes, please.

Cisco U.

Tech learning, shaped to you.

cisco

Data Center

ACI Technologies

Take a deep dive into ACI technologies, architecture and troubleshooting.

START

Feb 6 | 08:30
TECDCN-2840
Next Generation ACI Data Center Architecture, Deployment and Operations

Feb 7 | 08:30
BRKDCN-1601
Introduction to ACI

Feb 7 | 11:30
BRKDCN-2906
Introduction to Infrastructure as Code for ACI with Ansible and Terraform

Feb 7 | 14:00
BRKDCN-1688
How to operate your Nexus and ACI networks from the Cloud with Nexus Cloud

Feb 7 | 17:00
BRKDCN-2910
Why You Shouldn't Fear Upgrading Your ACI Fabric - The Handbook!

Feb 8 | 10:30
BRKDCN-2673
Nexus-as-Code - Kickstart your automation with ACI

Feb 8 | 12:00
BRKDCN-2949
Cisco ACI Multi-Pod Design and Deployment

Feb 8 | 14:30
BRKDCN-2980
ACI Multi-Site Architecture and Deployment

Feb 9 | 08:30
BRKDCN-2950
Nexus Cloud: How to manage your Nexus Data Center from the cloud

Feb 9 | 10:45
BRKDCN-3900
A Network Engineer's Blueprint for ACI Forwarding

Feb 9 | 13:45
BRKDCN-3982
ACI L4-L7 Policy-Based Redirect (PBR) Deep Dive and Tips

Feb 9 | 15:45
BRKDCN-3612
Secure Firewall in ACI

Feb 10 | 11:00
FINISH
BRKDCN-2969
Managing your data center network with ServiceNow

If you are unable to attend a live session, you can watch it On Demand after the event

cisco Live!

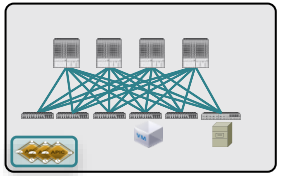
Agenda

- Overview, Use Cases, and supported Topologies
- APIC Cluster Deployment
- Inter-Pod Connectivity
- Control and Data Planes
- Connecting to External Networks
- Network Services Integration
- Remote Leaf

Overall Design Principles (AZs and Regions)

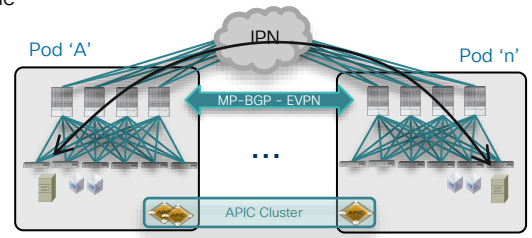
ACI Fabric and Policy Domain Evolution

ACI Single Pod Fabric



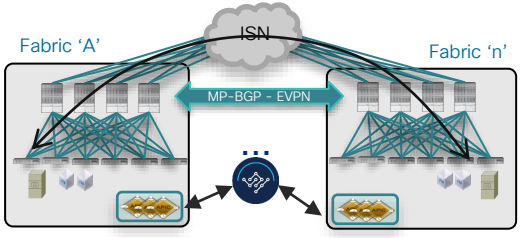
ACI 1.0 – Leaf/Spine Single Pod Fabric

ACI Multi-Pod Fabric



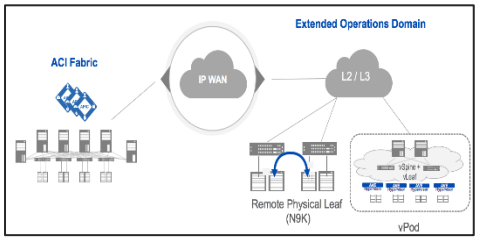
ACI 2.0 – Multiple Networks (Pods or Availability Zones) in a single Fabric (Region)

ACI Multi-Site



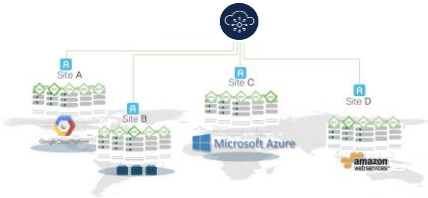
ACI 3.0 – Multiple Fabrics (Regions) interconnected in the same Multi-Site Orchestrator domain

ACI Remote Leaf



ACI 3.1/4.0 – Remote Leaf extends a Fabric to remote locations

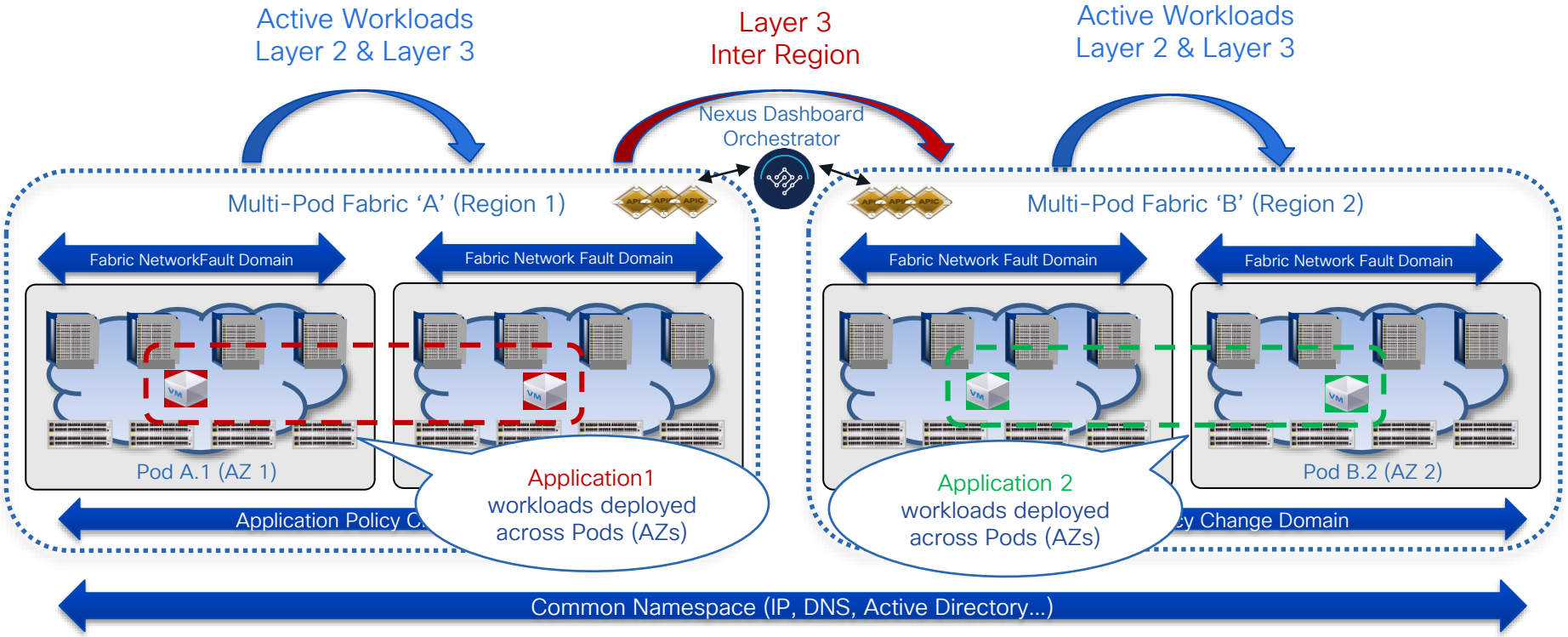
Cloud Network Controller



ACI 4.1 & 4.2 – ACI Extensions to Public Cloud

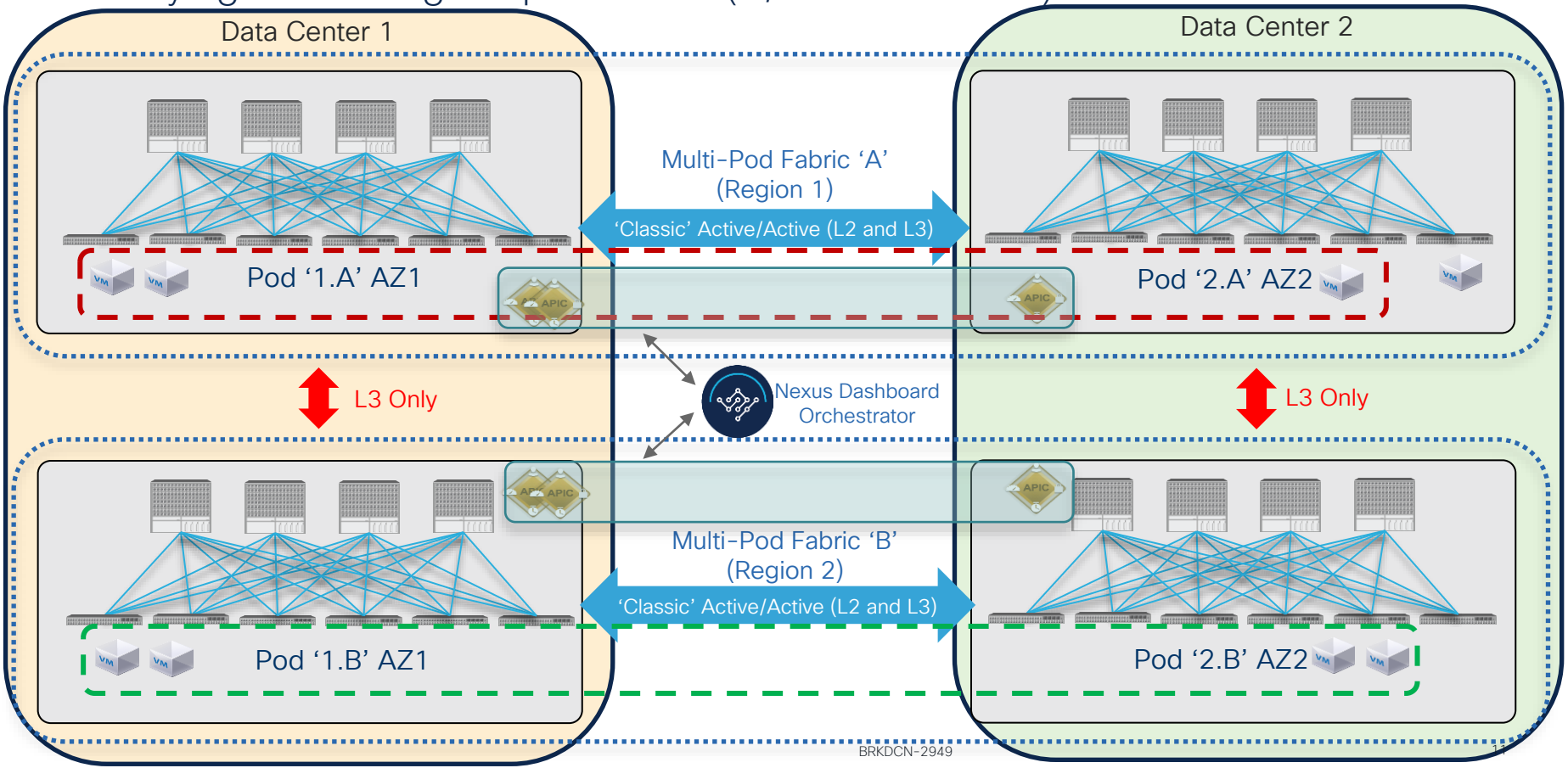
Systems View (How do these things relate)

Change and Network Fault Domain Isolation



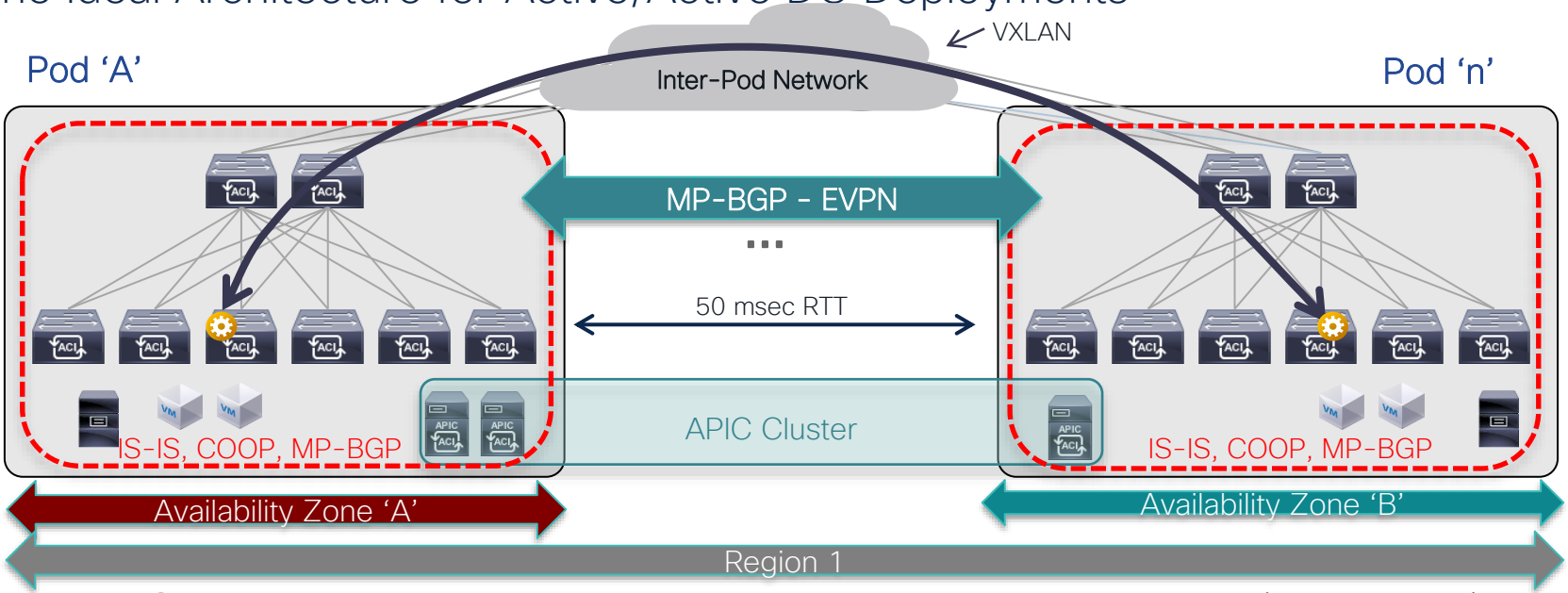
Multi-Pod + Multi-Site

Satisfying Conflicting Requirements (A/A DCs and DR)



ACI Multi-Pod

The Ideal Architecture for Active/Active DC Deployments

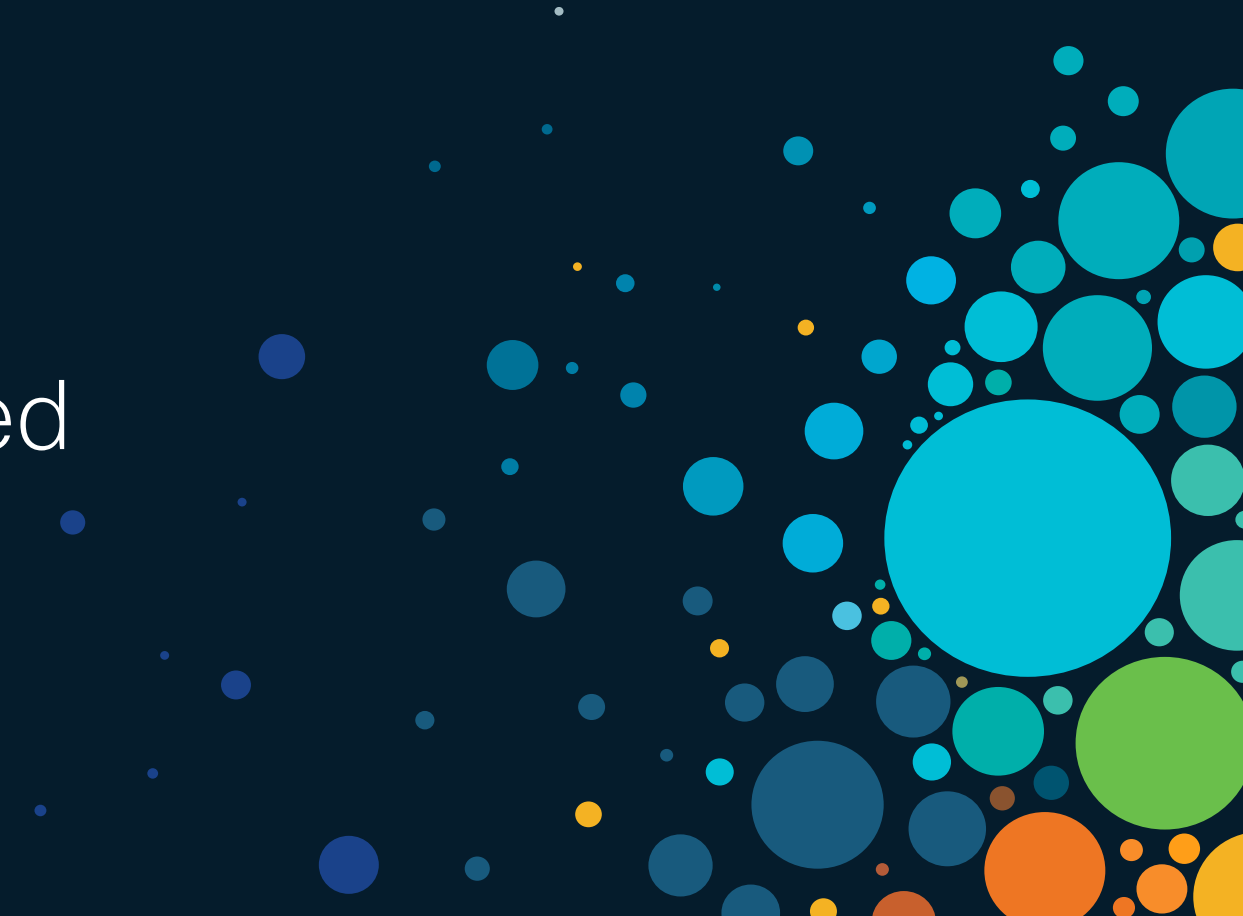


- Multiple ACI Pods connected by an IP Inter-Pod L3 network, each Pod consists of leaf and spine nodes
- Managed by a single APIC Cluster
- Single Management and Policy Domain
- Forwarding control plane (IS-IS, COOP) fault isolation
- Data Plane VXLAN encapsulation between Pods
- End-to-end policy enforcement

ACI Multi-Pod Deep Dive

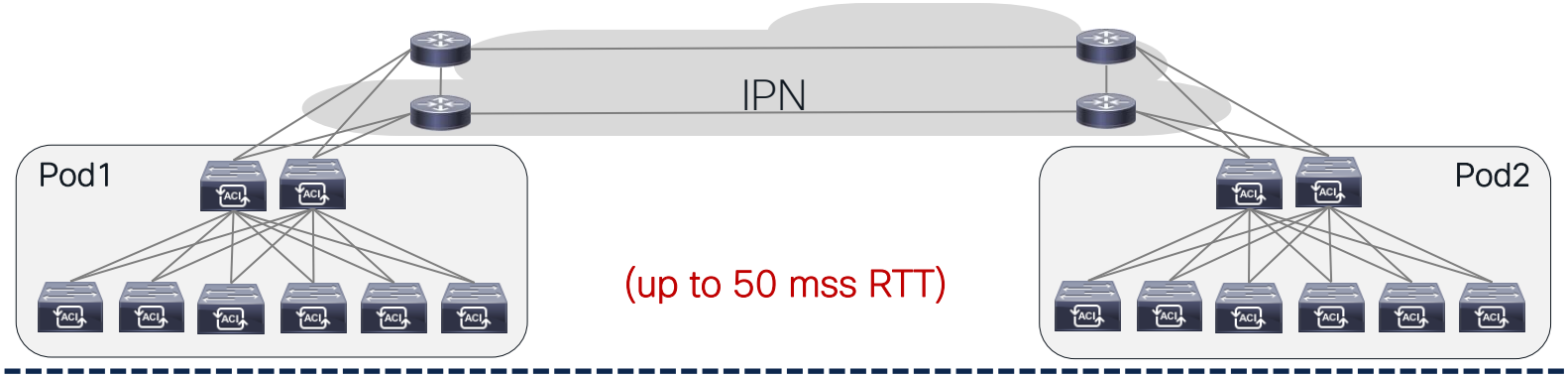
cisco Live!

