Cisco ACI Design and Deployment with Multi-Pod and Multi-Site

Maurizio Portolani
Technical Marketing Engineer, Data Center Networking

Session ID: BRKDCN-2617

Cisco Webex App

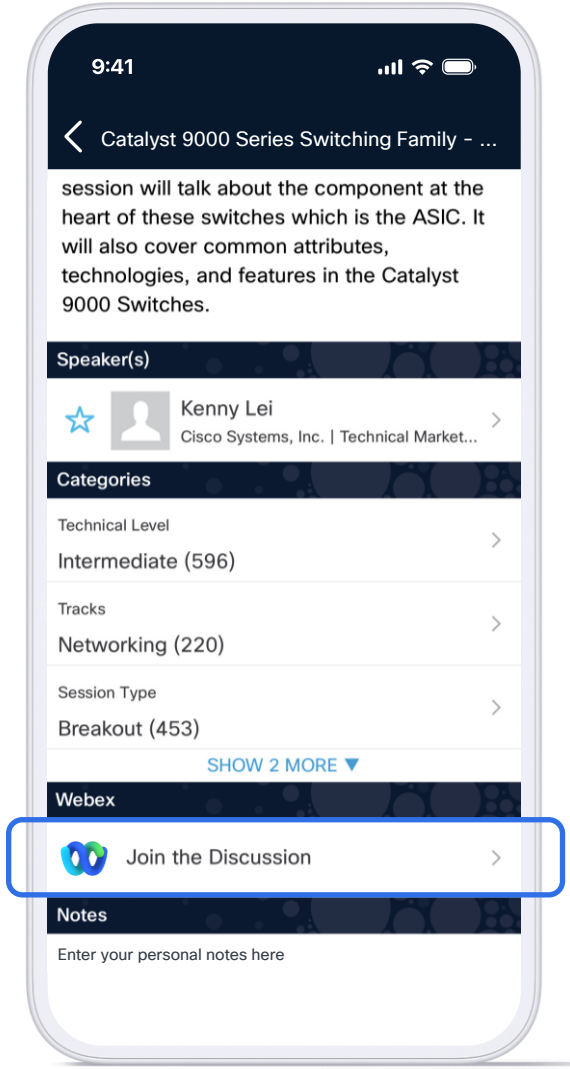
Questions?

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 by the speaker until 14 November 2025.



https://ciscolive.ciscoevents.com/
ciscolivebot/**#BRKDCN-2617**

© 2025 Cisco and/or its affiliates. All rights reserved.

Agenda

01 Introduction

02 IPN/ISN Requirements

03 Multi-Pod APIC Cluster Deployment Considerations

04 ND Orchestrator (NDO) Deployment Considerations

05 Multi-Pod Control and Dataplane

06 Multi-Site Control and Dataplane

07 External L3 Domain with Multipod

08 External L3 Domain with Multi-Site

09 Network Services Integration with Multi-Pod

10 Network Services Integration with Multi-Site

© 2025 Cisco and/or its affiliates. All rights reserved.

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 and Multi-Site architectures
- Initial assumption:
  - The audience already has a good knowledge of ACI main concepts (Tenant, BD, EPG, L3Out, etc.)

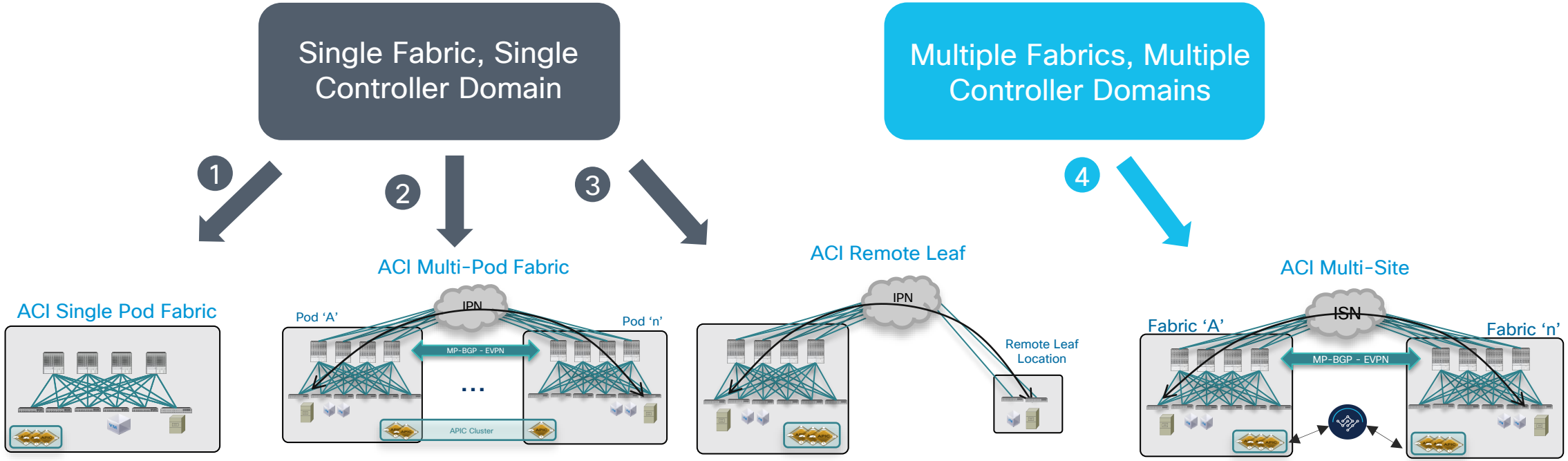
© 2025 Cisco and/or its affiliates. All rights reserved.

Introduction

© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Architectural Options

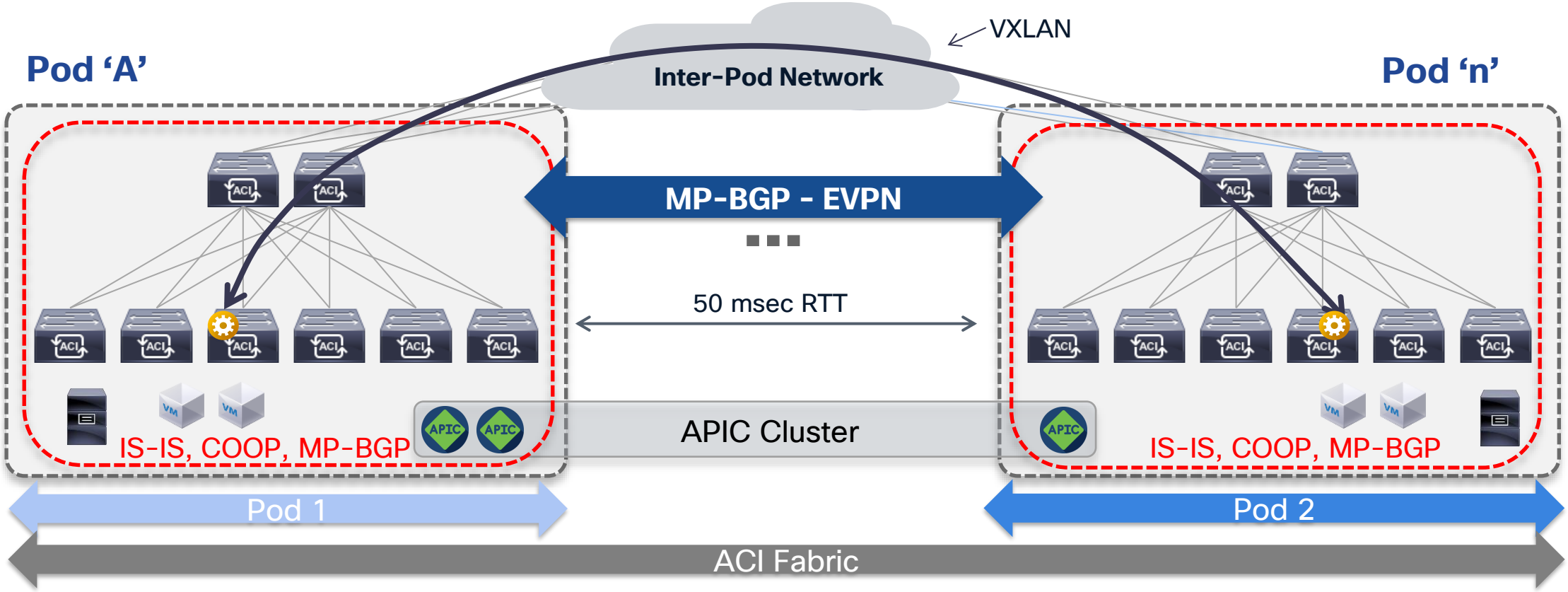
Fabric and Policy Domain Evolution



© 2025 Cisco and/or its affiliates. All rights reserved.