Overview, Use Cases and Supported Topologies

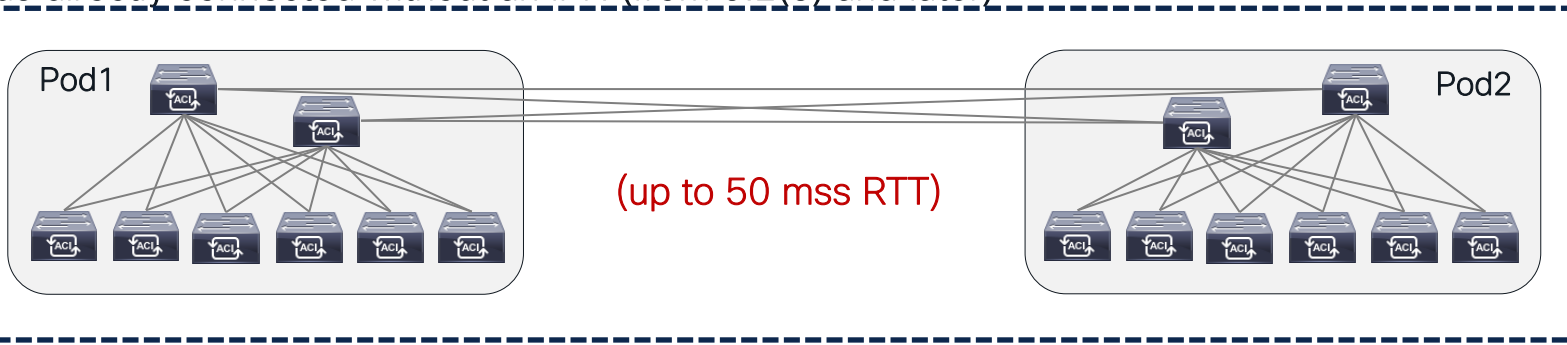


Multi-Pod Supported Topologies

Pods connected via an Inter-Pod Network (IPN)



Pods directly connected without an IPN (from 5.2(3) and later)



Multi-Pod Spines Back-to-Back

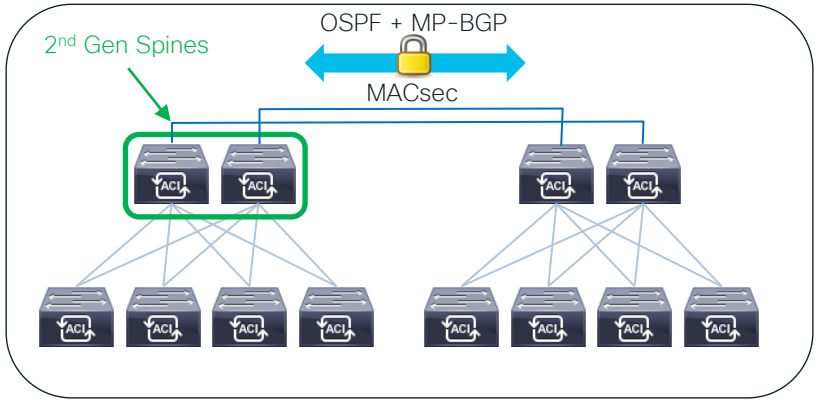
Problem Description

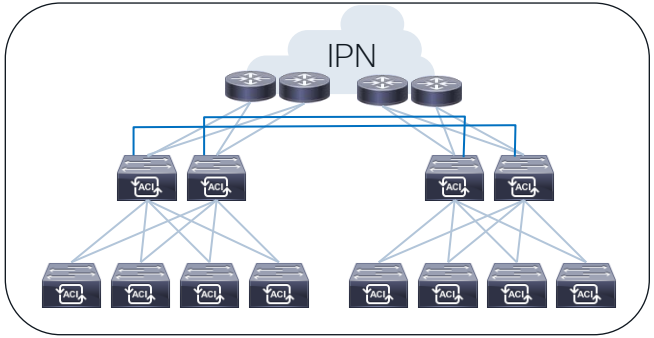
ACI Release 5.2(3)

- Many customers deploy ACI Multi-Pod fabrics with only two Pods and are not using any other features that require spine IPN connectivity (Multi-Site, Remote leaf, GOLF, Cloud ACI)
- These customers may have small to medium size fabrics and the requirement to build an additional network (IPN) for inter-Pod connectivity is an added cost (plus the need to operate it)
- The ACI Multi-Pod Spines back-to-back option removes the requirement to build and operate an additional network for inter-Pod connectivity

Multi-Pod Spines Back-to-Back

Guidelines and Restrictions





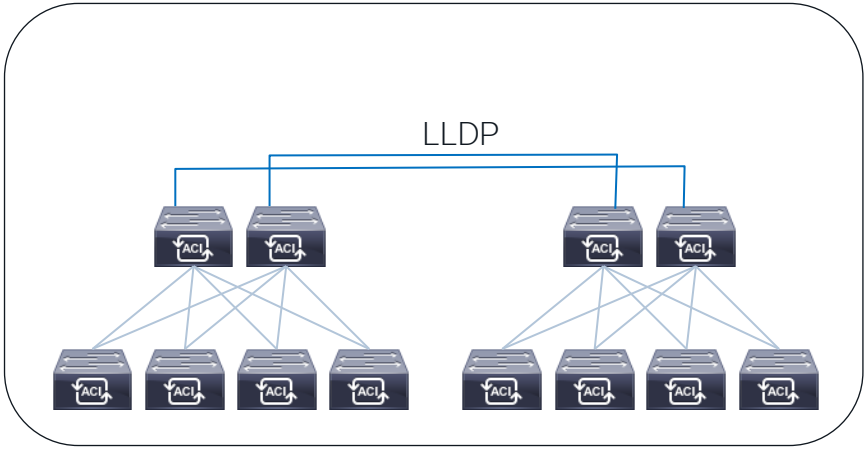
Back-to-Back + IPN (Migration Only)

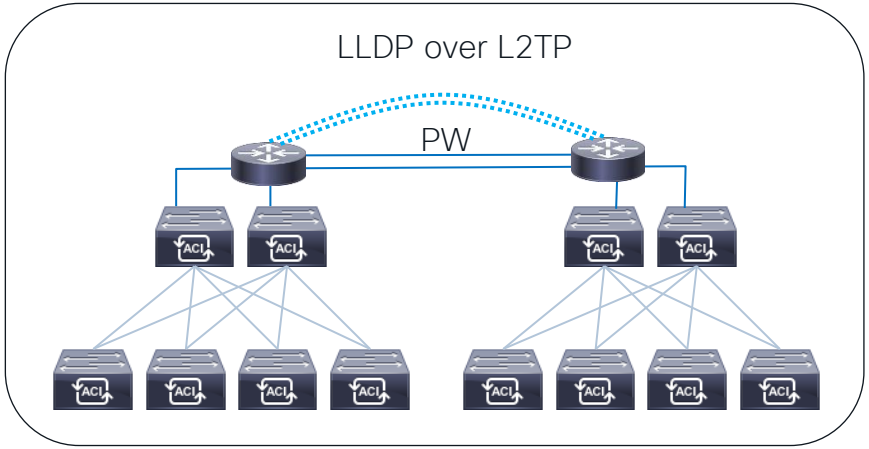
- Support is limited to a topology with 2 Pods leveraging 2nd generation spines only
- OSPF underlay peering, MP-BGP overlay peering between the spines in separate Pods
  - No need for PIM-Bidir (spines do not run PIM)
- MACsec encryption supported across Pods
- Not compatible functions
  - ACI Multi-Site
  - Remote Leaf
  - GOLF
  - Cloud ACI
  - APIC connectivity via L3 network
- Back-to-Back + IPN only supported for migration purposes (migration is disruptive)

Multi-Pod Spines Back-to-Back

Supported Topologies

- Back-to-back spine connectivity must be point-to-point (physical or logical)
- Spines discover back-to-back connections via LLDP
- Links can be directly connected or must support tunneling of LLDP packets





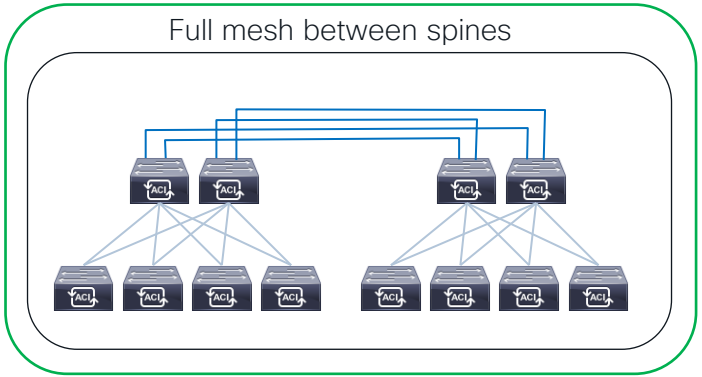
Multi-Pod Spines Back-to-Back

Supported Topologies

It is not mandatory for all spines in a Pod to connect to all the spines in the other Pod, the design decision must be made based on resiliency/bandwidth considerations

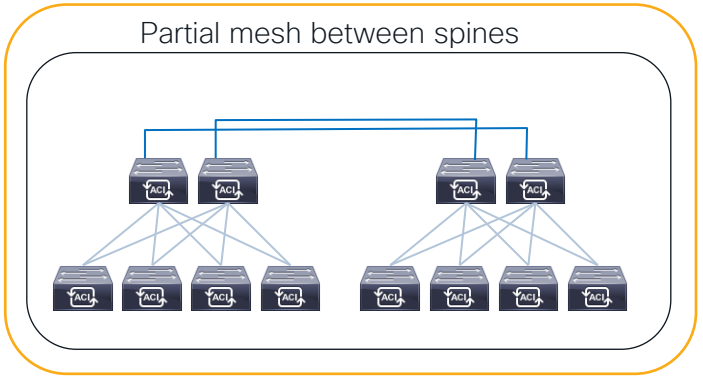
Recommended

Full mesh between spines



Supported

Partial mesh between spines



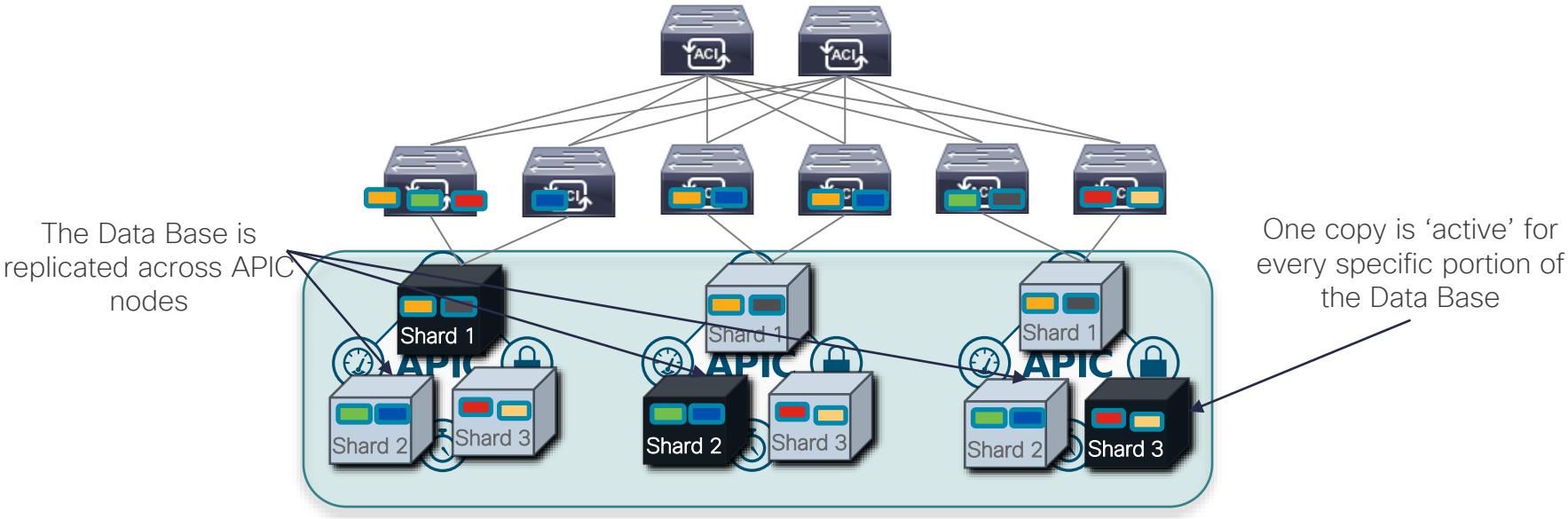
ACI Multi-Pod

SW/HW Support and Scalability Values

- All existing Nexus 9000 HW supported as leaf and spine nodes*
- Maximum number of supported ACI leaf nodes (across all Pods)
  - Up to 80 leaf nodes supported with a 3 node APIC cluster
  - 200 leaf nodes (across Pods) with a 4 node APIC cluster (from ACI release 4.1)
  - 300 leaf nodes (across Pods) with a 5 node APIC Cluster
  - 400 leaf nodes (across Pods) with a 7 node APIC Cluster (from ACI release 2.2(2e))
  - 500 leaf nodes (across Pods) with a 7 node APIC Cluster (from ACI release 4.2(4))
  - Maximum 400 leaf nodes per Pod (from ACI release 4.2(4))
  - Up to 6 spines per Pod, 50 spines per Fabric (from ACI release 6.0(1))
- Maximum number of supported Pods
  - 4 in 2.0(1)/2.0(2) releases
  - 6 in 2.1(1) release
  - 10 in 2.2(2e) release
  - 12 in 3.0(1) release
  - 25 in 6.0(1) release

APIC Cluster Deployment Considerations

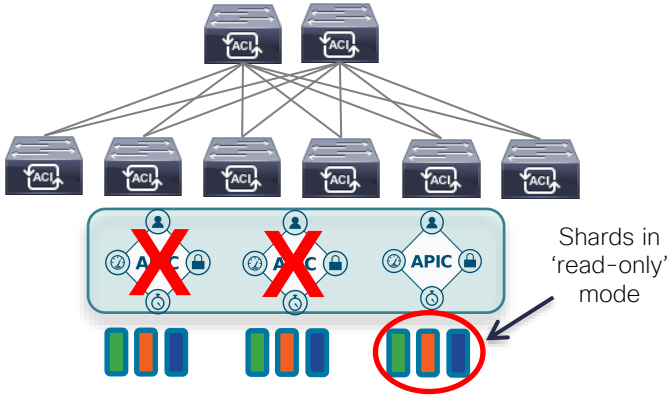
APIC – Distributed Multi-Active Data Base



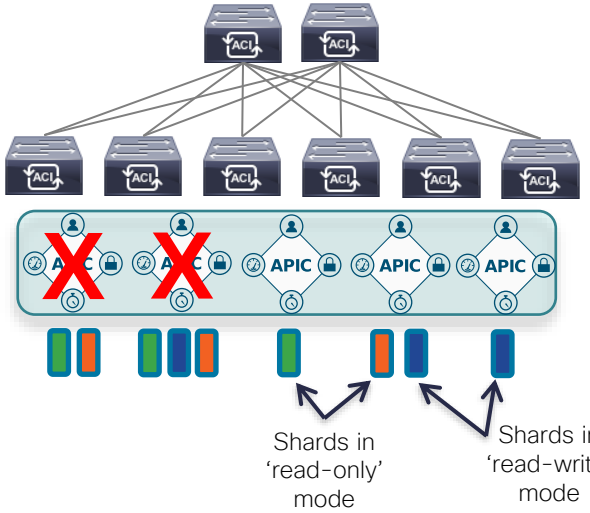
- Processes are active on all nodes (not active/standby)
- The Data Base is distributed as active + 2 backup instances (shards) for every attribute

APIC Cluster Deployment Considerations

Single Pod Scenario



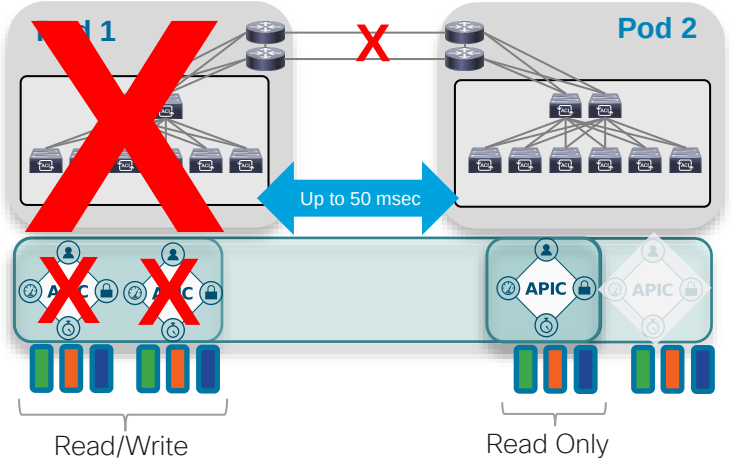
- APIC will allow read-only access to the DB when only one node remains active (standard DB quorum)
- Hard failure of two nodes cause all shards to be in 'read-only' mode (of course reboot etc. heals the cluster after APIC nodes are up)



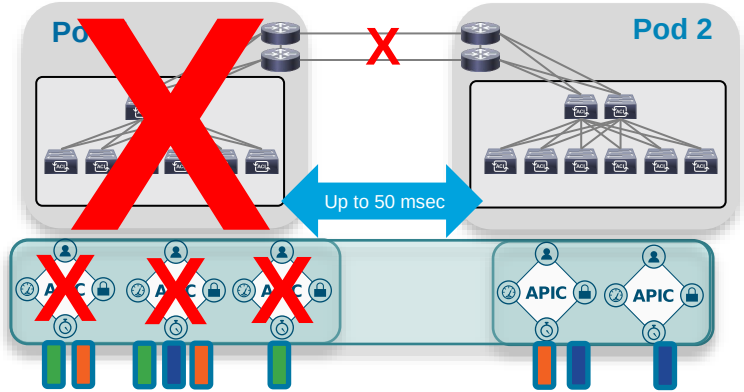
- Additional APIC will increase the system scale (up to 7* nodes supported) but does not add more redundancy
- Hard failure of two nodes would cause inconsistent behaviour across shards (some will be in 'read-only' mode, some in 'read-write' mode)

APIC Cluster Deployment Considerations

Multi-Pod – 2 Pods Scenario



- Pod isolation scenario: changes still possible on APIC nodes in Pod1 but not in Pod2
- Pod hard failure scenario: recommendation is to activate a standby node to make the cluster fully functional again



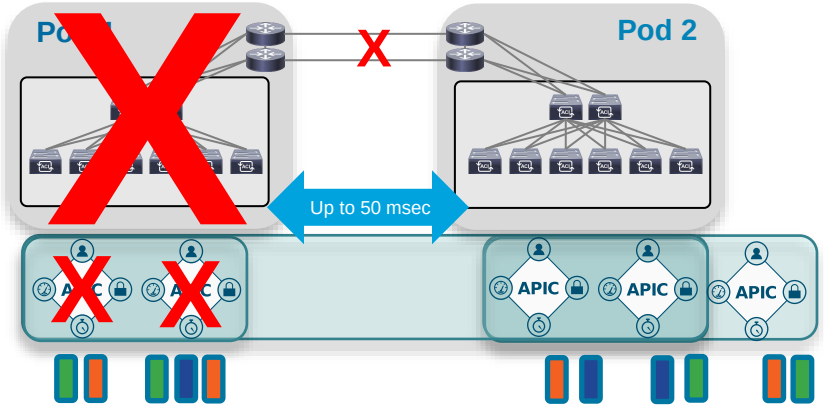
- Pod isolation scenario: same considerations as with single Pod (different behaviour across shards)
- Pod hard failure scenario: may cause the loss of information for the shards replicated across APIC nodes in the failed Pod

  Possible to restore the whole fabric state to the latest taken configuration snapshot ('ID Recovery' procedure – needs BU and TAC involvement)

APIC Cluster Deployment Considerations

ACI Release 4.1(1)

What about a 4 Nodes APIC Cluster?

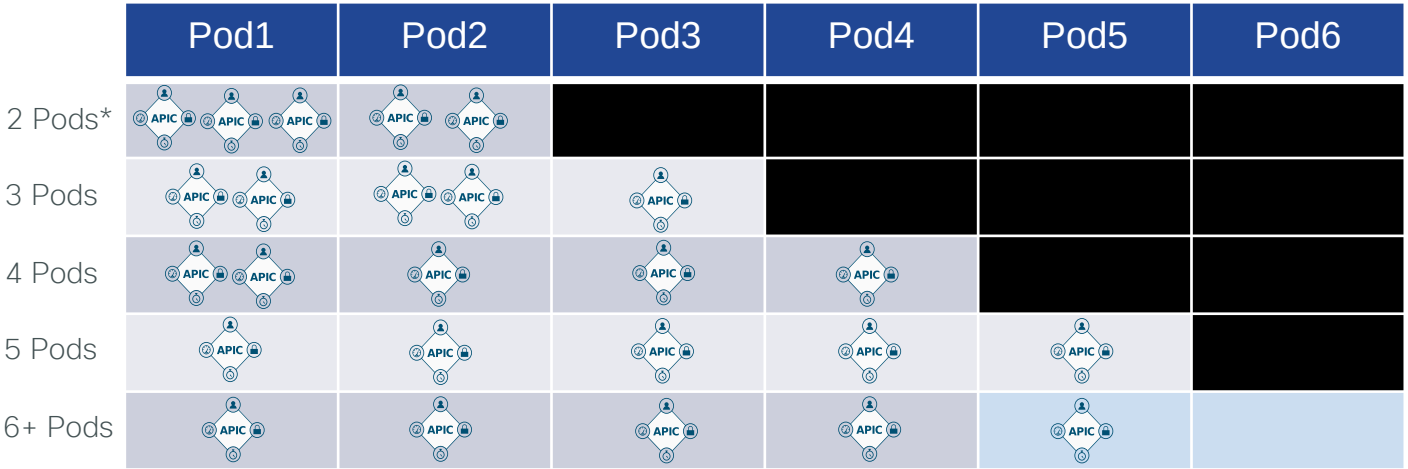


- Intermediate scalability values compared to a 3 or 5 nodes cluster scenario (up to 200 leaf nodes supported)
- Pod isolation scenario: same considerations as with 5 nodes (different behaviour across shards)
- Pod hard failure scenario
  - No chance of total loss of information for any shard
  - Can bring up a standby node in the second site to regain full majority for all the shards

APIC Cluster Deployment Considerations

Deployment Recommendations

- Main recommendation: deploy a 3 nodes APIC cluster when less than 80 leaf nodes are deployed across Pods
- From 4.1(1) can deploy 4 nodes if the scalability requirements are met
- When 5 (or 7) nodes are really needed for scalability reasons, follow the rule of thumb of never placing more than two APIC nodes in the same Pod (when possible):



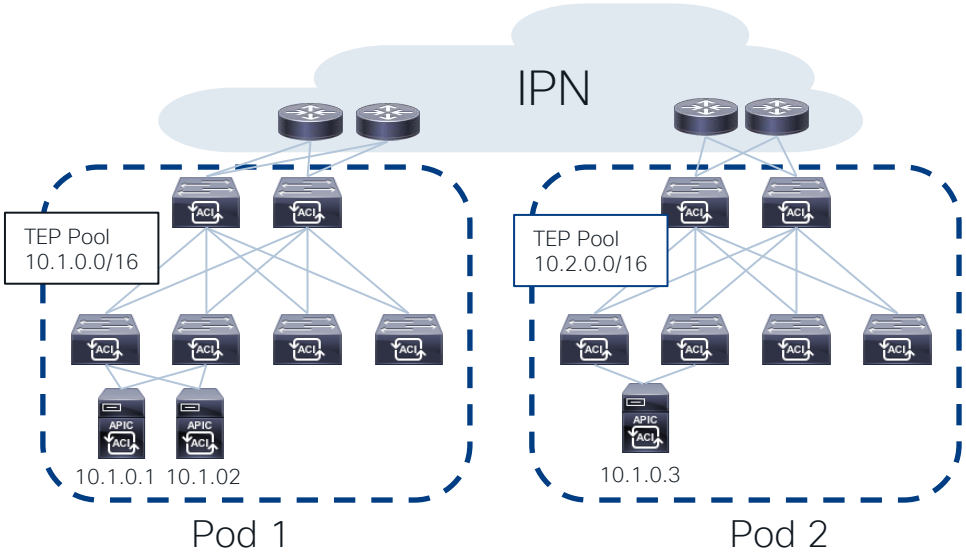
*'ID Recovery' procedure possible for recovering of lost information

APIC Connectivity over L3 Network

cisco Live!

APIC Connectivity Options

APIC Cluster directly connected to fabric

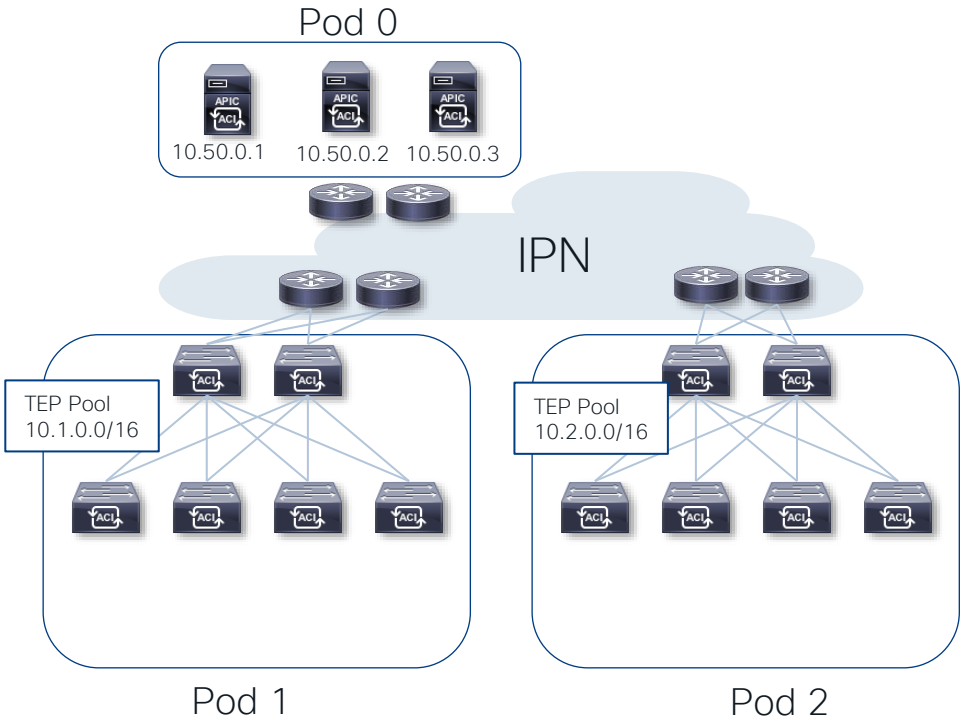


- APICs can be placed in any pod
- APIC fabric IP addresses are always assigned from pod 1 TEP pool
- Recommended to distribute APICs across pods so loss of a pod does not bring down the entire cluster

APIC Connectivity Options

APIC cluster connected over L3 Network

ACI Release 5.2(1)

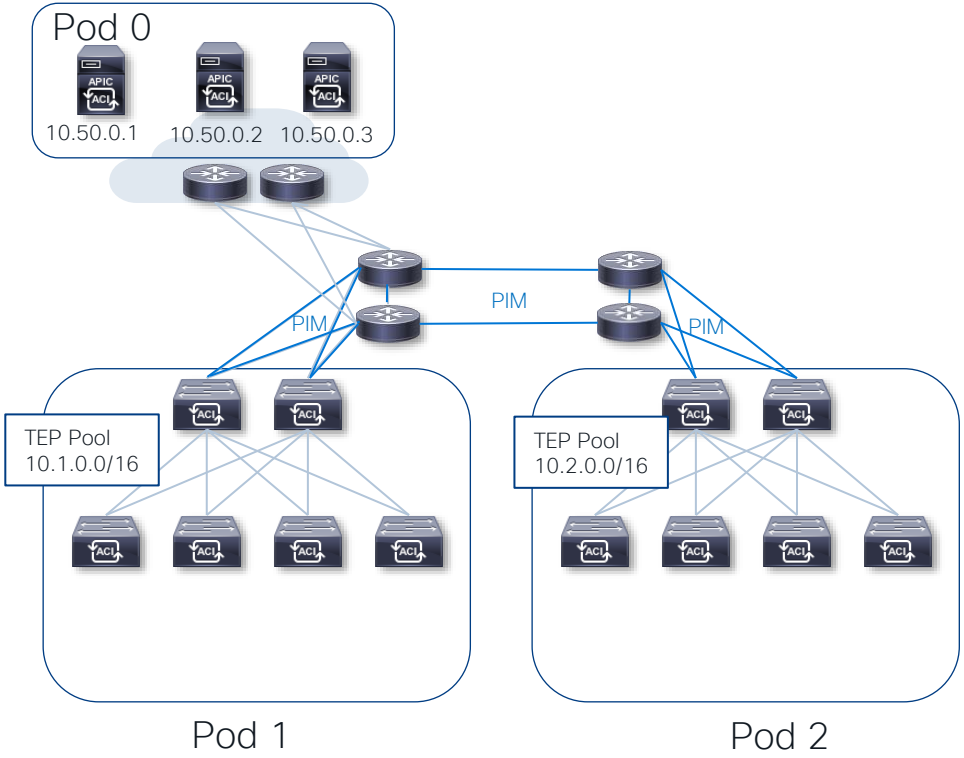


- APICs do not need to be directly connected to the leaf switches. Can be placed in L3 network that has IP reachability to the spines via IPN
- APICs will be part of pod 0. Pod 0 is a special pod that only contains APICs and no fabric switches
- APIC fabric IP addresses are user configurable. Not assigned from any pod TEP range
- APIC fabric IPs can be in the same or different subnet per APIC
- APICs can be geographically distributed within the Multi-Pod 50 msec distance requirement

APIC Connectivity Options

ACI Release 5.2(1)

APIC cluster connected over L3 Network, IPN Multicast Requirement

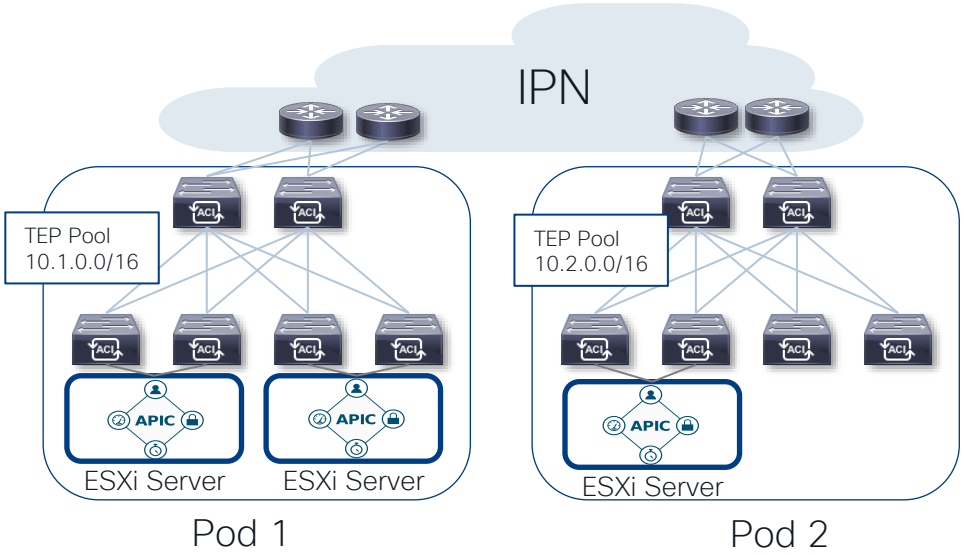


- Multicast (PIM Bidir) is only required for inter-pod BUM traffic
- If APIC cluster over L3 network is managing only one pod, multicast is not required in the IPN
- If it is a Multi-Pod fabric, multicast is only required on the links interconnecting the pods

APIC Connectivity Options

Virtual APIC cluster

ACI Release 6.0(2)

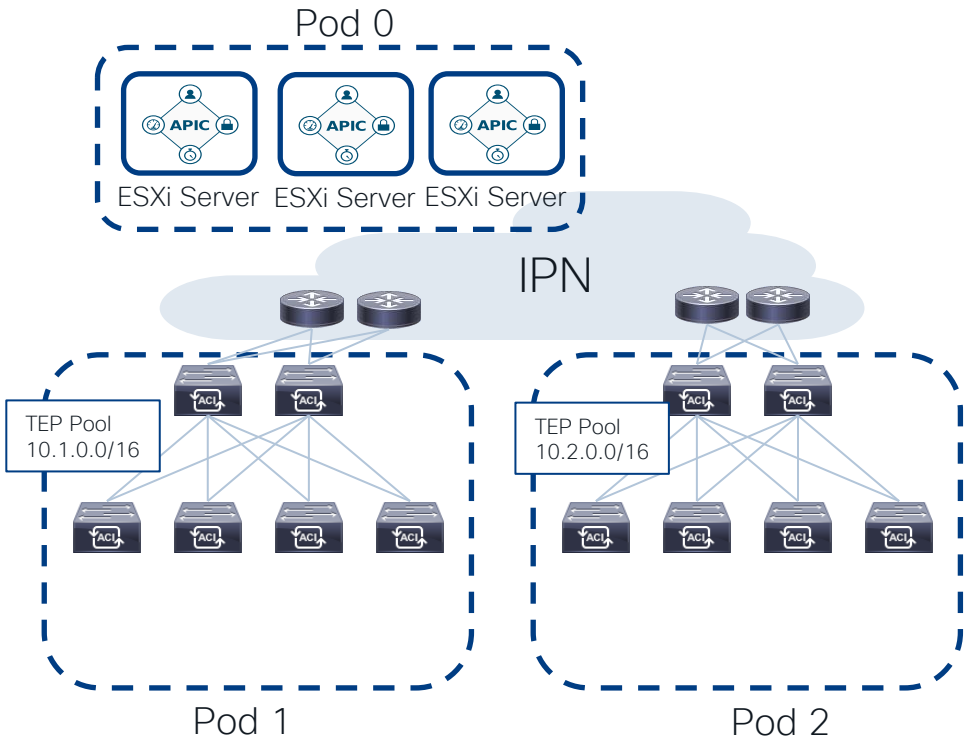


- Virtual APIC cluster (all virtual APICs)
- Runs as a VM on an ESXi hypervisor
- ESXi server directly connected to fabric
- No mixed cluster support. Must be all virtual or all physical
- Supports all types of deployments, Remote Leaf, Multi-Pod, Multi-Site.

APIC Connectivity Options

Virtual APIC cluster over L3 Network

ACI Release 6.0(2)



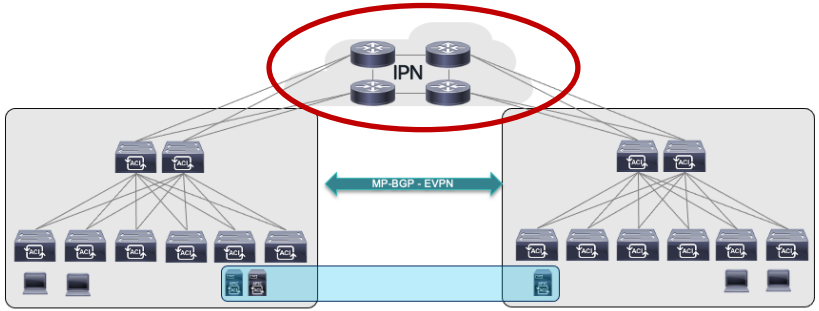
- Virtual APIC over L3 Network
- Same or different IP addresses per APIC same as physical APIC over L3 network
- Cannot mix virtual APIC over L3 Network with directly connected virtual APIC

Inter-Pod Connectivity Deployment Considerations

cisco Live!