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

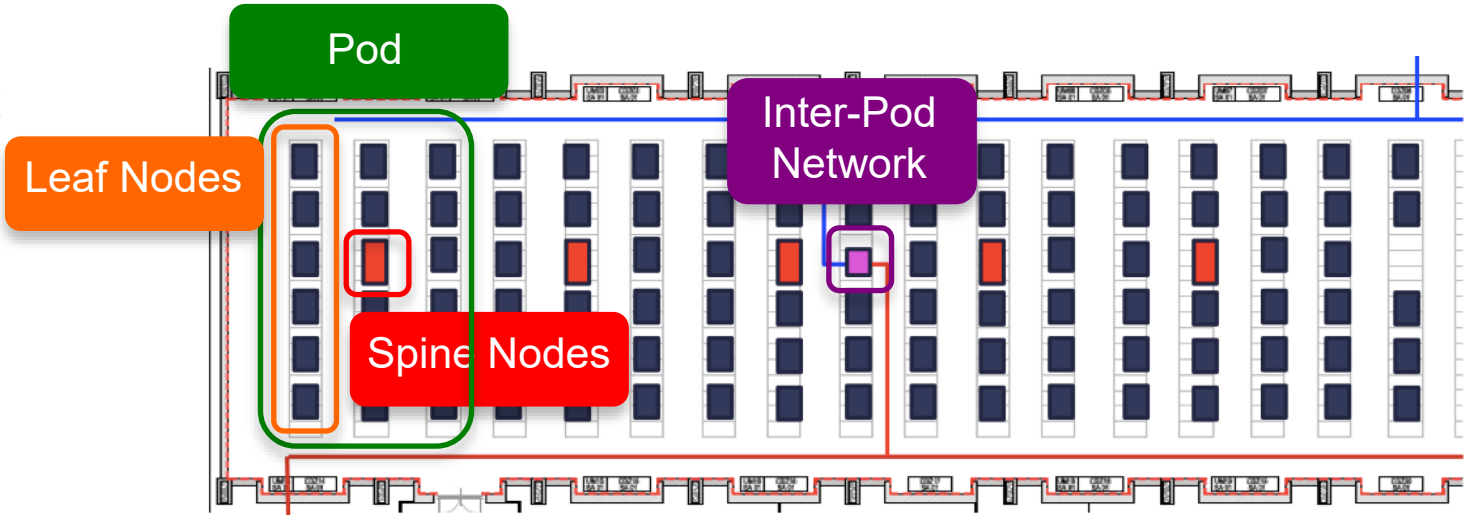
© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Pod

Use Cases

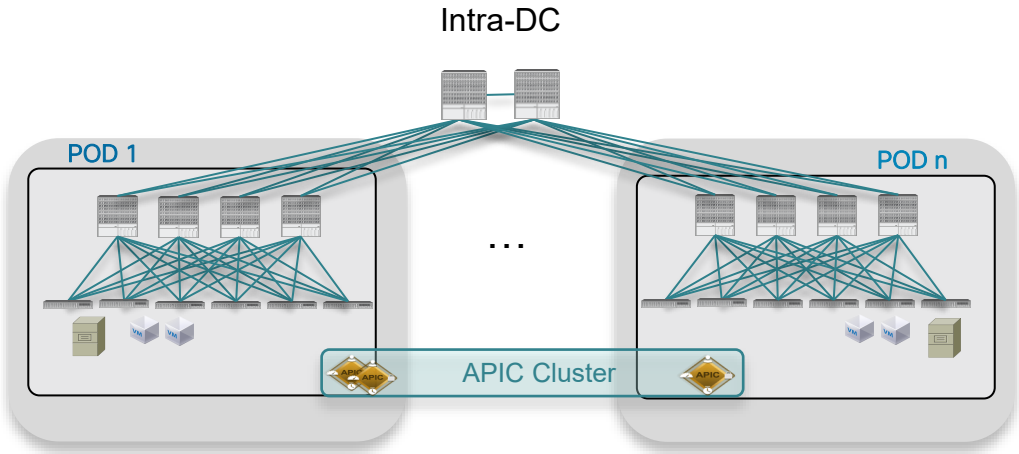
- Handling 3-tiers physical cabling layout

  Cable constrain (multiple buildings, campus, metro) requires a second tier of “spines”



- Building a distributed DC fabric

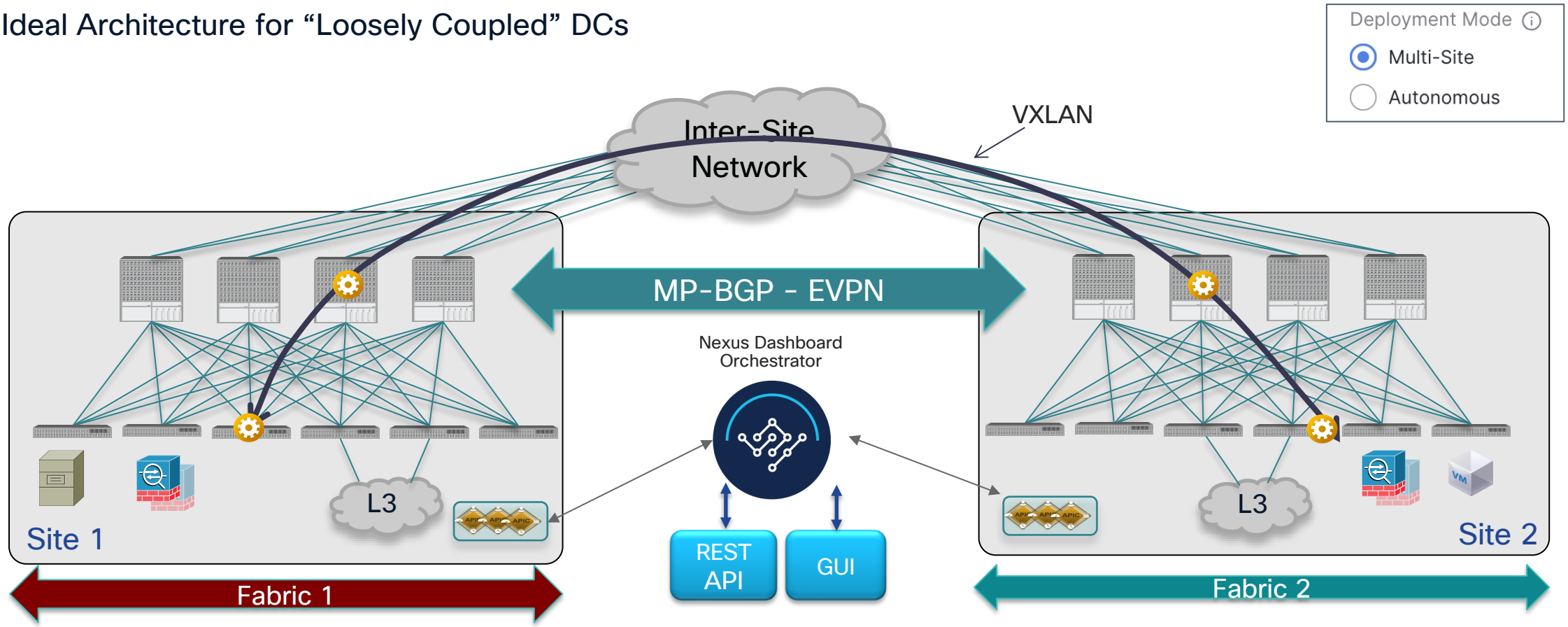
  Same DC location, Metro Area (dark fiber, DWDM), L3 core



© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Site

The Ideal Architecture for “Loosely Coupled” DCs



- Separate ACI Fabrics with independent APIC clusters
- No latency limitation between Fabrics
- Nexus Dashboard Orchestrator pushes cross-fabric configuration to multiple APIC clusters providing scoping of all configuration changes