ACI Multi-Pod

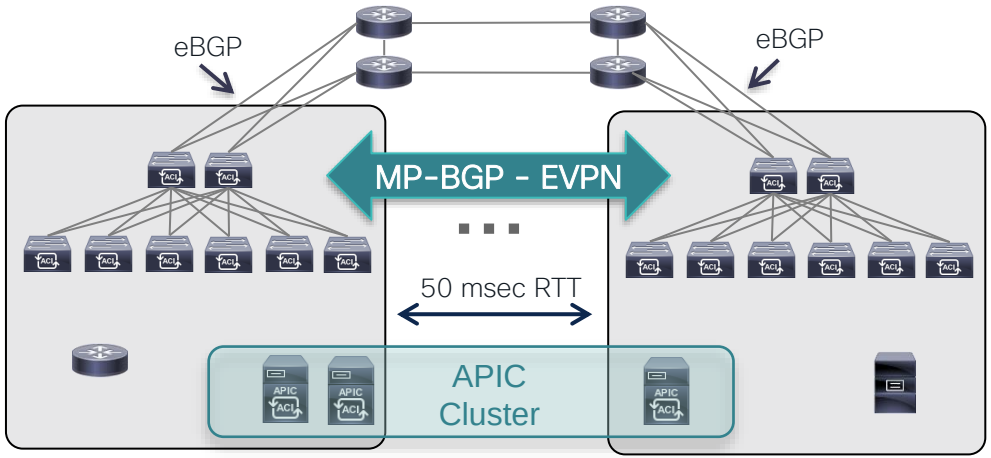
Inter-Pod Network (IPN) Requirements



- Not managed by APIC, must be separately configured (day-0 configuration)
- IPN topology can be arbitrary, not mandatory to connect to all spine nodes
- Main requirements:
  - Multicast BiDir PIM → needed to handle Layer 2 BUM* traffic
  - OSPF or BGP to peer with the spine nodes and learn VTEP reachability
  - Increase MTU support to handle VXLAN encapsulated traffic
  - DHCP-Relay

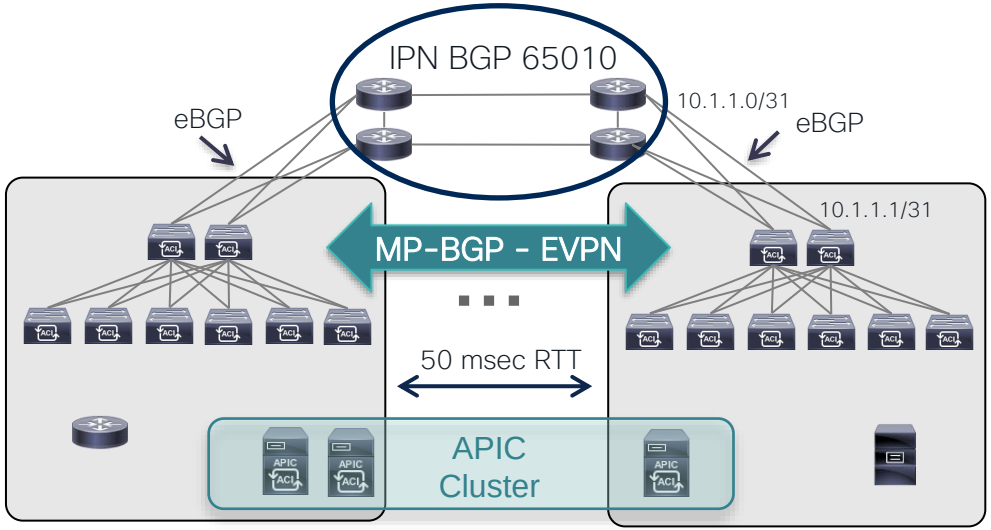
* Broadcast, Unknown unicast, Multicast

BGP Underlay Support for IPN links



- From ACI 5.2(3) you can use either OSPF and/or BGP for IPN connectivity
- Infra L3Out interfaces can be configured with OSPF, BGP, or both protocols at the same time (typically used for migration)
- Only eBGP is supported
- Supported for Multi-Pod, Remote Leaf, Multi-Site, and APIC over L3 Network
- When both protocols are configured, BGP routes will be preferred due to lower admin distance

BGP Underlay Support for IPN links



- Configure BGP 'disable-peer-as-check' if Nexus switches are used for IPN
- Nexus switches will not advertise prefixes to peer if peer AS is already in the AS PATH. 'disable-peer-as-check' turns off this behavior

Sample IPN configuration (Nexus 9000)

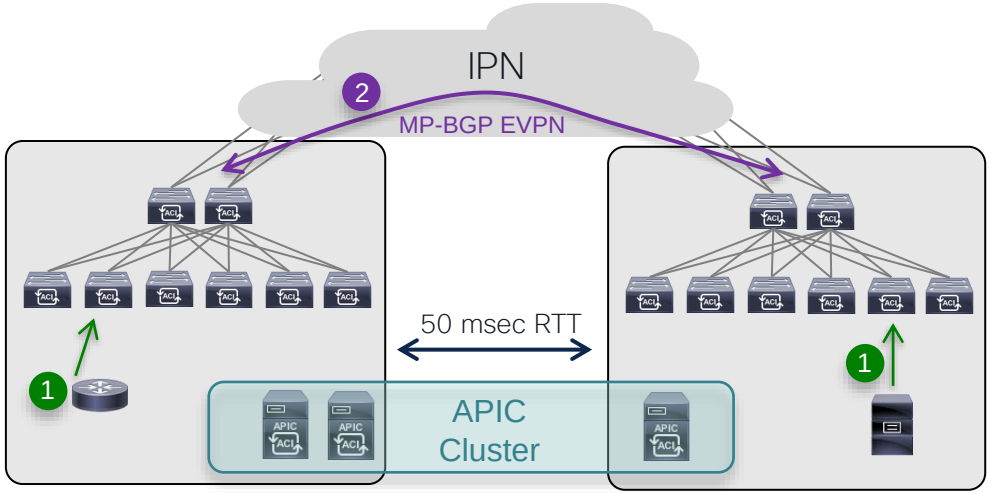
```
feature bgp

router bgp 65010
  router-id 10.10.10.1
  vrf IPN
    address-family ipv4 unicast
    neighbor 10.1.1.1
      remote-as 65001
      address-family ipv4 unicast
        disable-peer-as-check
```

ACI Multi-Pod and MTU

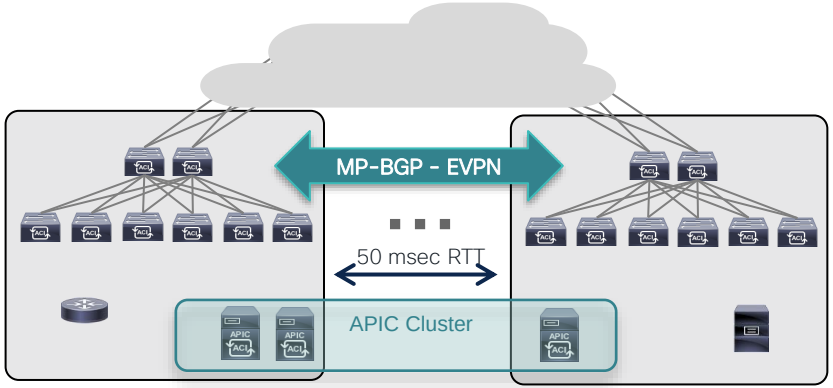
Different MTU Meanings

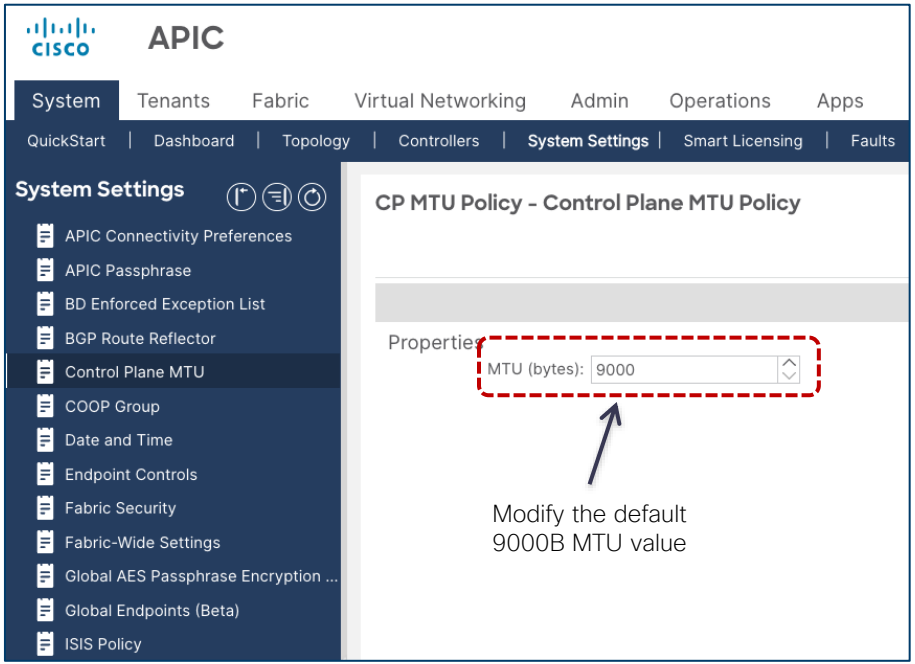
1. Data Plane MTU: MTU of the traffic generate by endpoints (servers, routers, service nodes, etc.) connected to ACI leaf nodes
   - Need to account for 50B of overhead (VXLAN encapsulation) for inter-Pod communication
2. Control Plane MTU: for CPU generated traffic like EVPN across sites
   - The default value is **9000B,** can be tuned to the maximum MTU value supported in the ISN



ACI Multi-Pod and MTU

Tuning CP MTU for EVPN Traffic across Pods



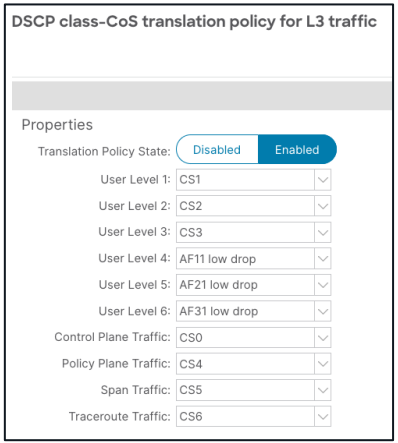


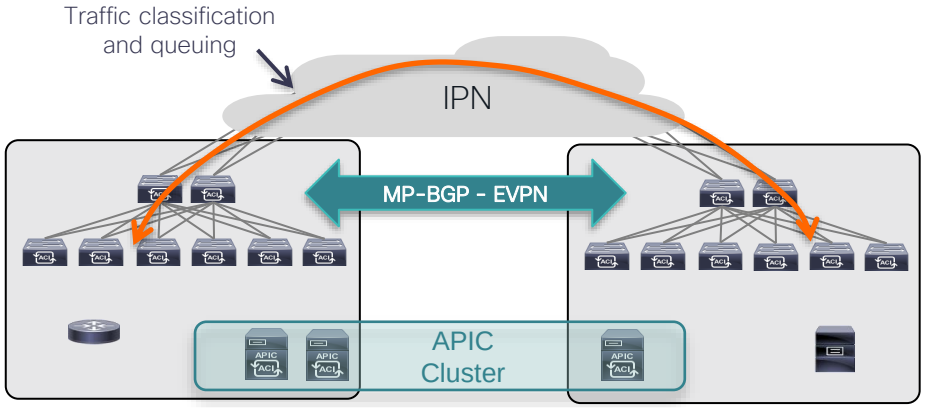
- Control Plane MTU can be set leveraging the "CP MTU Policy" on APIC
- The required MTU in the IPN would then depend on this setting and on the Data Plane MTU configuration
  - Always need to consider the VXLAN encapsulation overhead for data plane traffic (50/54 bytes)

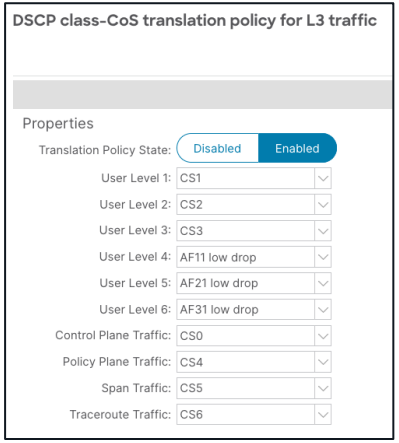
ACI Multi-Pod and QoS

Inter-Pod QoS Behavior

- Traffic across sites should be consistently prioritized (as it happens intra-site)
- To achieve this end-to-end consistent behavior it is required to configure DSCP-to-CoS mapping in the ‘infra’ Tenant
  - Allows to classify traffic received on the spines from the IPN based on outer DSCP value
  - Without the DSCP-to-CoS mapping configuration, classification for the same traffic will be CoS based (preserving CoS value in the IPN is harder)
- The traffic can also then be properly treated inside the IPN (classification/queuing)
  - Recommended to always prioritize at least Policy and Control Plane traffic



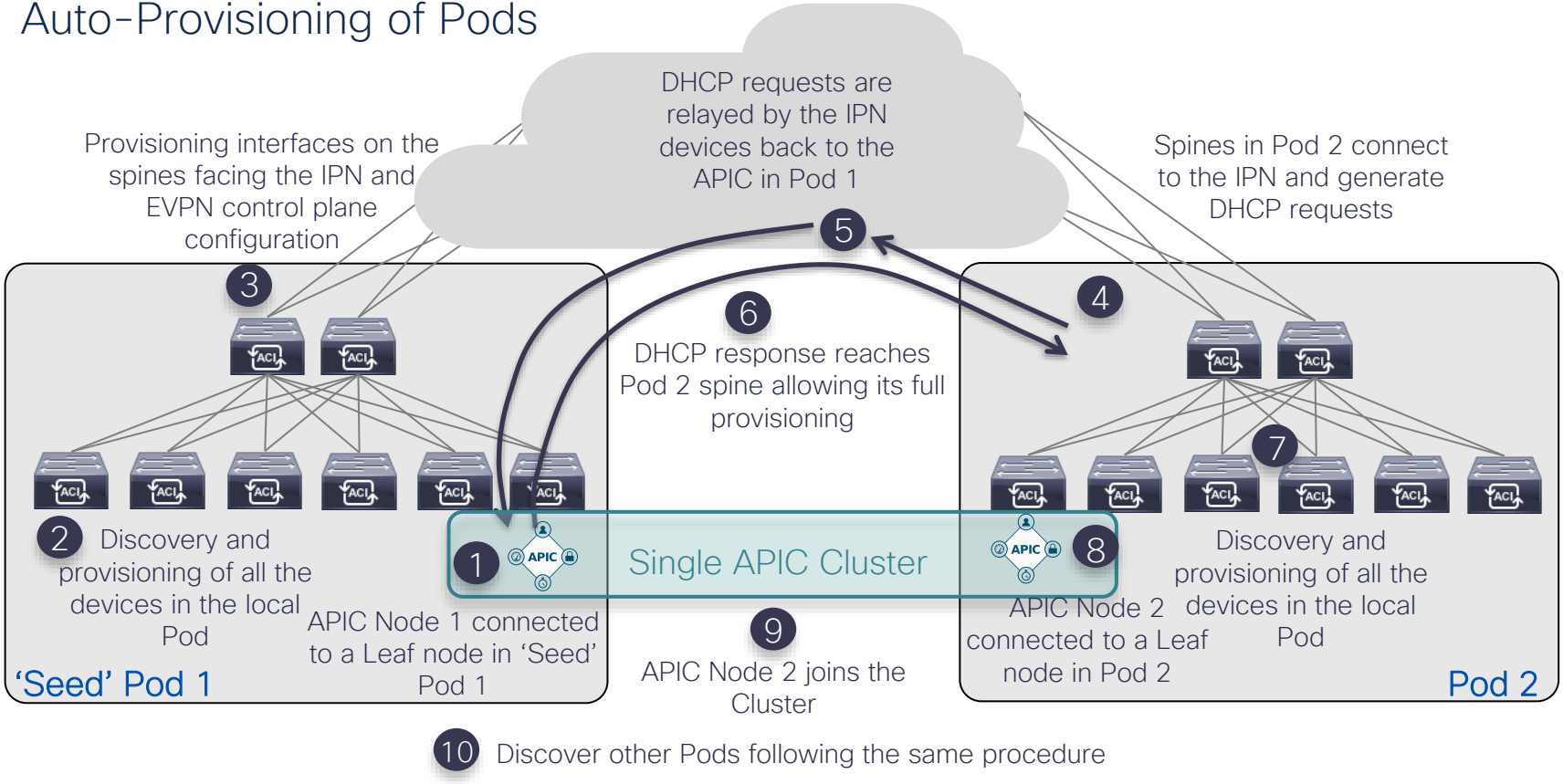




Control and Data Planes

ACI Multi-Pod

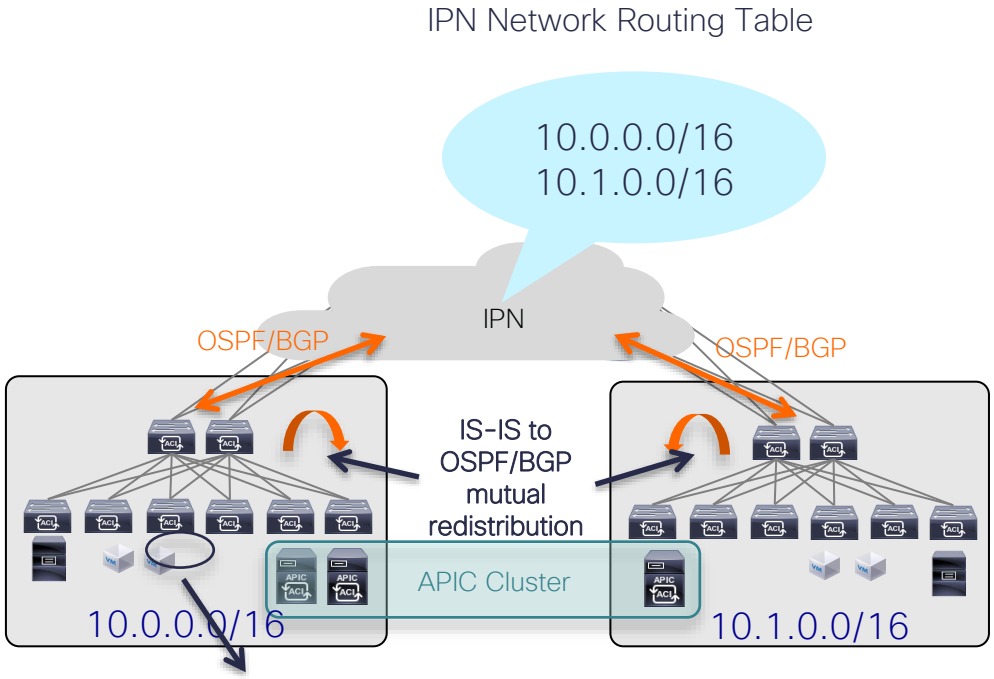
Auto-Provisioning of Pods



ACI Multi-Pod

IPN Control Plane

- Separate IP address pools for VTEPs assigned by APIC to each Pod
  - Summary routes advertised toward the IPN via OSPF or BGP routing
  - IS-IS convergence events local to a Pod not propagated to remote Pods
- Spine nodes redistribute other Pods summary routes into the local IS-IS process
  - Needed for local VTEPs to communicate with remote VTEPs



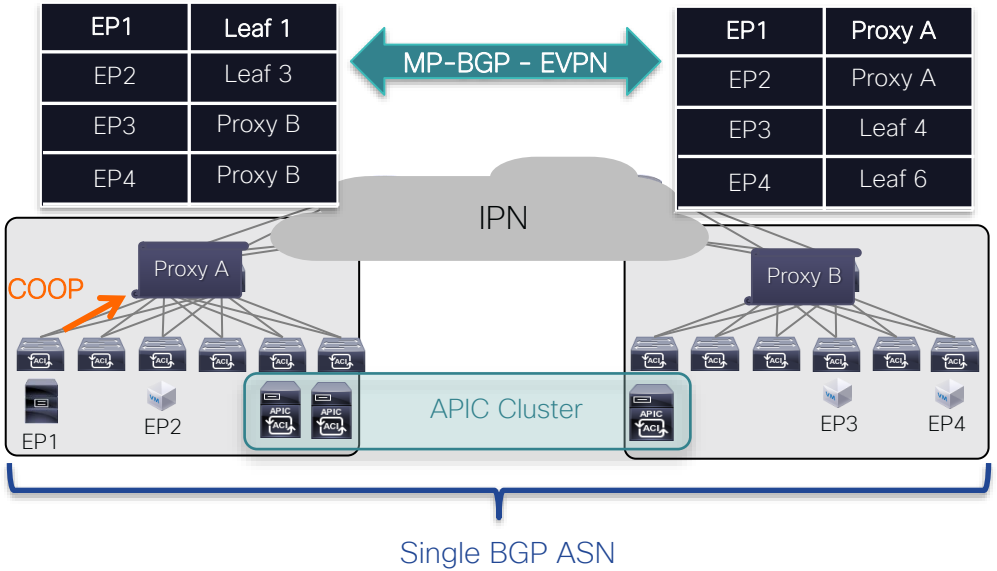
Leaf Node Underlay VRF

IP Prefix	Next-Hop
10.1.0.0/16	Pod1-S1, Pod1-S2, Pod1-S3, Pod1-S4

ACI Multi-Pod

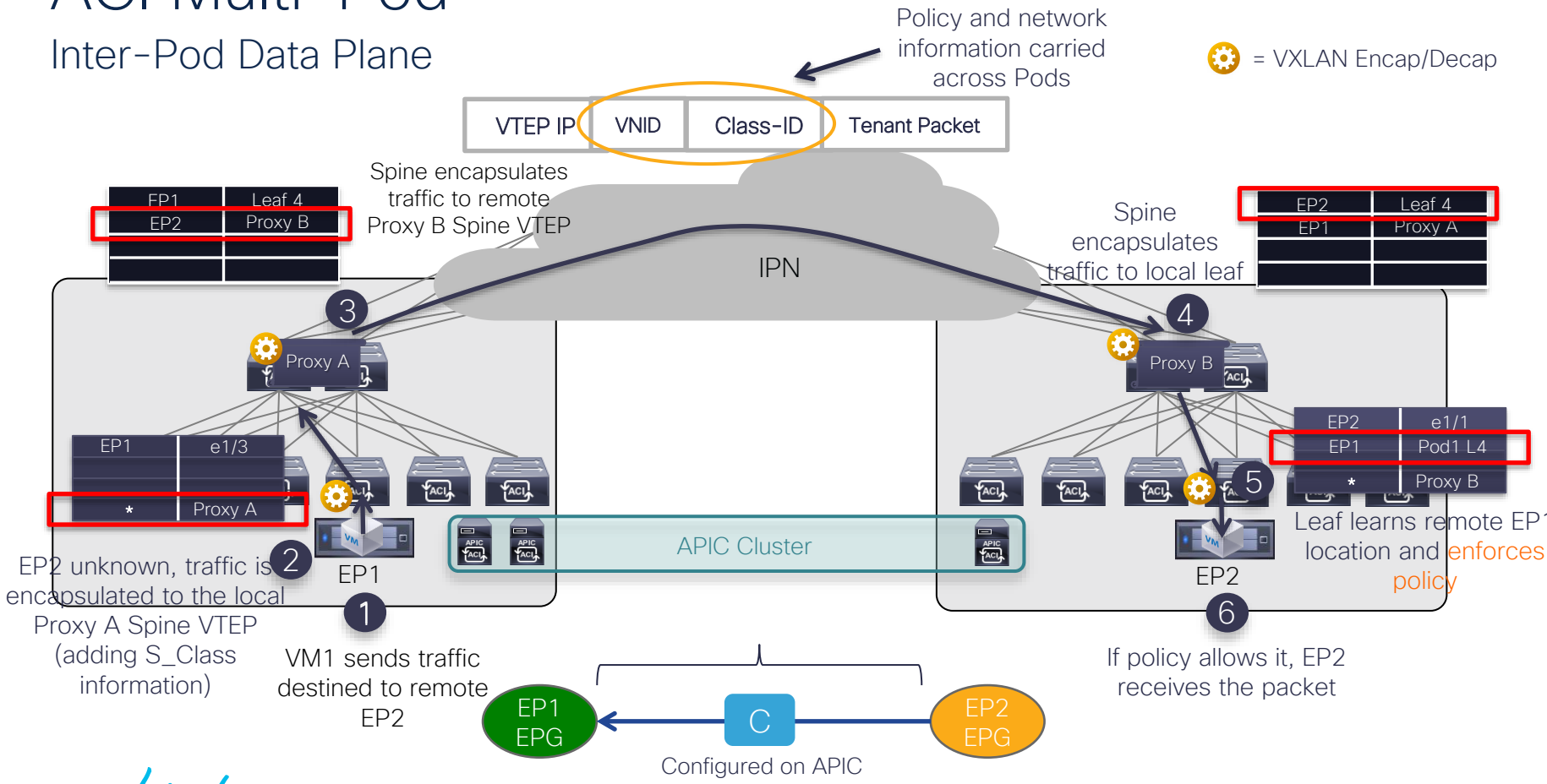
Inter-Pod MP-BGP EVPN Control Plane

- MP-BGP EVPN to sync Endpoint (EP) and Multicast Group information
  - All remote Pod entries associated to a Proxy VTEP next-hop address (not part of local TEP Pool)
  - Same BGP AS across all the Pods
- iBGP EVPN sessions between spines in separate Pods
  - Full mesh MP-iBGP EVPN sessions between local and remote spines (default behavior)
  - Optional RR deployment (recommended one RR in each Pod for resiliency)



ACI Multi-Pod

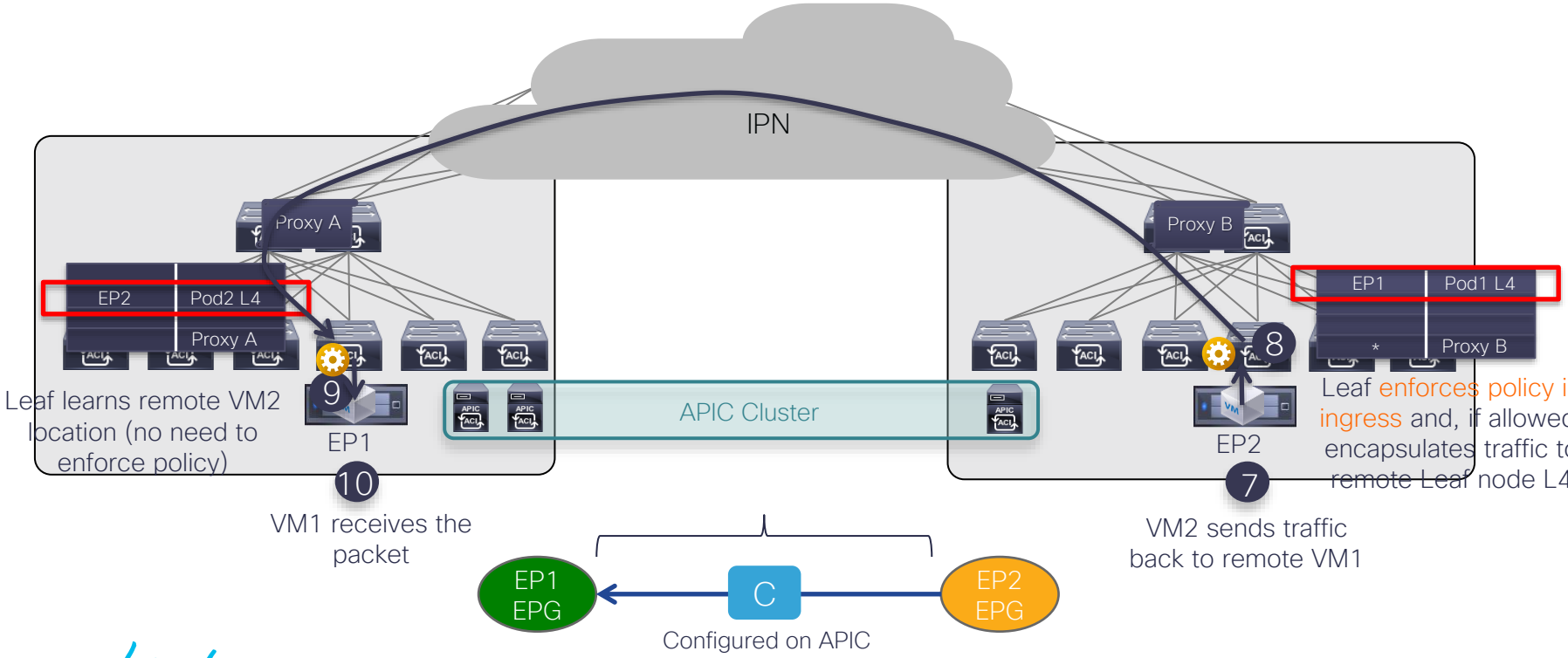
Inter-Pod Data Plane



ACI Multi-Pod

Inter-Pod Data Plane (2)

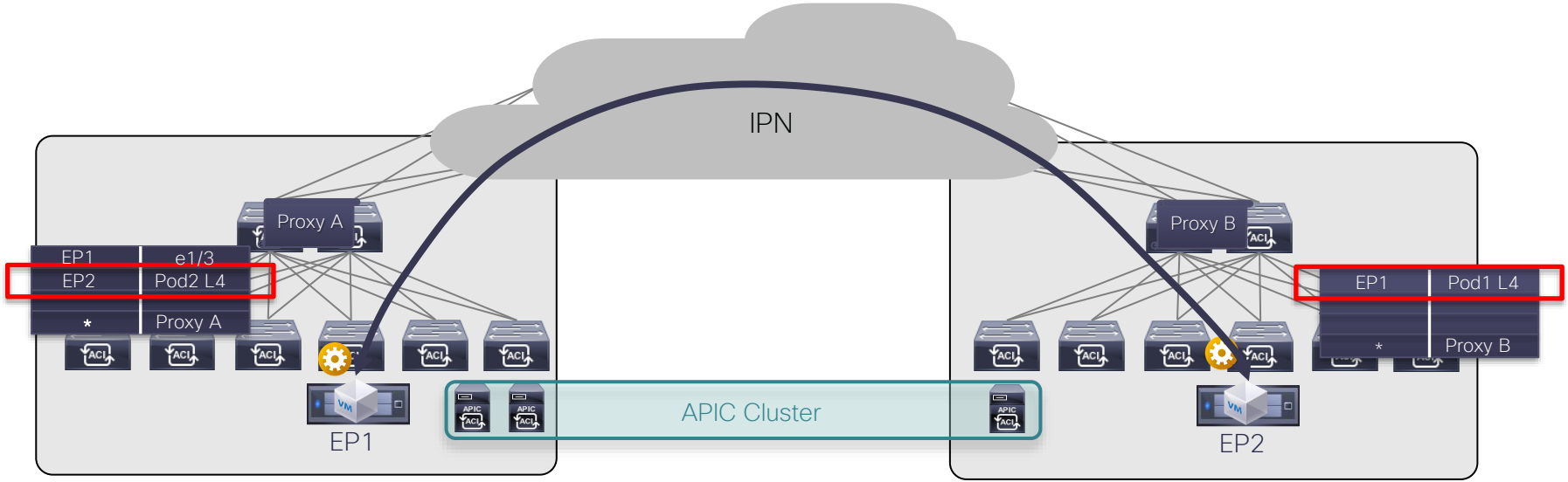
= VXLAN Encap/Decap



ACI Multi-Pod

Inter-Pod Data Plane (3)

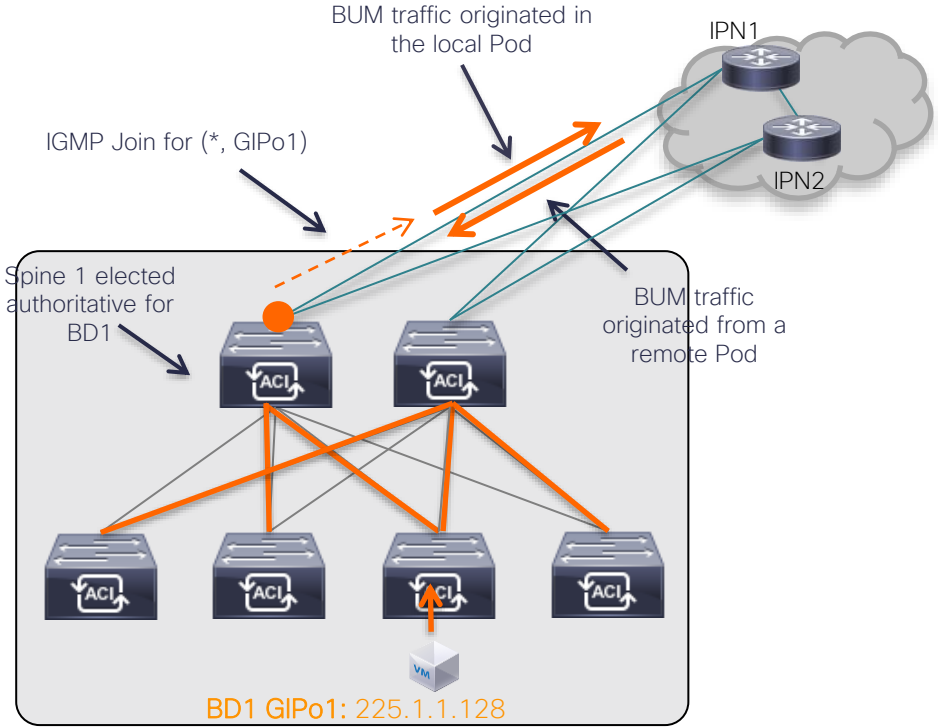
= VXLAN Encap/Decap



11 From this point EP1 to EP2 communication is encapsulated Leaf to Leaf (VTEP to VTEP) and policy always applied at the ingress leaf (applies to both L2 and L3 communication)

ACI Multi-Pod

Use of Multicast for Inter-Pod Layer 2 BUM Traffic

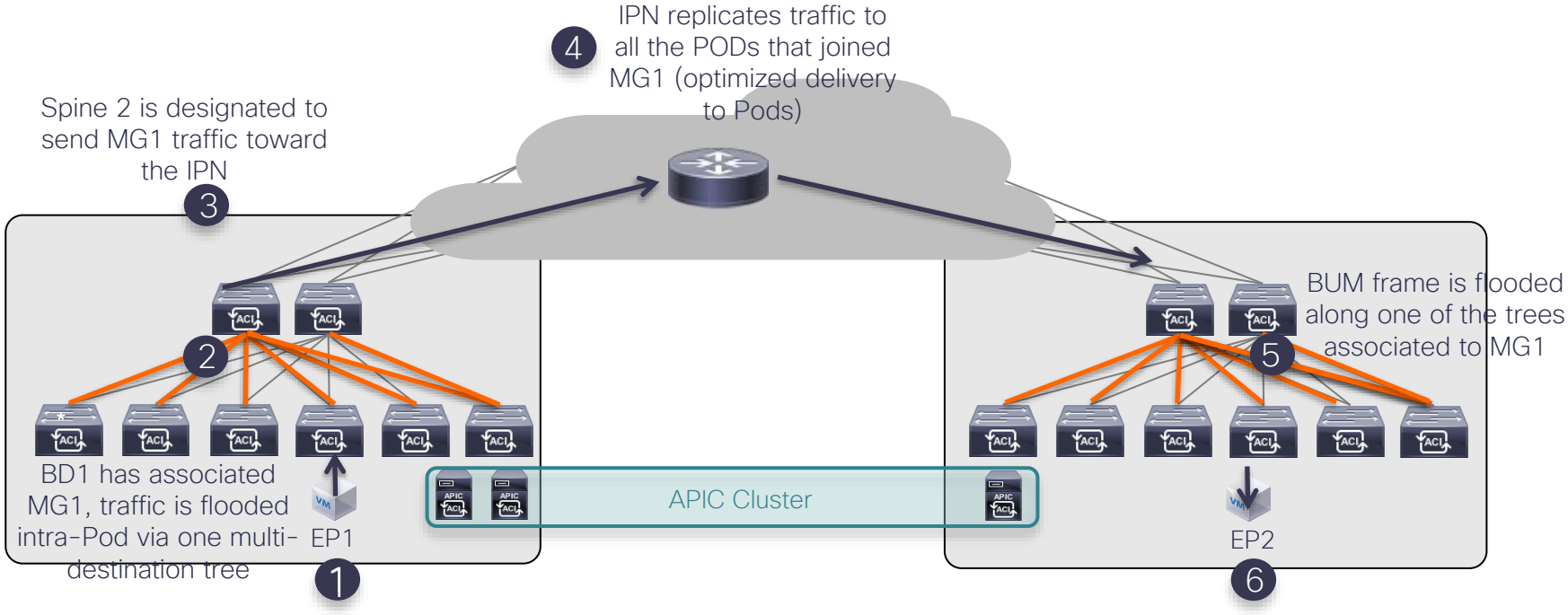


- Ingress replication for BUM* traffic not supported with Multi-Pod
- PIM Bidir is the only validated and supported option
  - Scalable: only a single (*,G) entry is created in the IPN for each BD
  - Fast-convergent: no requirement for data-driven multicast state creation
- A spine is elected authoritative for each Bridge Domain:
  - Generates an IGMP Join on a specific link toward the IPN
  - Always sends/receives BUM traffic on that link