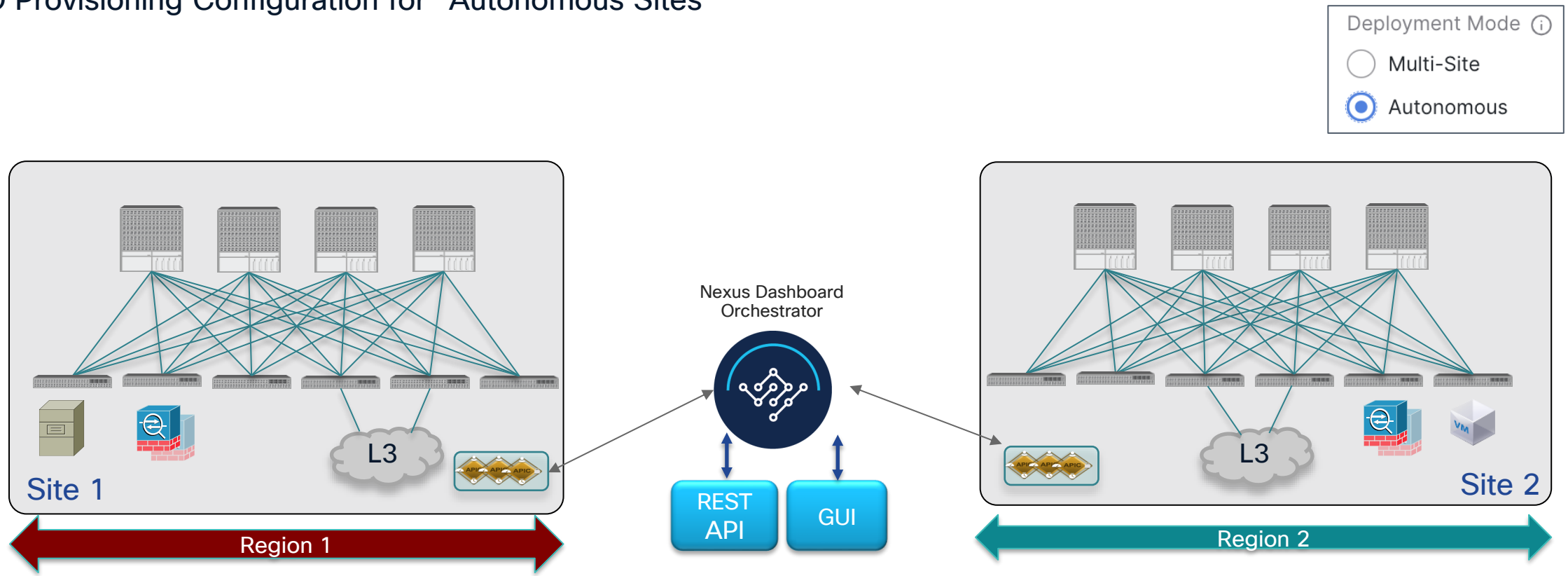
- MP-BGP EVPN control-plane between sites
- Data-Plane VXLAN encapsulation across sites
- End-to-end policy definition and enforcement

© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Site

Cisco Nexus Dashboard Orchestrator 4.x

NDO Provisioning Configuration for “Autonomous Sites”



- If the fabrics are operated as independent (“autonomous”) sites, NDO could still be used as a single point of provisioning
- No use of ISN and VXLAN EVPN for east-west communication

- Intersite Layer 3 communication still possible via the L3Out data path
  - NDO can be used to “replicate” configuration across sites by associating the same “autonomous template” to up to 100 fabrics

© 2025 Cisco and/or its affiliates. All rights reserved.

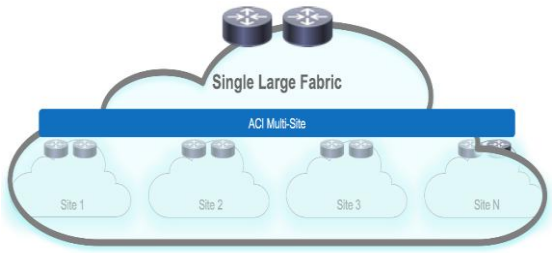
cisco

ACI Multi-Site Architecture

Most Common Use Cases

- Compartmentation/Scale

  Building Multiple Fabrics inside a single Data Center

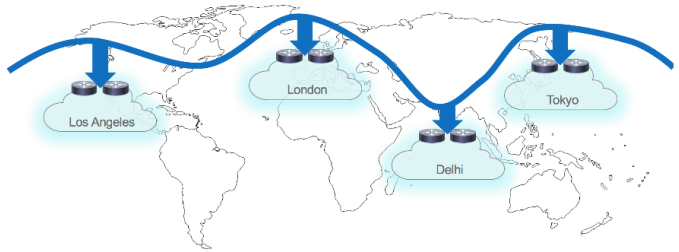


Optimized and controlled L2/L3 connectivity (including optimized/controlled BUM forwarding), scale out total number of leaf nodes (SP use case)

- Data Center Interconnect (DCI)

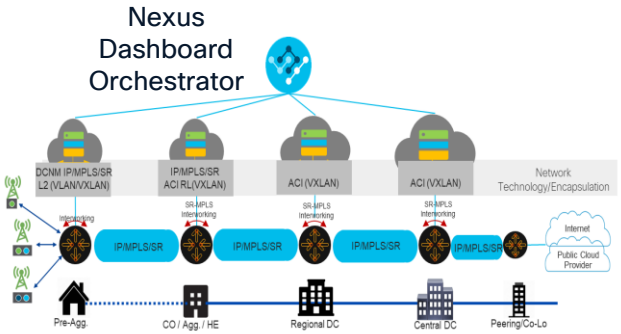
  Extend connectivity/policy between ‘loosely coupled’ DC sites

  Disaster Recovery and IP mobility use cases



- SP 5G Telco DC/Cloud*

  Centralized DC Orchestration for “Autonomous Fabrics”



*May also apply to Enterprise deployments

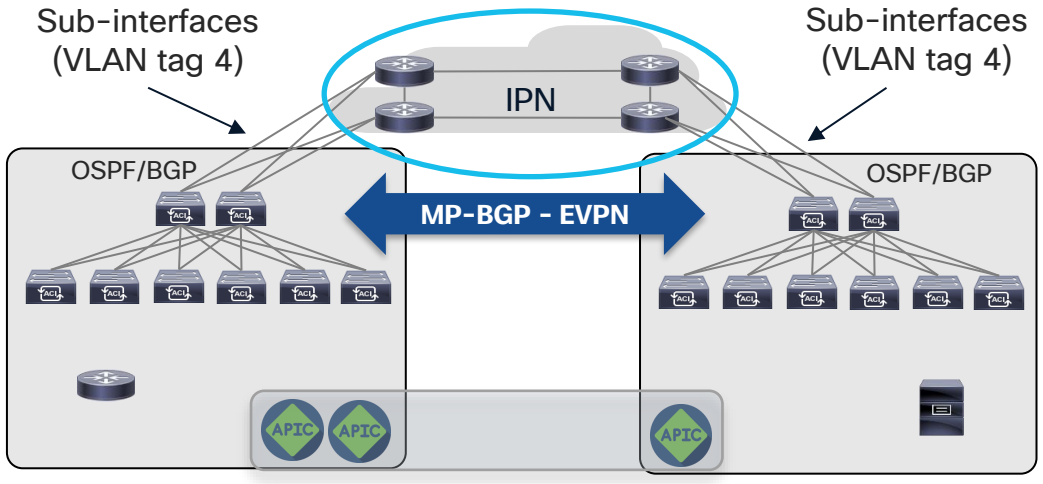
© 2025 Cisco and/or its affiliates. All rights reserved.