BUM: Broadcast, Unknown Unicast, Multicast

ACI Multi-Pod

Use of Multicast for Inter-Pod BUM Traffic

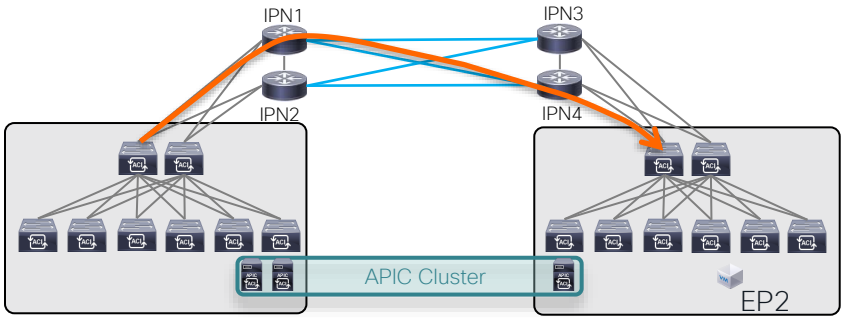


BUM: Layer 2 Broadcast, Unknown Unicast, Multicast

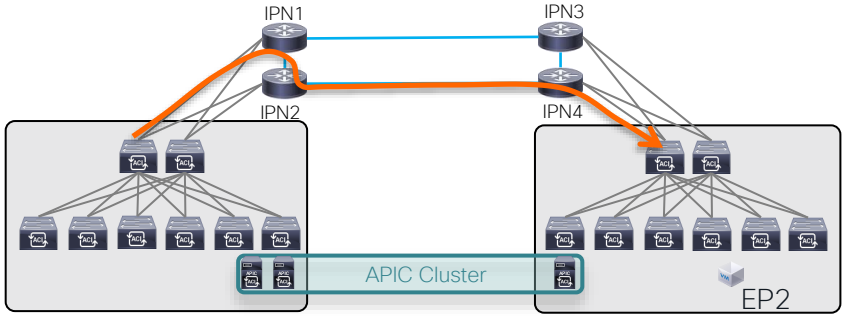
ACI Multi-Pod

PIM Bidir for BUM – Supported Topologies

Full Mesh between remote IPN devices

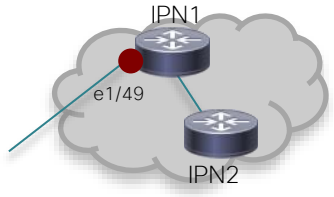


Directly connect local IPN devices



- Create full-mesh connections between IPN devices
- More costly for geo-dispersed Pods, as it requires more links between sites
- Alternatively, connect local IPN devices with a port-channel interface (for resiliency)
- In both cases, it is **critical** to ensure that the preferred path toward the RP from any IPN devices is not via a spine
- Recommendation is to increase the OSPF cost of the interfaces between IPN and spines

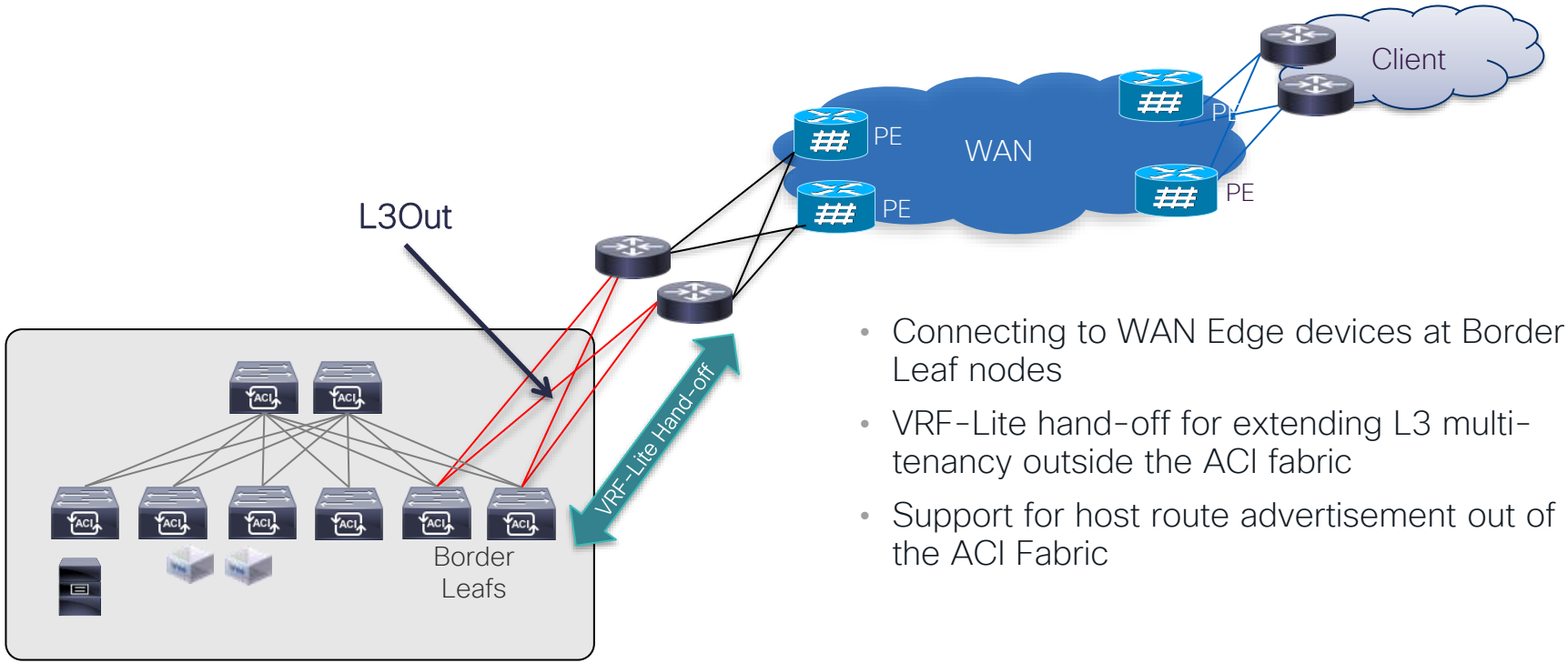
```
interface Ethernet1/49.4
  description L3 Link to Pod1-Spine1
  mtu 9150
  encapsulation dot1q 4
  ip address 192.168.1.1/31
  ip ospf cost 100
  ip ospf network point-to-point
  ip router ospf IPN area 0.0.0.0
  ip pim sparse-mode
  ip dhcp relay address 10.1.0.2
  ip dhcp relay address 10.1.0.3
```



Connecting to the External Layer 3 Domain

Connecting ACI to Layer 3 Domain

‘Traditional’ L3Out on the BL Nodes



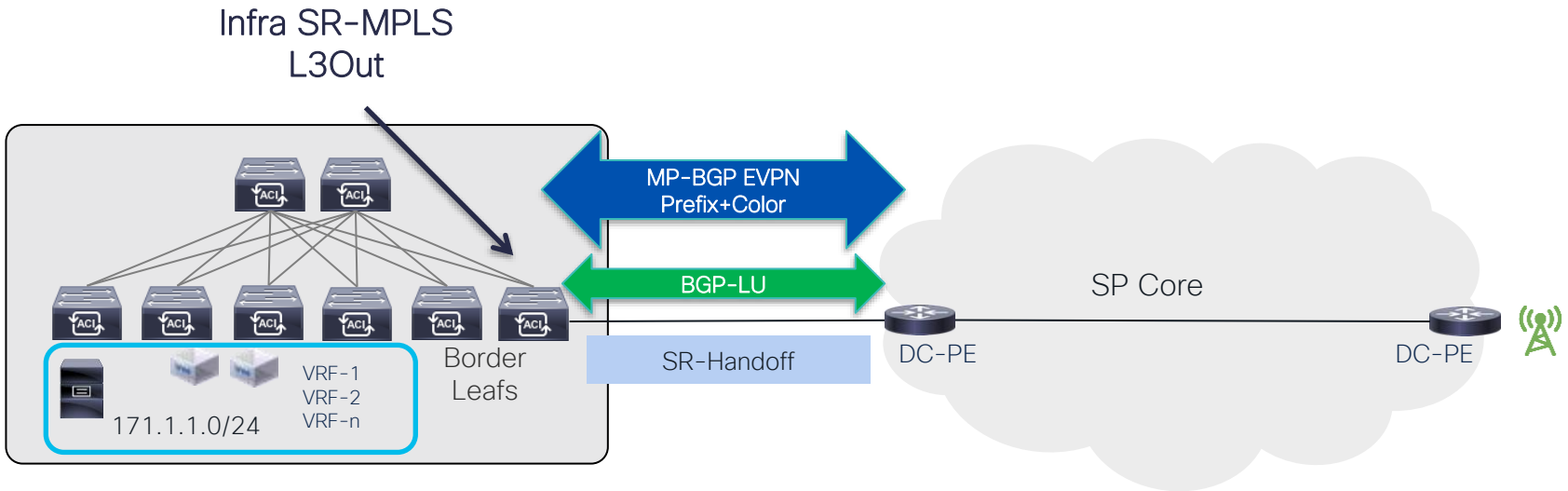
- Connecting to WAN Edge devices at Border Leaf nodes
- VRF-Lite hand-off for extending L3 multi-tenancy outside the ACI fabric
- Support for host route advertisement out of the ACI Fabric

Connecting ACI to Layer 3 Domain

‘SR-MPLS Handoff’

ACI Release 5.0(1)

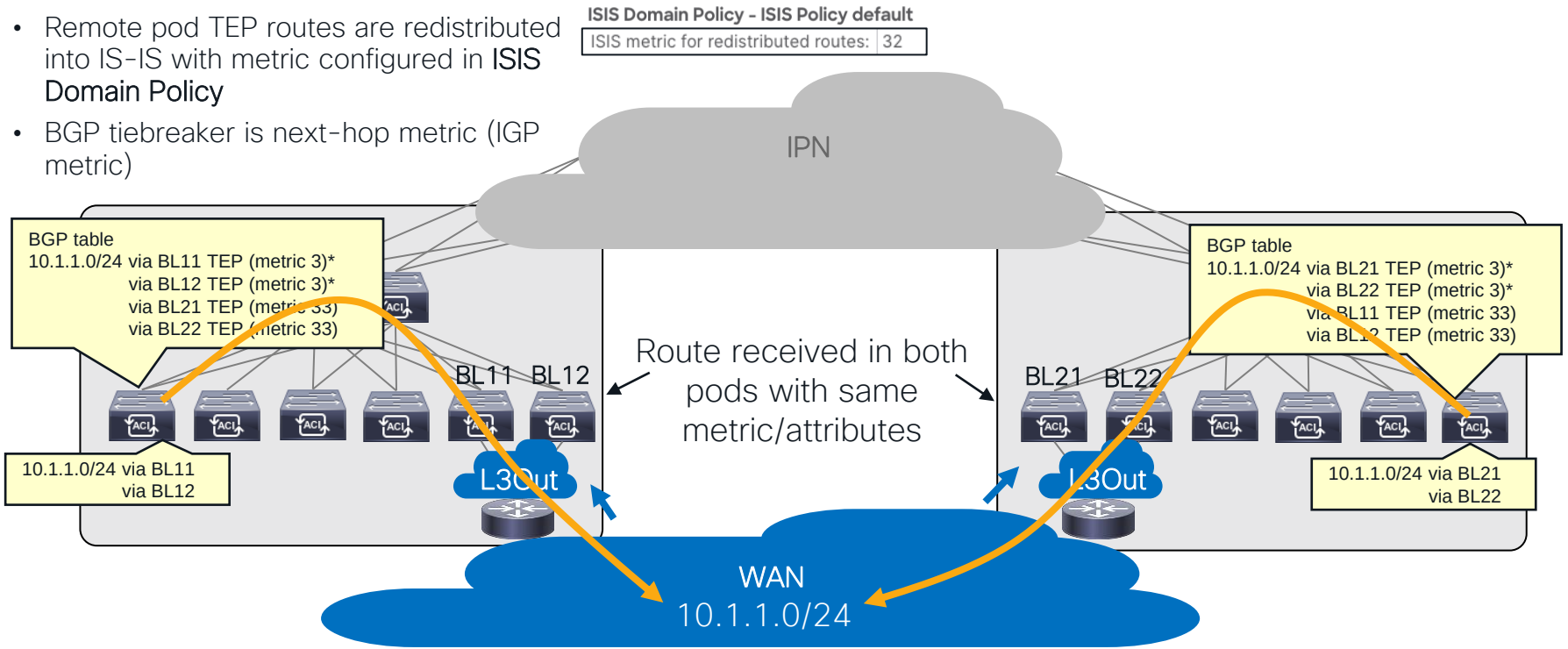
- Border Leafs connect to PE router in SP core
- Single BGP EVPN session for all VRFs
- ACI BL is advertising EVPN type-5 routes with BGP color community



Connecting to the External L3 Domain

Local L3Outs preferred over L3Outs in remote pods

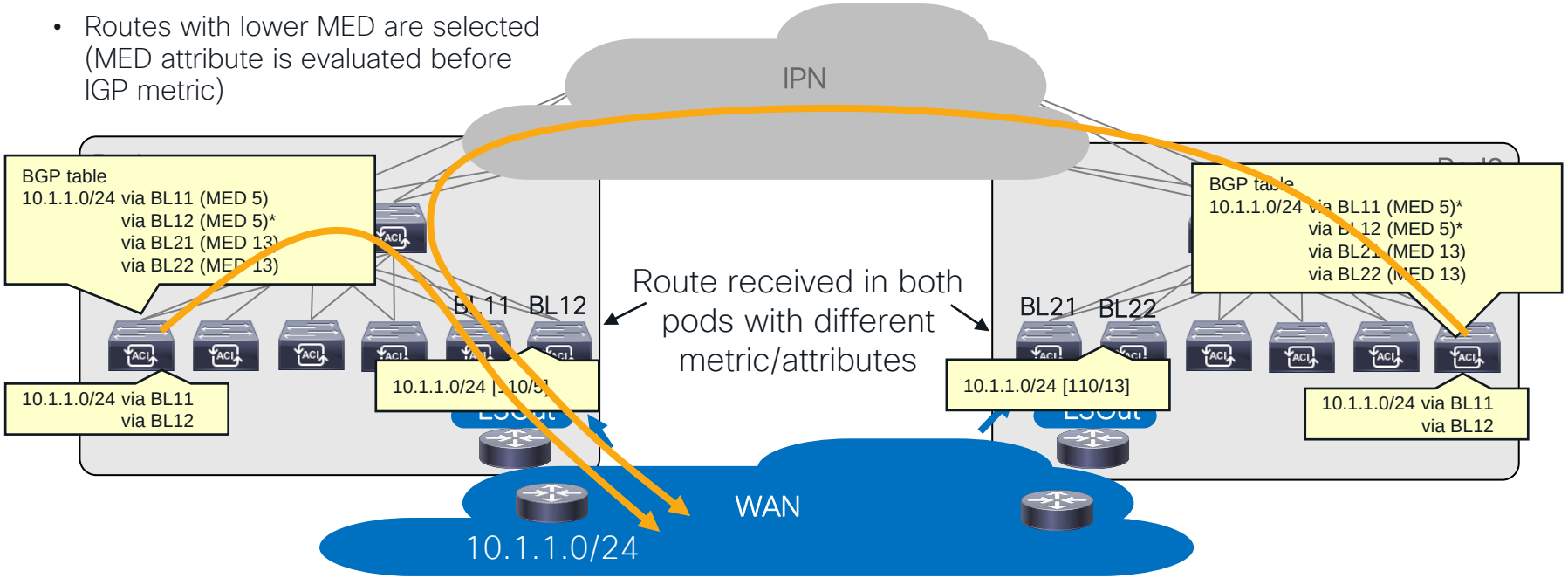
- Remote pod TEP routes are redistributed into IS-IS with metric configured in **ISIS Domain Policy**
- BGP tiebreaker is next-hop metric (IGP metric)



Connecting to the External L3 Domain

Remote pod L3Out may be used if it has a better external metric

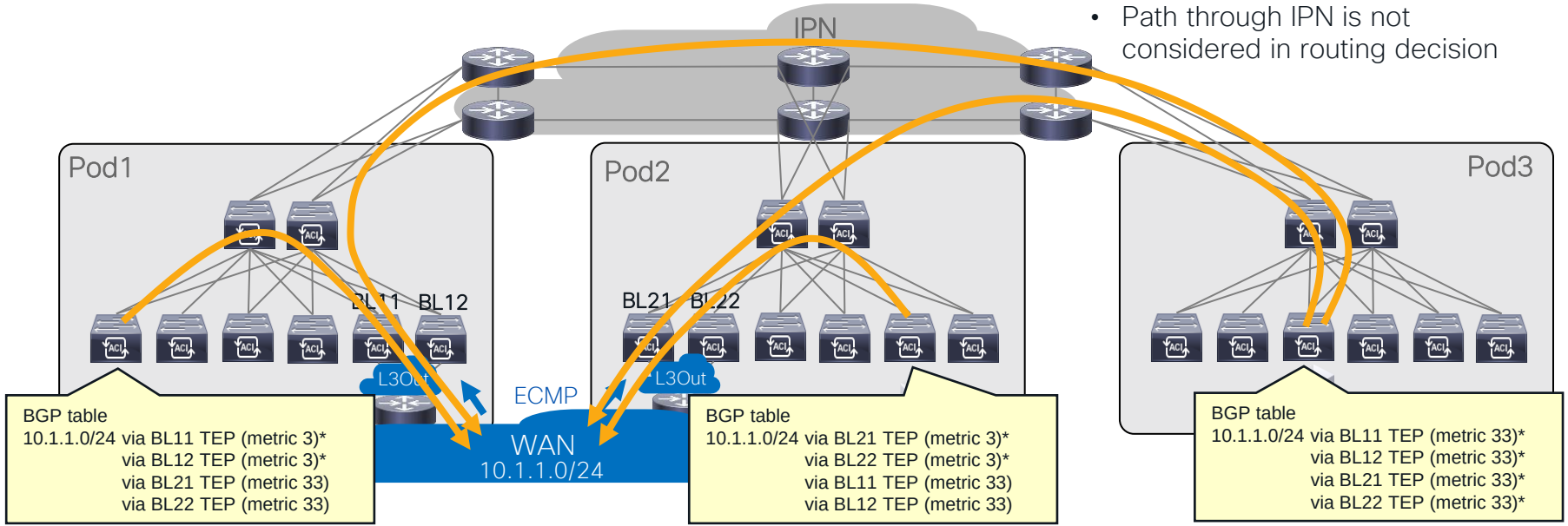
- BGP MED is set to OSPF metric when redistributed into fabric
- Routes with lower MED are selected (MED attribute is evaluated before IGP metric)



Connecting Multi-Pod to the Layer 3 Domain

What happens when there are more than two pods?

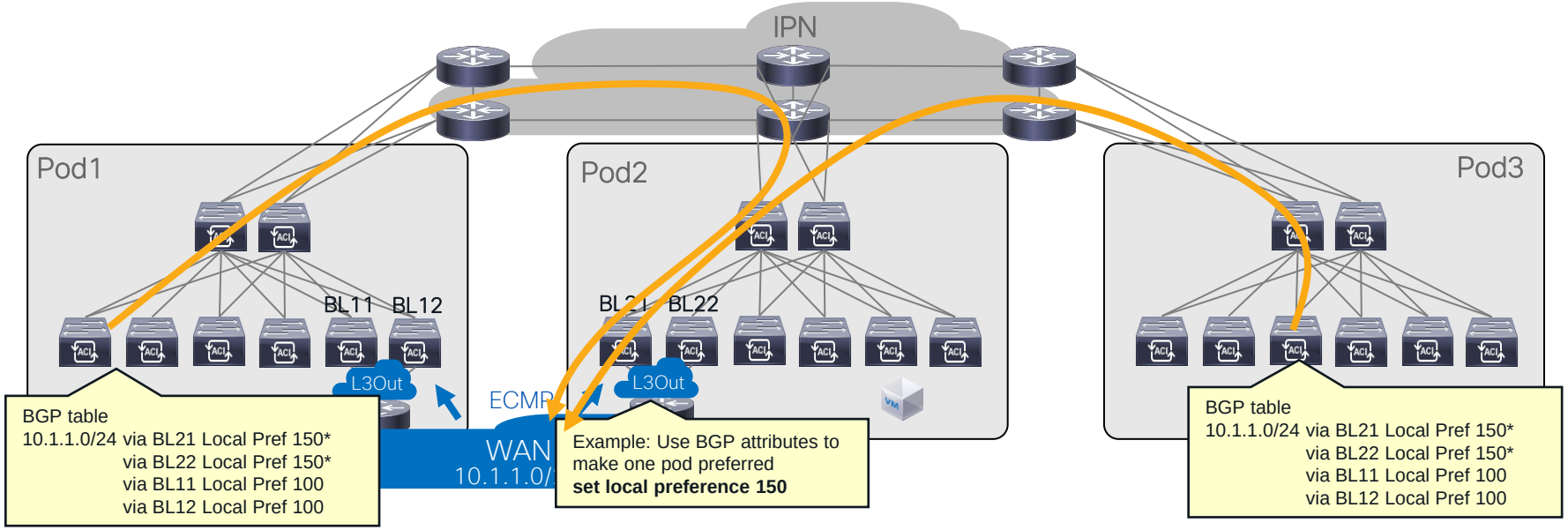
- Traffic flows are load balanced across all remote pods
- Path through IPN is not considered in routing decision



- A pod does not need a dedicated L3Out. Flows to external destinations can use an L3Out in another pod

Connecting Multi-Pod to the Layer 3 Domain

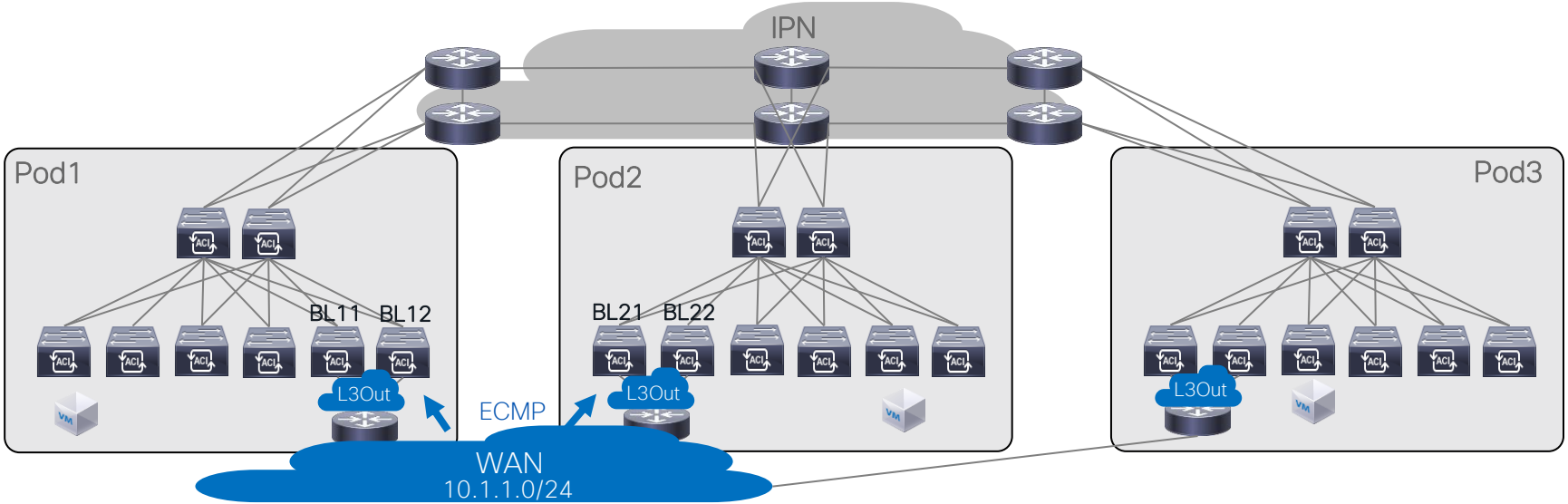
How to prefer one remote pod over another?



But change will affect all pods!

Connecting Multi-Pod to the Layer 3 Domain

How to prefer one remote pod over another?

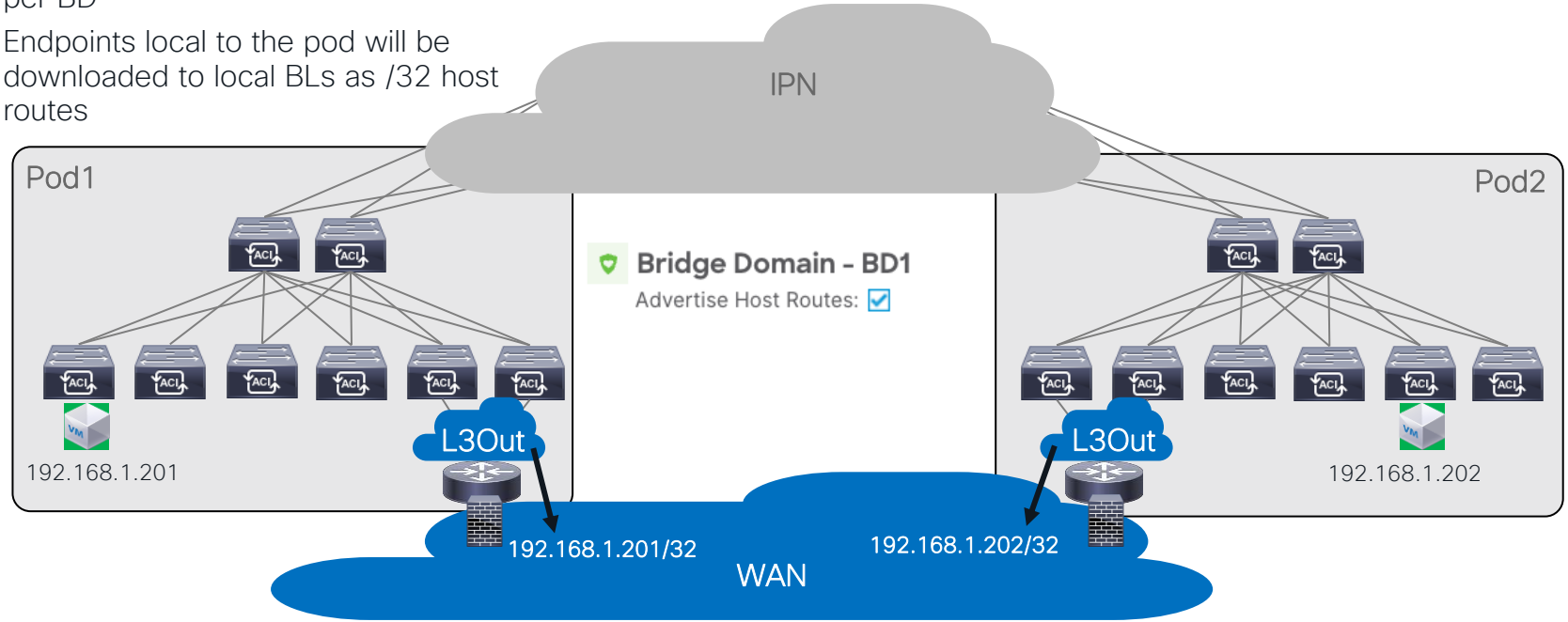


Adding a local L3out may be a better option

Connecting to the External L3 Domain

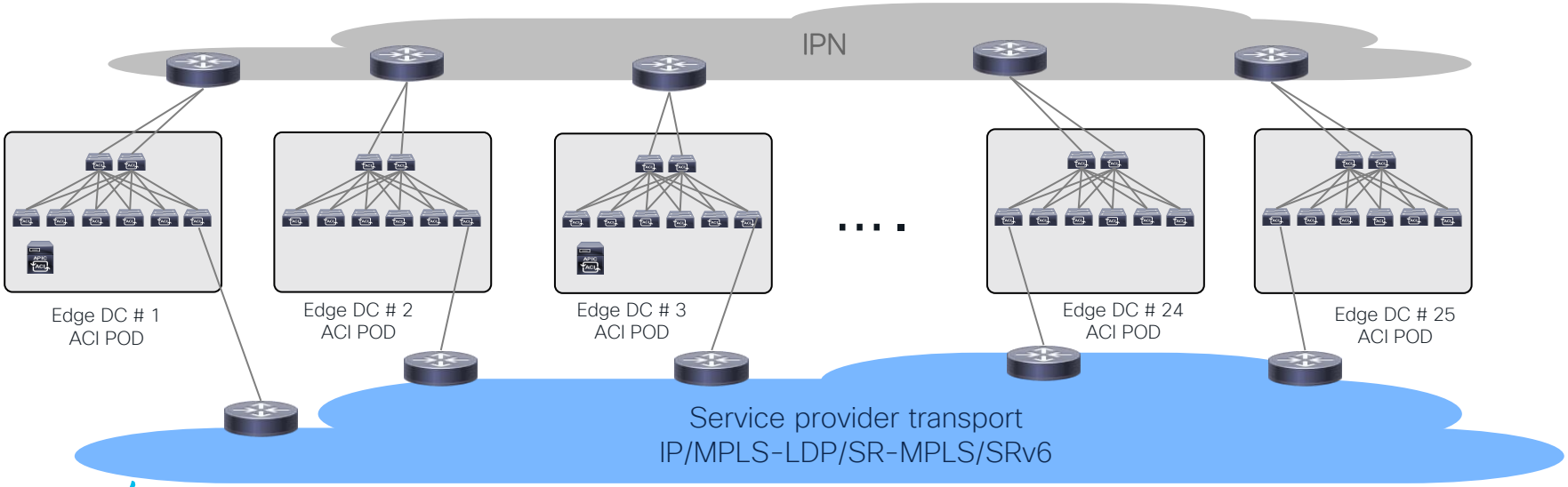
Influencing inbound path: Host route advertisement

- Host route advertisement can be enabled per BD
- Endpoints local to the pod will be downloaded to local BLs as /32 host routes



Edge DCs with Multi-Pod architecture

- APIC controllers are needed only in some Pods
- Communication across Pods is typically through SR-MPLS L3out
- 25 Pods per fabric is supported starting 6.0(1) release
- Leaf scale per fabric remains same. 2 Spines per Pod is supported
- Latency requirement remains same – 50 msec RTT requirement across APIC clusters and between switches and APIC
- No need to enable PIM-Bidir in IPN if L2 extension across Pod is not required



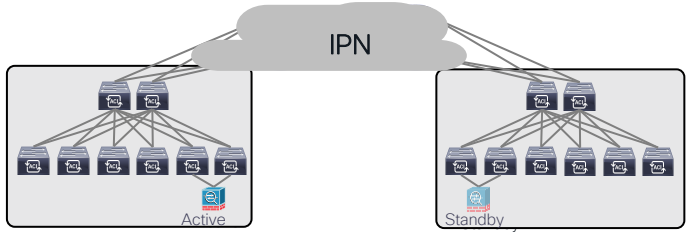
Network Services Integration

ACI Multi-Pod

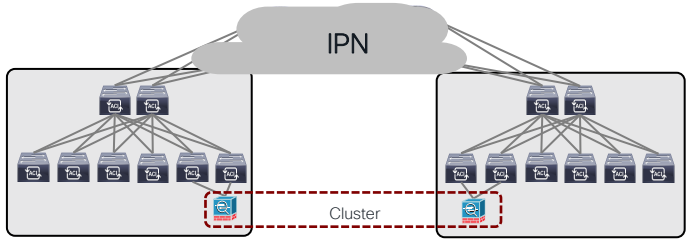
Design options

For your reference

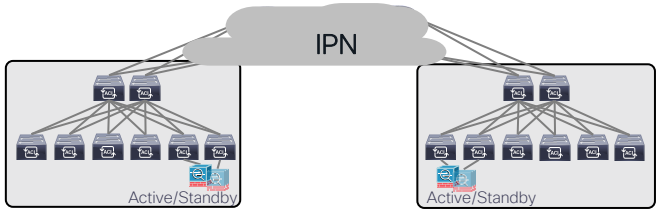
Typical options for an Active/Active DC use case



- Active and Standby pair deployed across Pods
- No issues with asymmetric flows



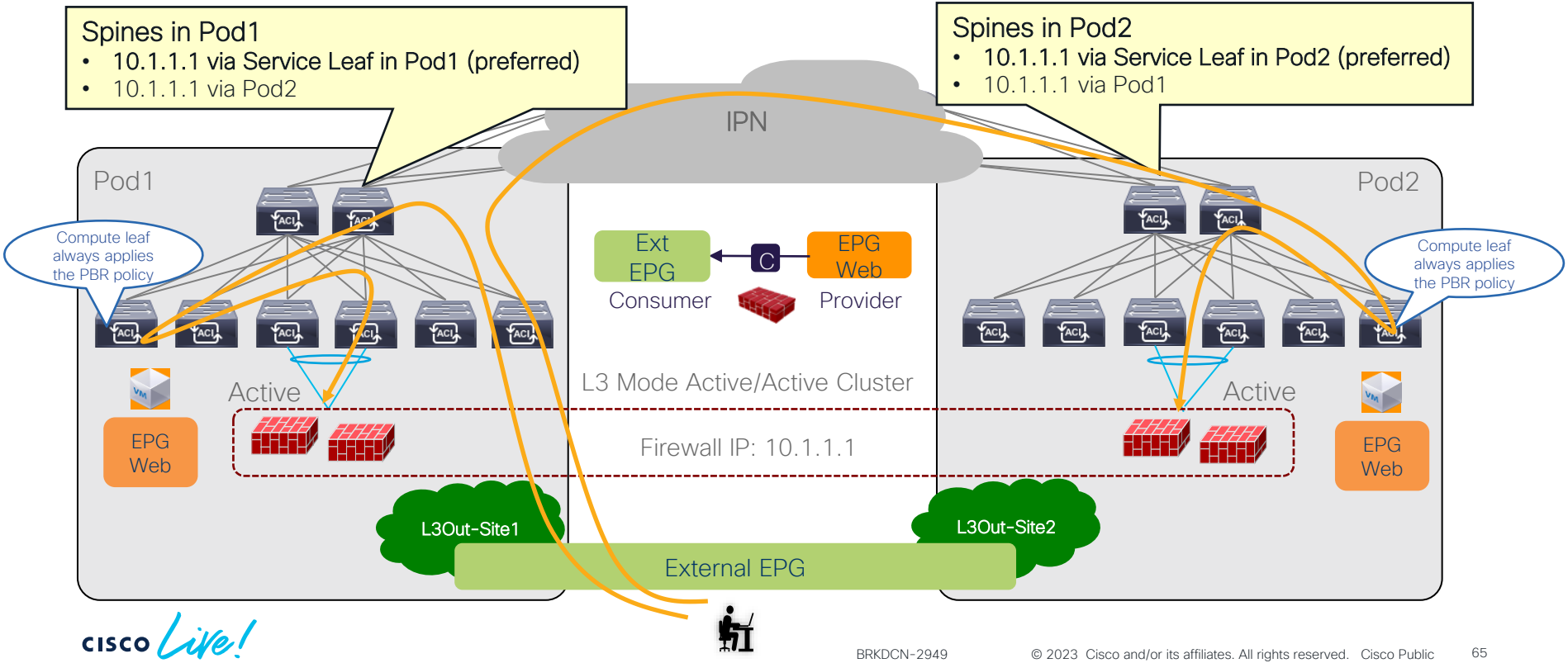
- Active/Active FW cluster nodes stretched across Sites (single logical FW)
- Requires the ability of discovering the same MAC/IP info in separate sites at the same time
- Supported from ACI release 3.2(4d) with the use of Service-Graph with PBR



- Independent Active/Standby pairs deployed in separate Pods
- Use of Symmetric PBR to avoid the creation of asymmetric paths crossing different active FW nodes

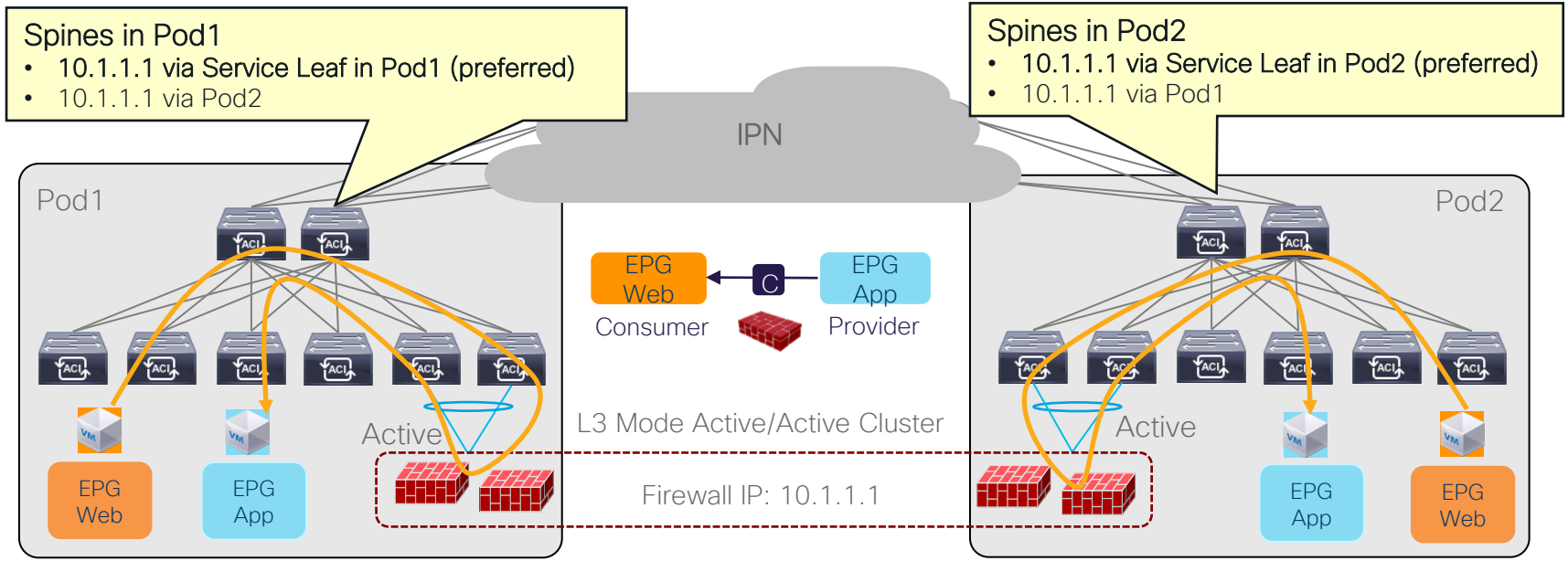
ACI Multi-Pod: Active/Active cluster across pods