Inter-Pod and Inter-Site Connectivity Deployment Considerations

© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Pod

Inter-Pod Network (IPN) Requirements



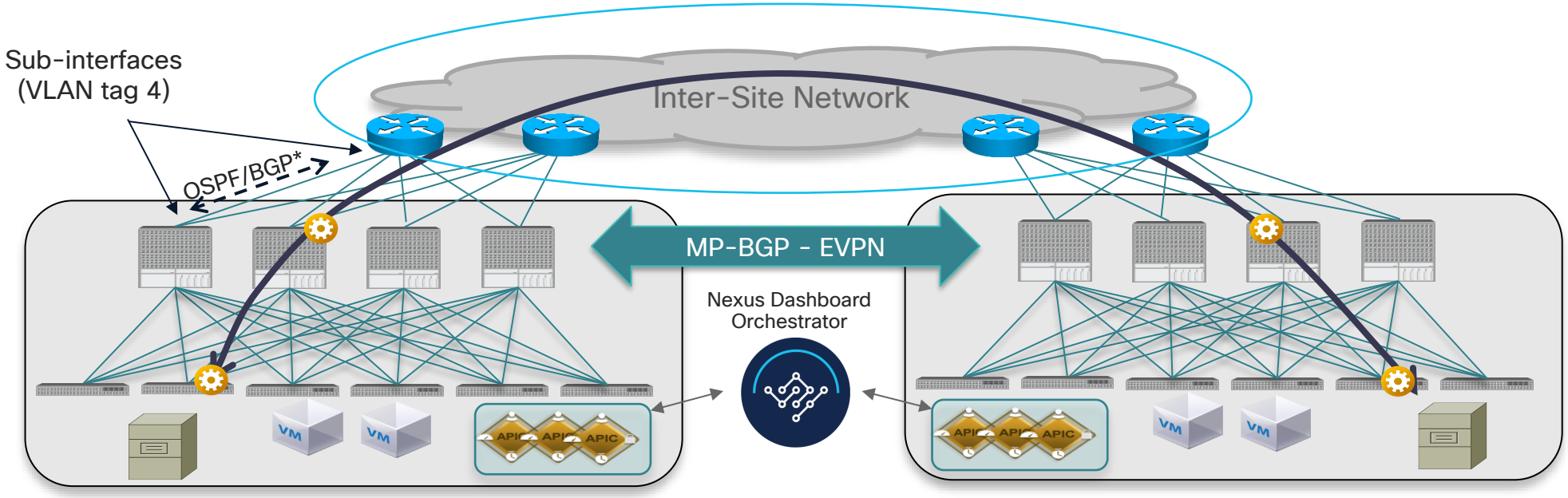
- **<u>Not managed</u> by APIC, must be separately configured (day-0 configuration)**
- **IPN topology can be arbitrary, not mandatory to connect to all spine nodes**
- **Main requirements:**
  - Multicast BiDir PIM → needed to handle Layer 2 BUM* traffic
  - OSPF or BGP to peer with the spine nodes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

* Broadcast, Unknown unicast, Multicast

© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Site

Inter-Site Network (ISN) Functional Requirements



- Not managed by APIC or NDO, must be independently configured (day-0 configuration)
- IP topology can be arbitrary, not mandatory to connect all the spine nodes to the ISN
- ISN main functional requirements:
  - ✓ OSPF/BGP* to peer with the spine nodes and exchange TEP address reachability

    Must use sub-interfaces (with VLAN tag 4) toward the spines
  - ✓ No multicast requirement for BUM traffic forwarding across sites
  - ✓ Increased end-to-end MTU support (at least 50/54 extra Bytes)

© 2025 Cisco and/or its affiliates. All rights reserved.

*Requires ACI 5.2(1)

cisco

What if the IPN/ISN Supports Only 1500B MTU Size?

© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Site/Multi-Pod and MTU Size

Different MTU Meanings

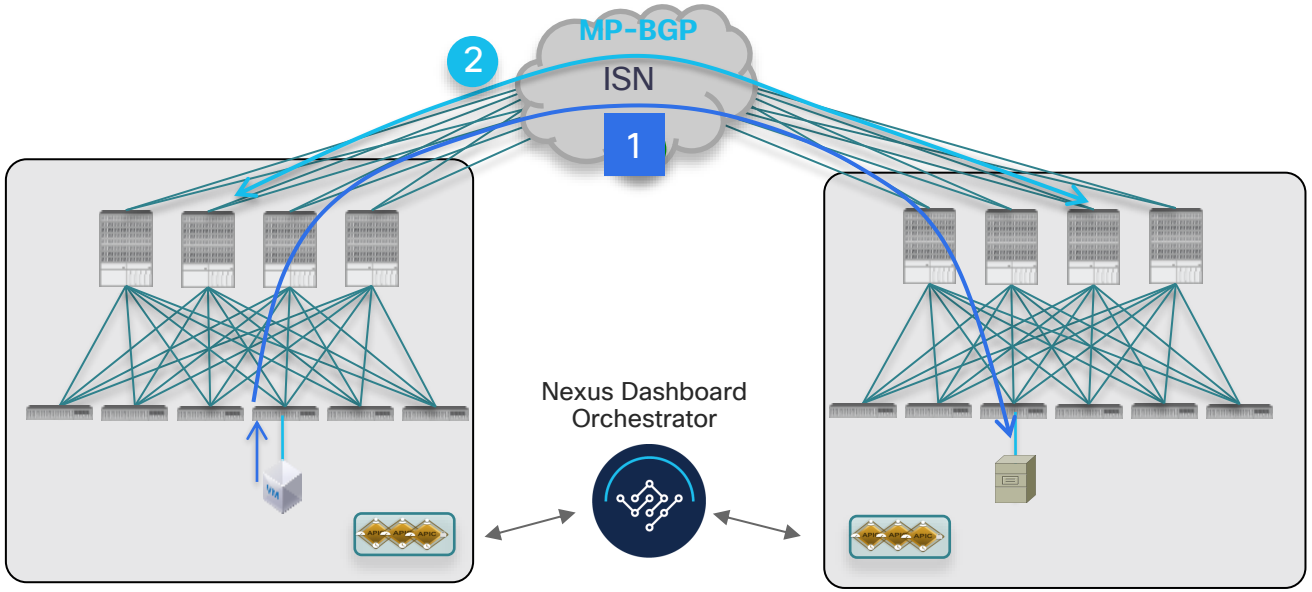
1. Data-Plane MTU: MTU of the traffic generated by endpoints (servers, routers, service nodes, etc.) connected to ACI leaf nodes

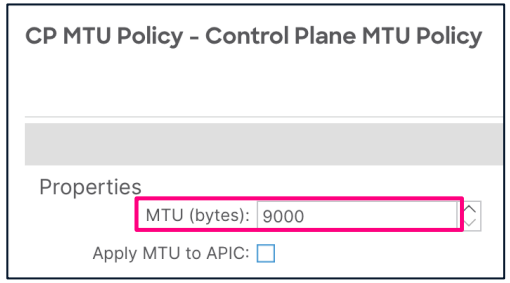
   Need to account for 50B of overhead (VXLAN encapsulation) for inter-site communication

2. Control-Plane MTU: for CPU generated traffic like MP-BGP sessions across sites

   Control plane traffic is not VXLAN encapsulated

   The default value is **9000B,** can be tuned on APIC to match the maximum MTU value supported in the ISN



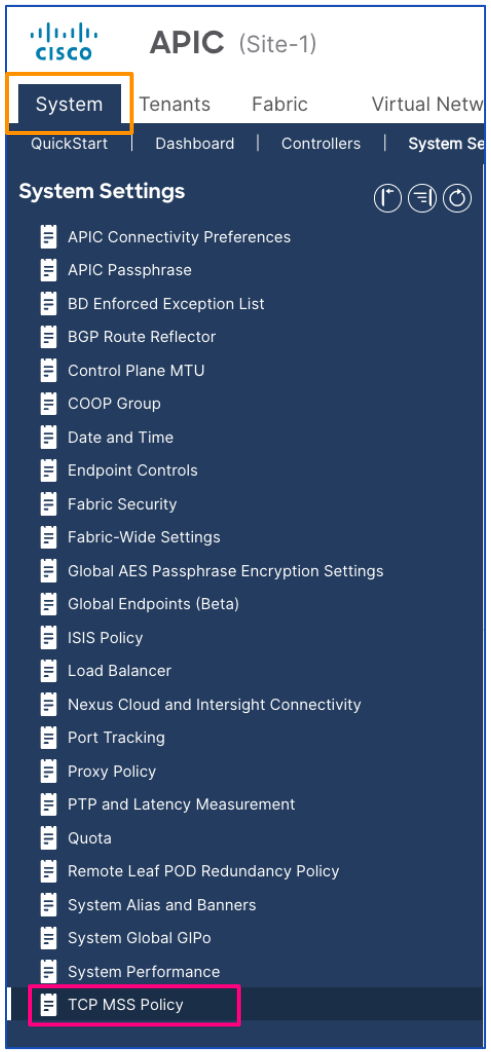


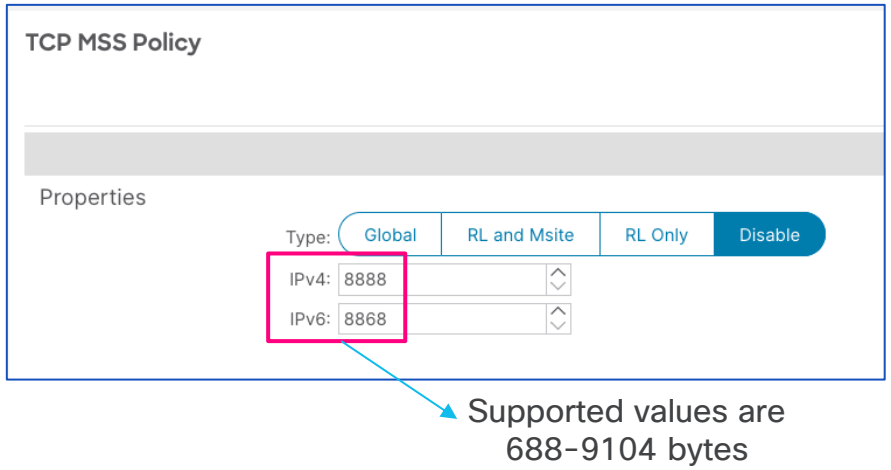
© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Site and MTU Size

Introducing the TCP-MSS Adjust Functionality

ACI Release 6.0(3)F





- TCP MSS adjust policy is enabled at System Settings level
- Supports different TCP MSS adjust setting for IPv4 and IPv6
- Supports three different options:
  1. **Global**: applies to all flows (Multi-Pod, Multi-Site, Remote Leafs)
  2. **RL and Msite**: applies to Multi-Site and Remote Leafs flows
  3. **RL Only**: applies only to Remote Leafs flows

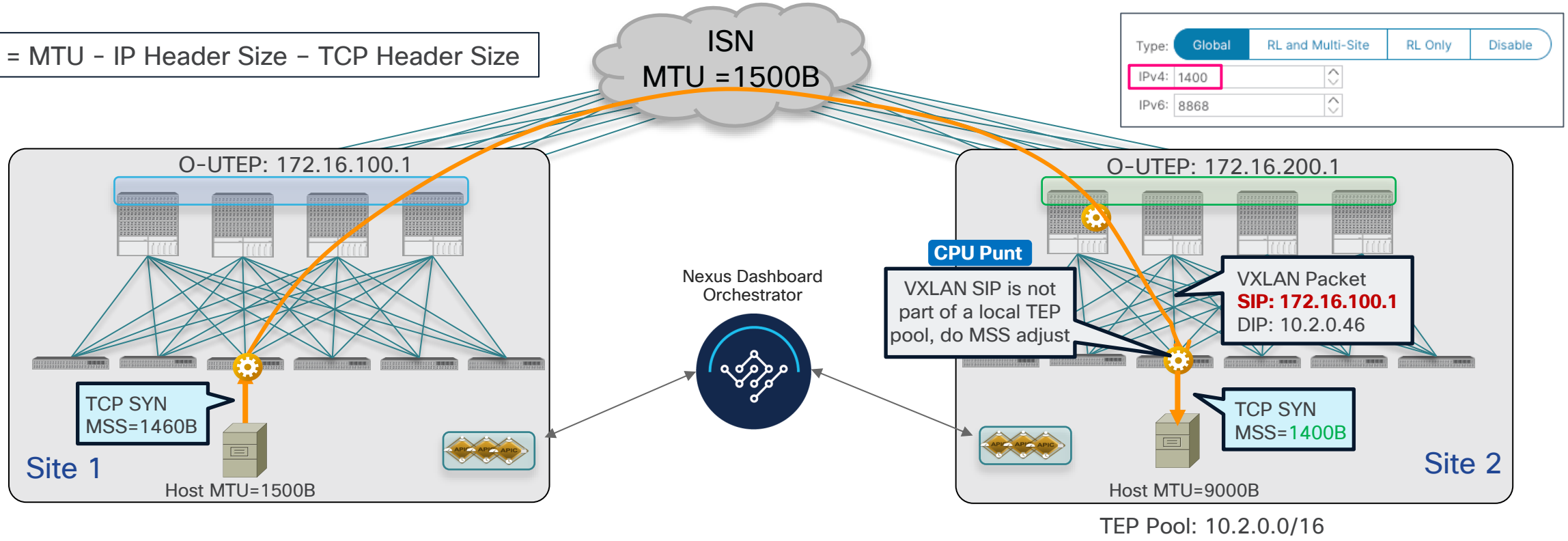
© 2025 Cisco and/or its affiliates. All rights reserved.

TCP-MSS Adjust Functionality

SYN Packet

ACI Release 6.0(3)F

MSS = MTU - IP Header Size - TCP Header Size



- TCP MSS adjust is always performed on the egress leaf node
- Adjusts TCP MSS value on SYN and SYN/ACK packets
- Checks for Source IP in the VXLAN header → TCP-MSS adjusts performed if the source IP is not part of the fabric's internal TEP pool

© 2025 Cisco and/or its affiliates. All rights reserved.

cisco

TCP-MSS Adjust Functionality

ACI Release 6.0(3)F

SYN/ACK Packet

- TCP MSS adjust is always performed on the egress leaf node
- Adjusts TCP MSS value on SYN and SYN/ACK packets
- Checks for Source IP in the VXLAN header → TCP-MSS adjusts performed if the source IP is not part of the fabric's internal TEP pool

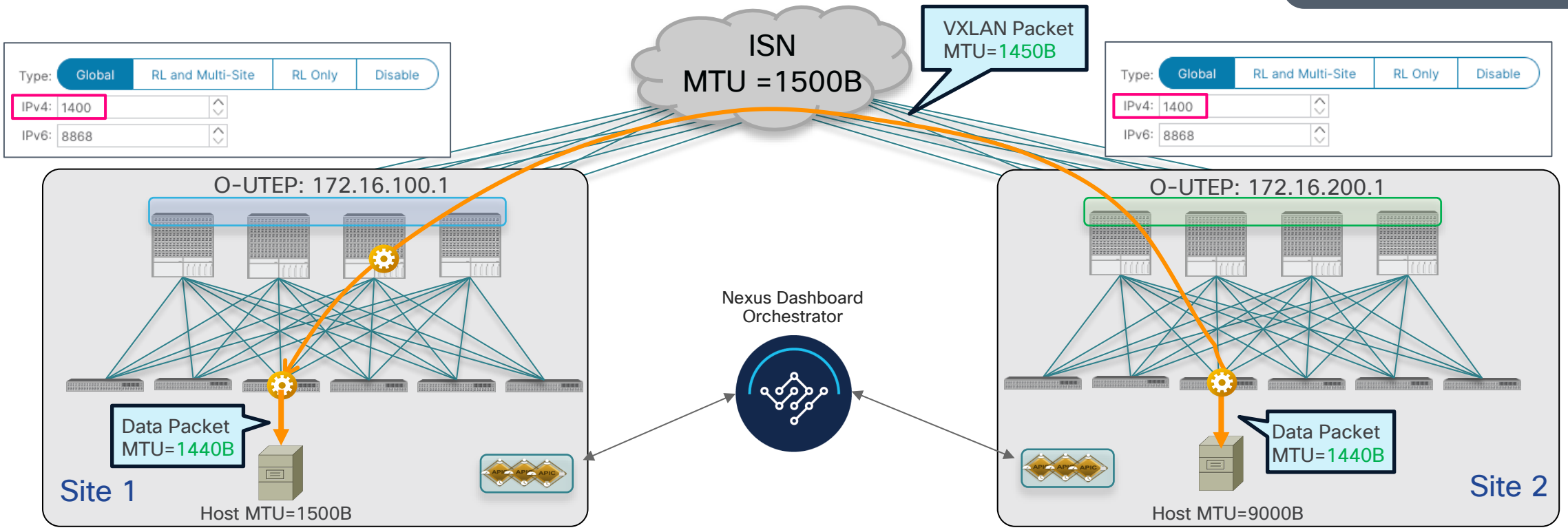
© 2025 Cisco and/or its affiliates. All rights reserved.