North-South Traffic Flow



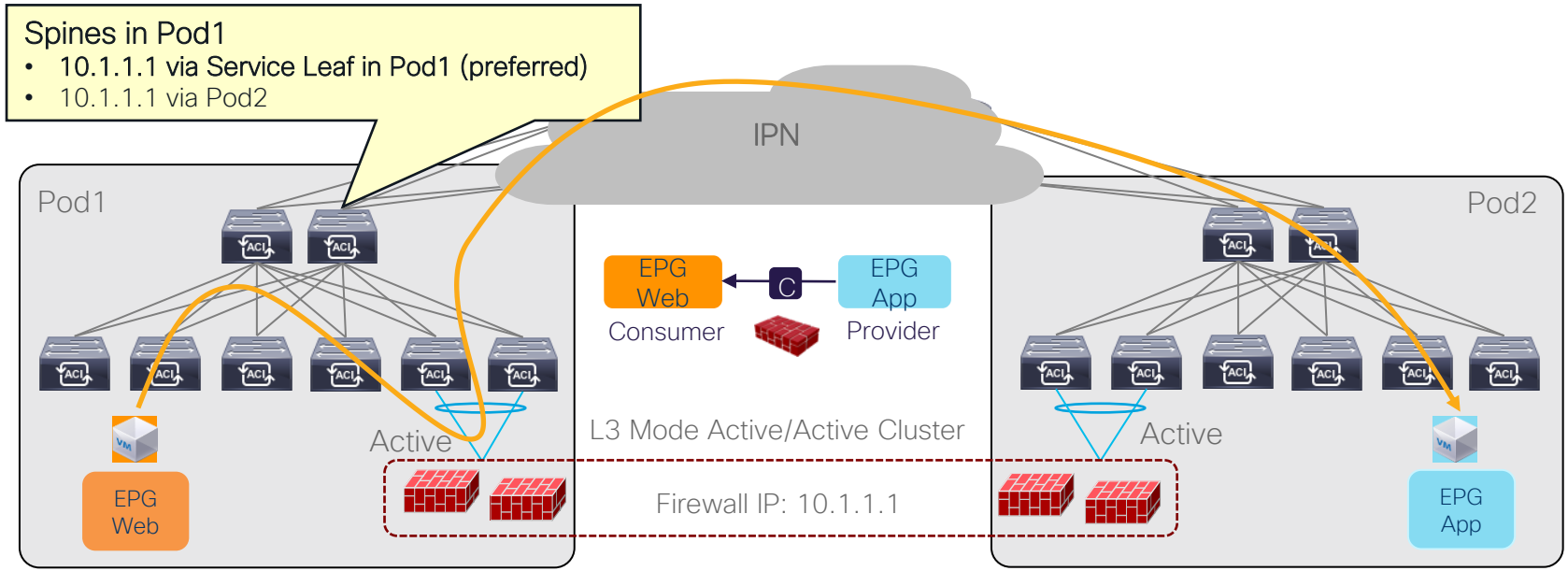
ACI Multi-Pod: Active/Active cluster across pods

East-West Traffic Flow (Intra-Pod)



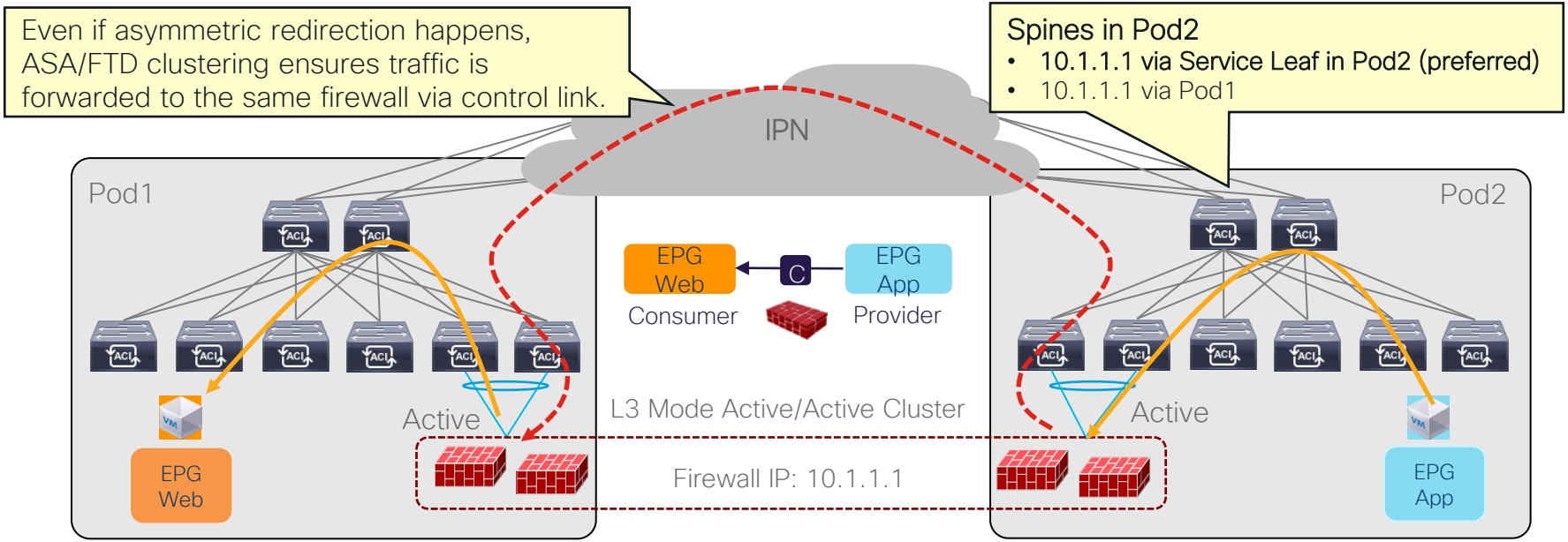
ACI Multi-Pod: Active/Active cluster across pods

East-West Traffic Flow (Inter-Pod) incoming traffic



ACI Multi-Pod: Active/Active cluster across pods

East-West Traffic Flow (Inter-Pod) return traffic

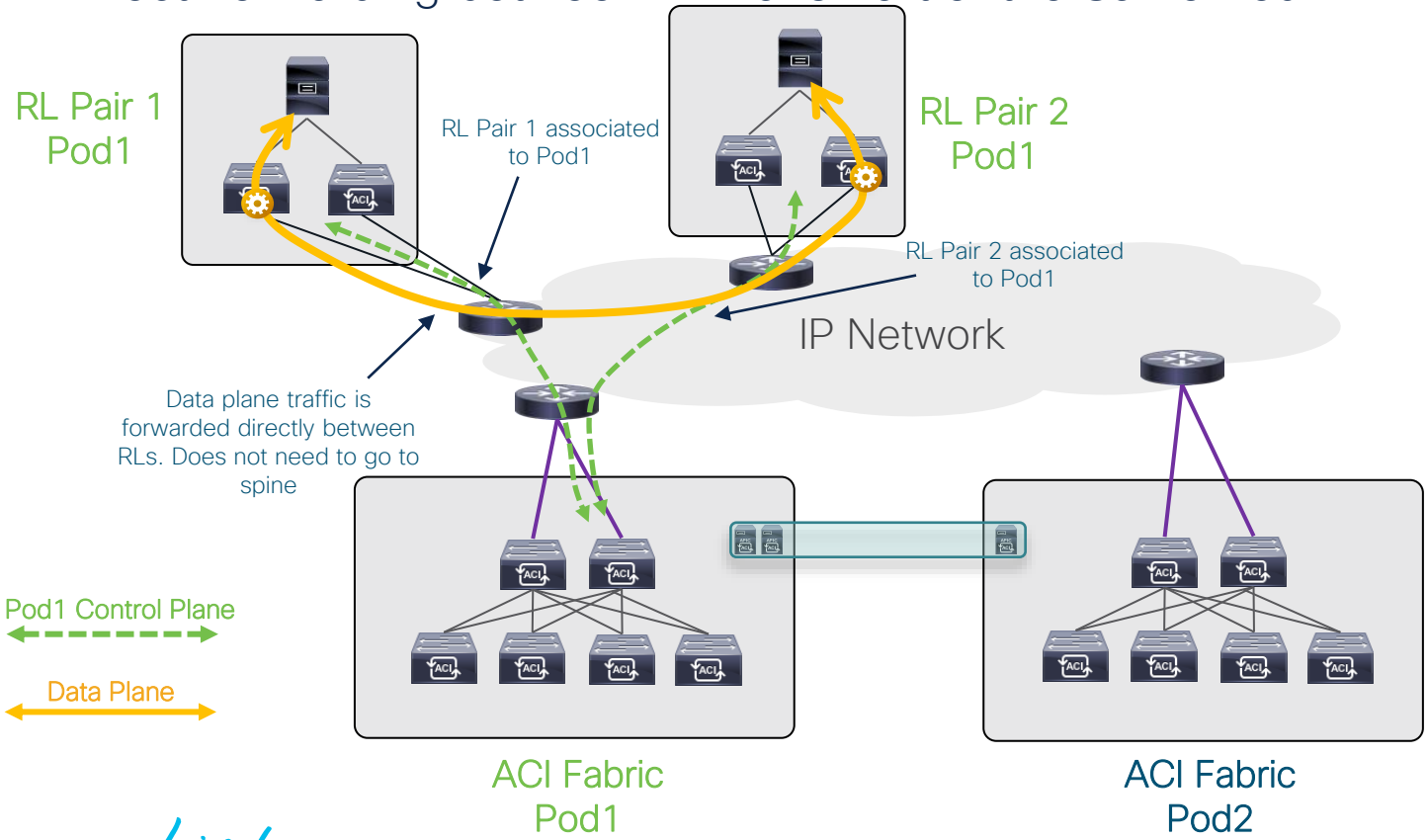


Multi-Pod with Remote Leaf

cisco Live!

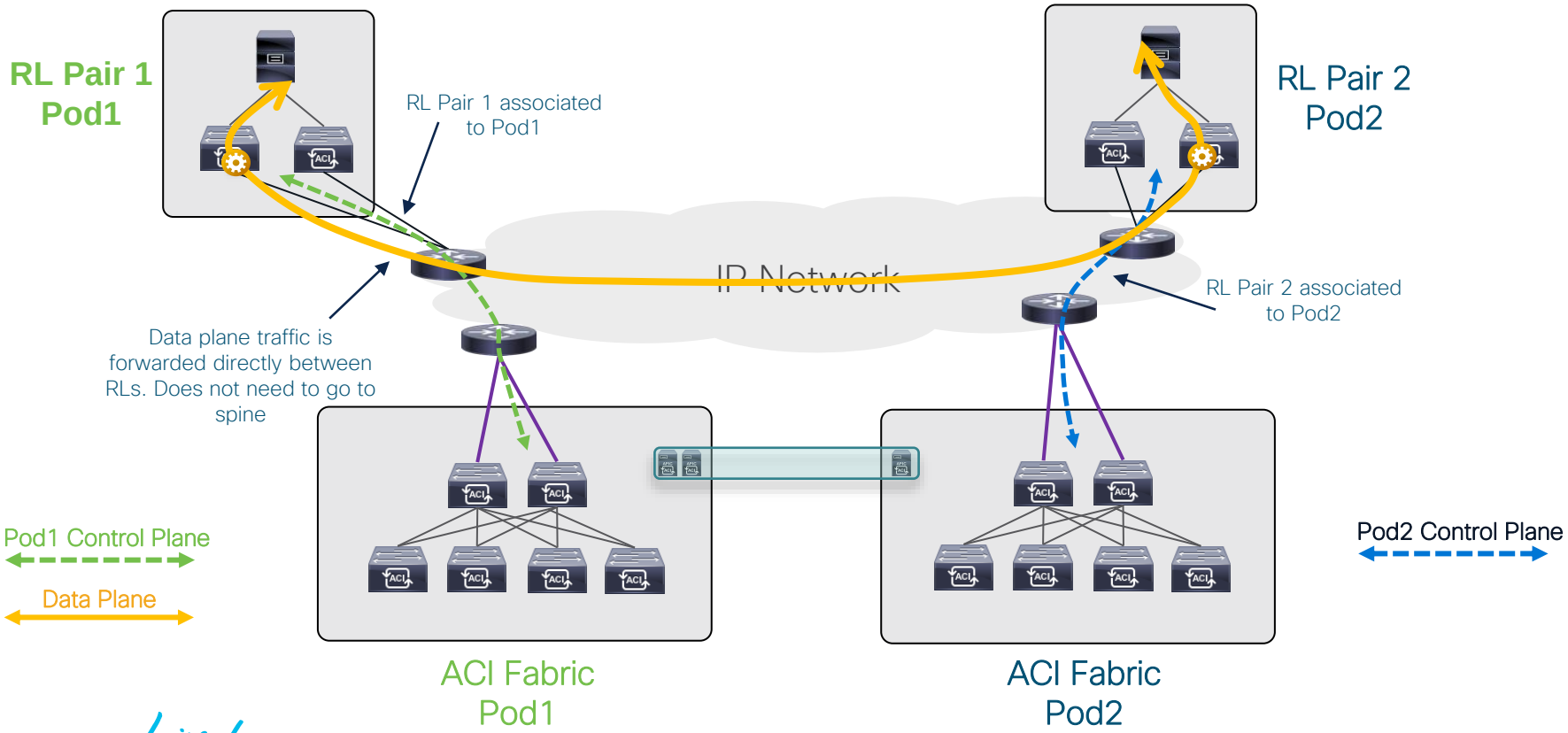
ACI Remote Leaf with Multi-Pod

Direct Forwarding between RL Pairs Part of the Same Pod



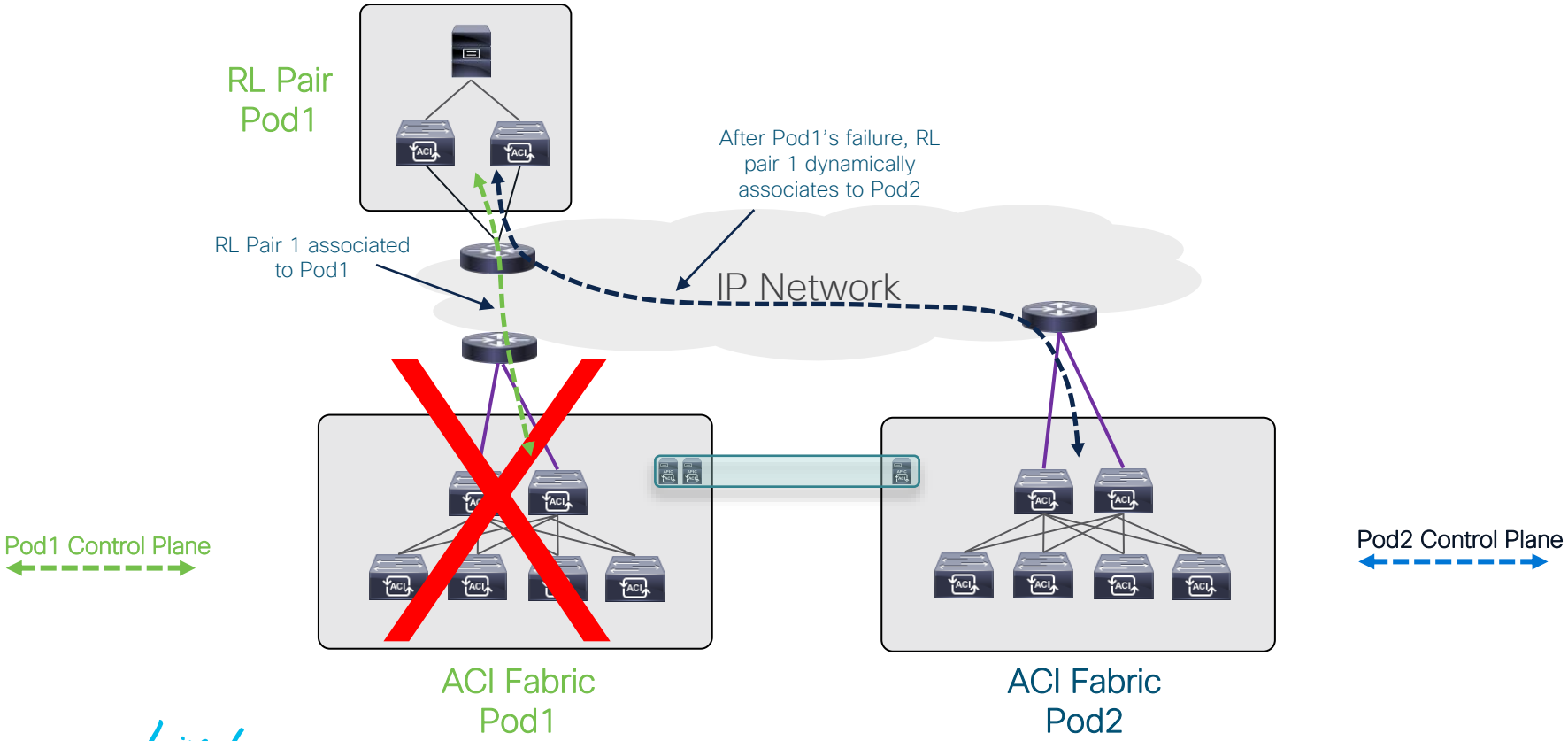
ACI Remote Leaf with Multi-Pod

Direct Forwarding between RL Pairs Part of Different Pods



ACI Remote Physical Leaf

RL Pair Resiliency in a Pod Failure Scenario



Useful Links

- ✓ ACI Multi-Pod White Paper

  http://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-737855.html?cachemode=refresh

- ✓ ACI Multi-Pod Configuration Paper

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

- ✓ ACI Multi-Pod and Service Node Integration White Paper

  https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-739571.html

Complete your Session Survey

- Please complete your session survey after each session. Your feedback is important.
- Complete a minimum of 4 session surveys and the Overall Conference survey (open from Thursday) to receive your Cisco Live t-shirt.

- All surveys can be taken in the Cisco Events Mobile App or by logging in to the Session Catalog and clicking the "Attendee Dashboard" at https://www.ciscolive.com/emea/learn/sessions/session-catalog.html

Continue Your Education

Visit the Cisco Showcase for related demos.

Book your one-on-one Meet the Engineer meeting.

Attend any of the related sessions at the DevNet, Capture the Flag, and Walk-in Labs zones.

Visit the On-Demand Library for more sessions at ciscolive.com/on-demand.


cisco
The bridge to possible
Thank you
cisco Live!


CISCO Live!
ALL IN