cisco

TCP-MSS Adjust Functionality

ACI Release 6.0(3)F

Inter-Site Data Packets



- As a result of the MSS negotiation, the endpoints generate packets for that TCP communication with total MTU 1440B (irrespectively of the local Host MTU)
- The VXLAN encapsulated traffic can be successfully forwarded across the ISN

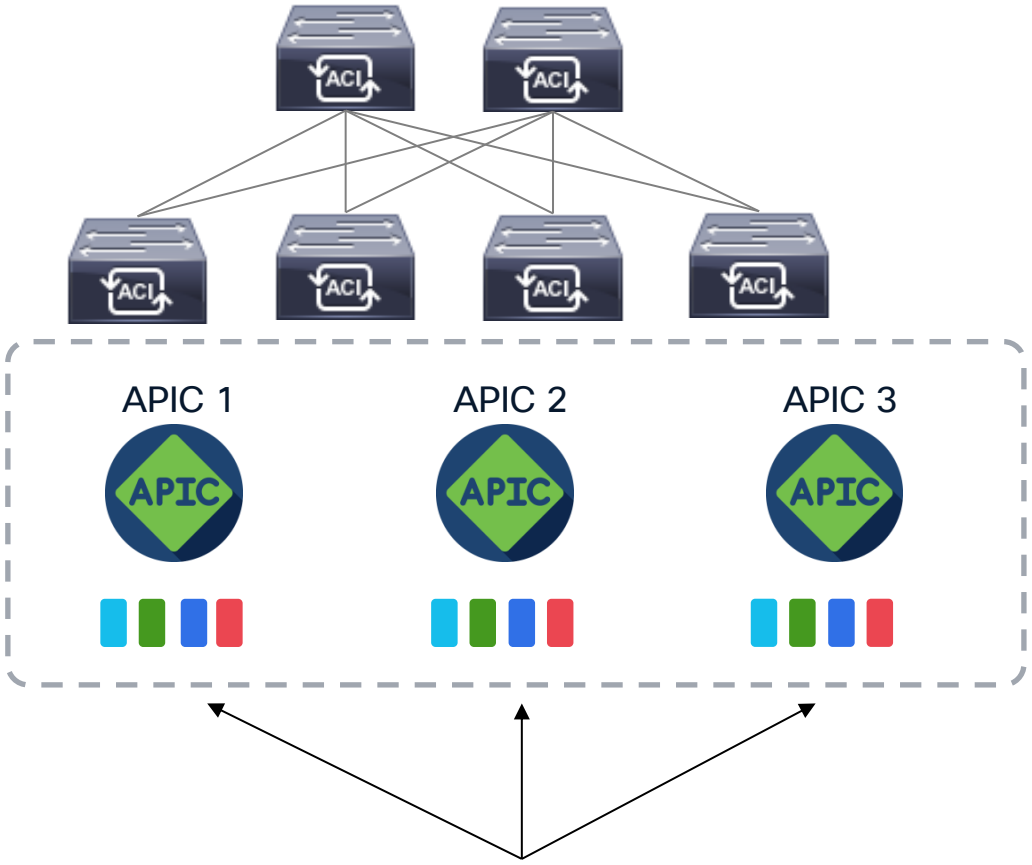
© 2025 Cisco and/or its affiliates. All rights reserved.

cisco

Multi-Pod APIC Cluster Deployment Considerations

© 2025 Cisco and/or its affiliates. All rights reserved.

APIC – Distributed Multi-Active Data Base



Sharding:
The data of the ACI fabric is divided into smaller, separate parts called shards and distributed across active APIC nodes.

Data shards

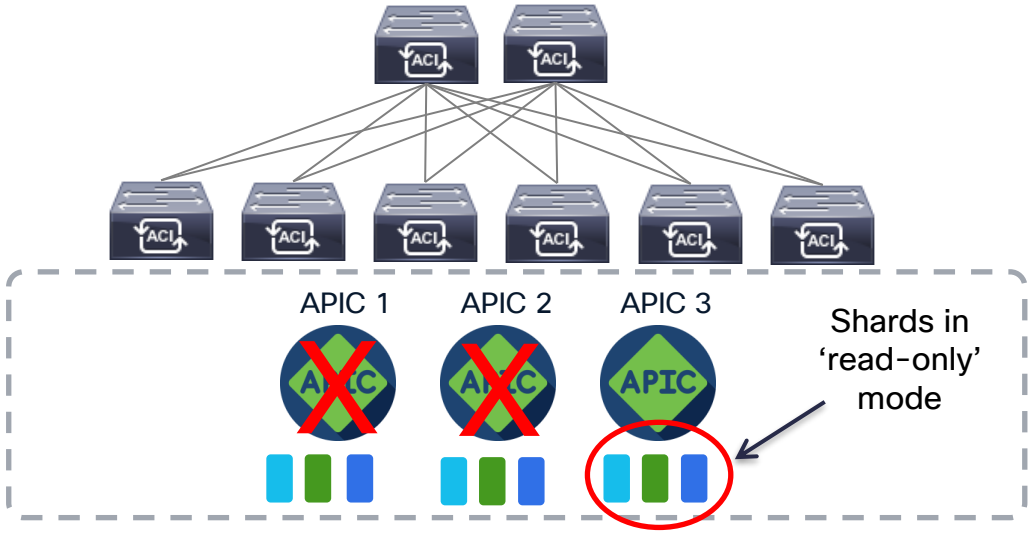
Replica:
Each shard is replicated up to three replicas regardless of the size of your cluster.

Each shard is replicated across three APICs

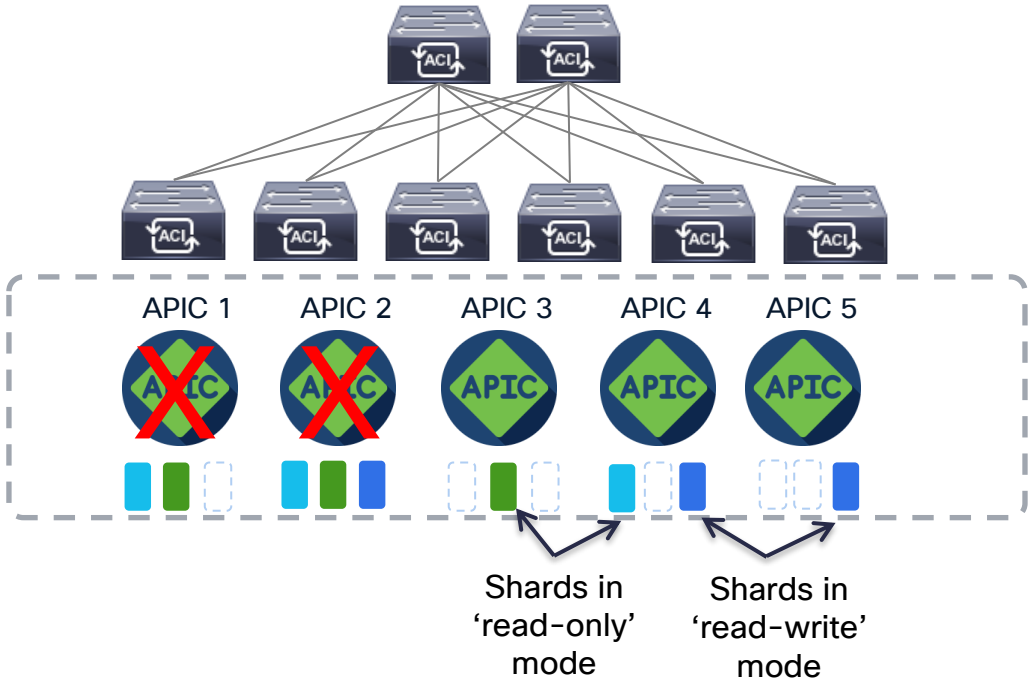
© 2025 Cisco and/or its affiliates. All rights reserved.

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)

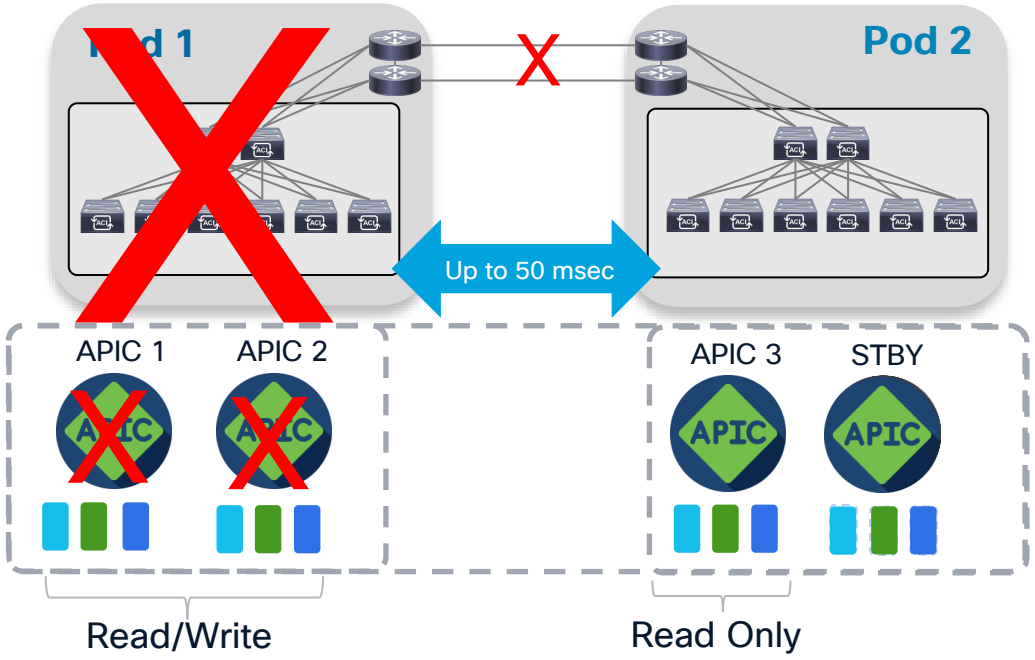


- Additional APIC will increase the system scale (up to 7* nodes supported) but does not add more redundancy
- Hard failure of two nodes would cause inconsistent behavior across shards (some will be in 'read-only' mode, some in 'read-write' mode)

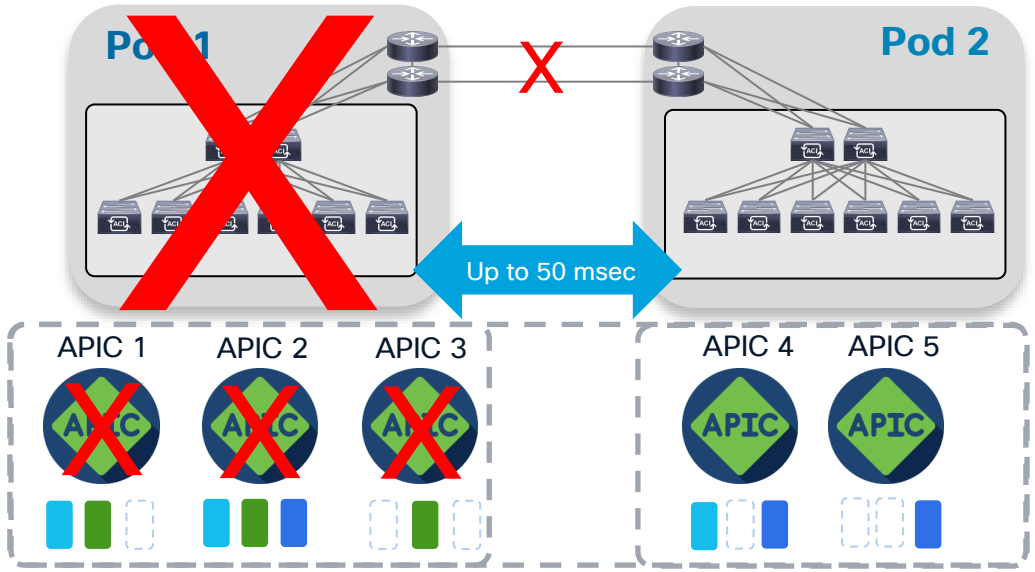
© 2025 Cisco and/or its affiliates. All rights reserved.

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

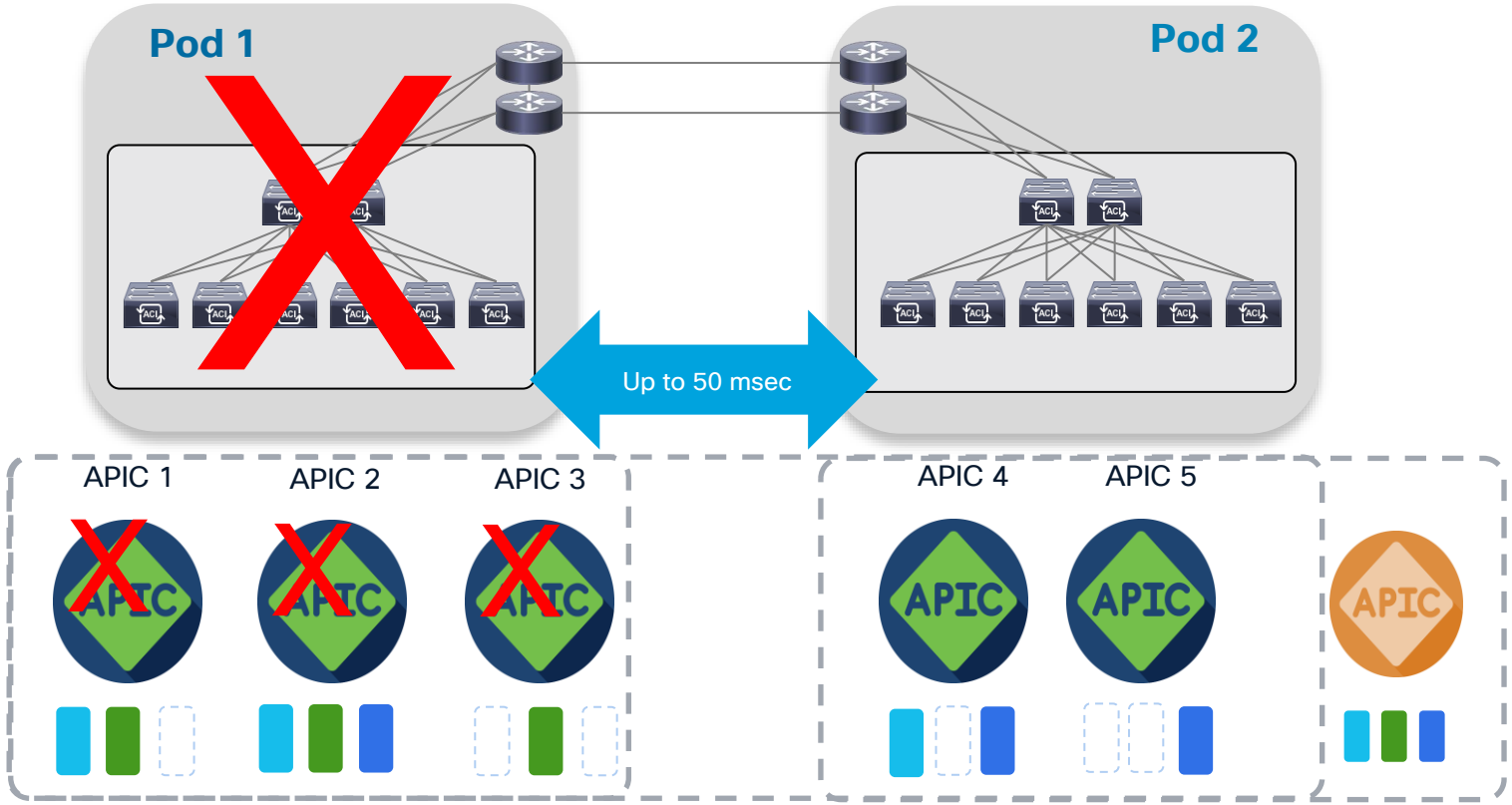


- Pod isolation scenario: same considerations as with single Pod (different behavior across shards)
- Pod hard failure scenario: may cause the loss of information for the shards replicated across APIC nodes in the failed Pod

  Possible to restore the whole fabric state to the latest taken configuration snapshot ('ID Recovery' procedure – needs BU and TAC involvement or Warm Standby from 6.1(2))

© 2025 Cisco and/or its affiliates. All rights reserved.

Warm Standby APIC Helps (available since ACI 6.1(2))



Warm Standby
(full data backup)

Rebuild the entire APIC cluster with the standby as new APIC1 via the data stored in the standby.

© 2025 Cisco and/or its affiliates. All rights reserved.

cisco

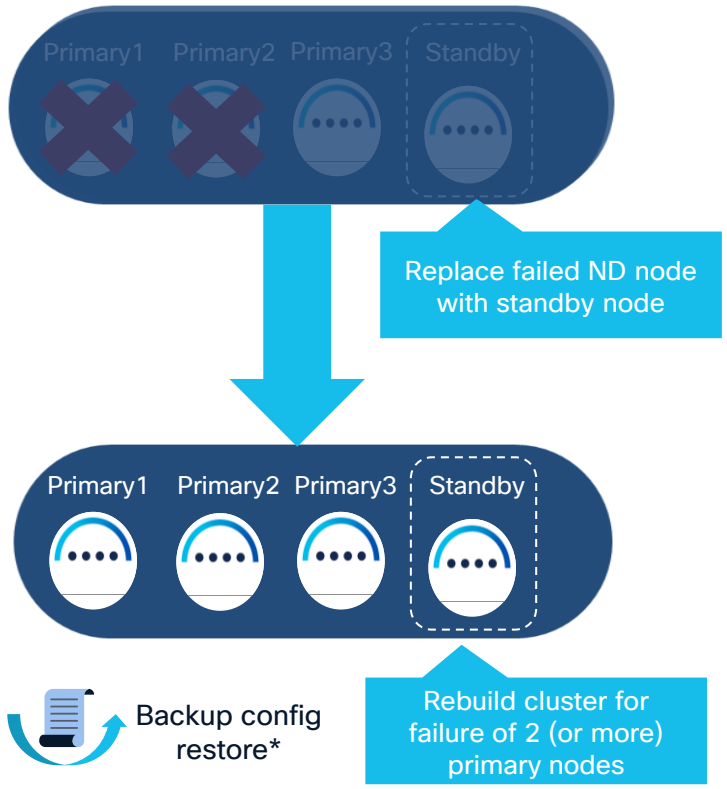
Nexus Dashboard Orchestrator (NDO) Deployment Considerations

© 2025 Cisco and/or its affiliates. All rights reserved.

Nexus Dashboard High Availability

Nexus Dashboard 3.2

- At least 2 ND active/primary nodes are needed to keep the cluster operational
- In case of a single primary node failure, the cluster will be operational but in a degraded state
- A standby node can be used to replace a failed primary node
- A standby node can be virtual or physical (must be of the same form factor as failed node, app or data for virtual)
- From ND 3.2 it is not possible to restore the failure of two primary nodes

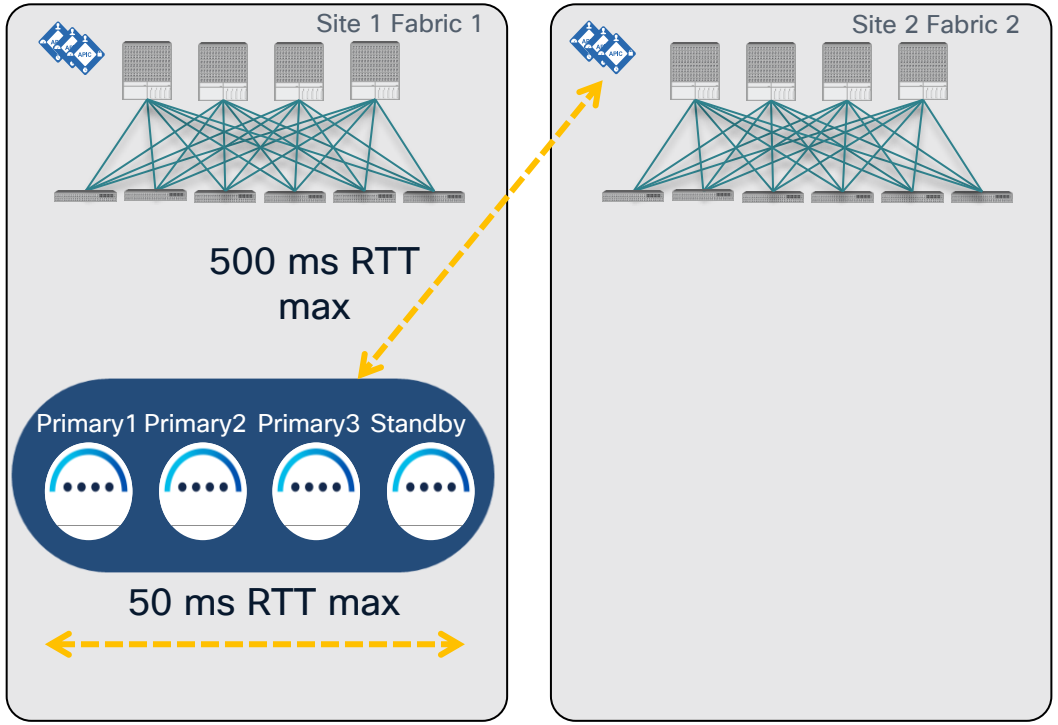


© 2025 Cisco and/or its affiliates. All rights reserved.

Nexus Dashboard Orchestrator High Availability

Centralized ND Cluster Deployment (Recommended)

- Low Latency communication facilitating quick data exchange (Services do require RTT between the nodes that ranges between 50ms and 150ms.)
- Up to 500ms RTT for managing fabric at remote site
- Simplified Network Topology
- In case of single ND node failure, restore a failed node with a standby node
- Recommended deployment mode for ND cluster. Distributed deployment is possible but adds more complexity without benefit

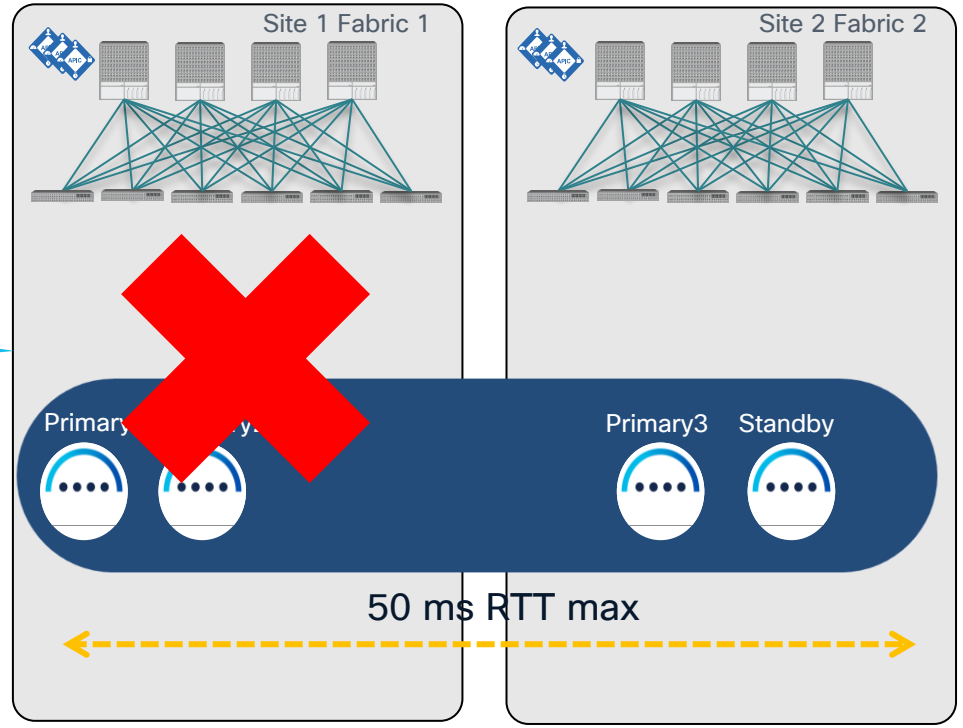


© 2025 Cisco and/or its affiliates. All rights reserved.

Nexus Dashboard Orchestrator High Availability

Distributed ND Cluster Deployment: Not recommended with two sites

- For two site deployments it is not recommended to deploy a distributed cluster
- Failure of two primary nodes in one site will make recovery of the cluster not possible
- Adds more complexity without benefit

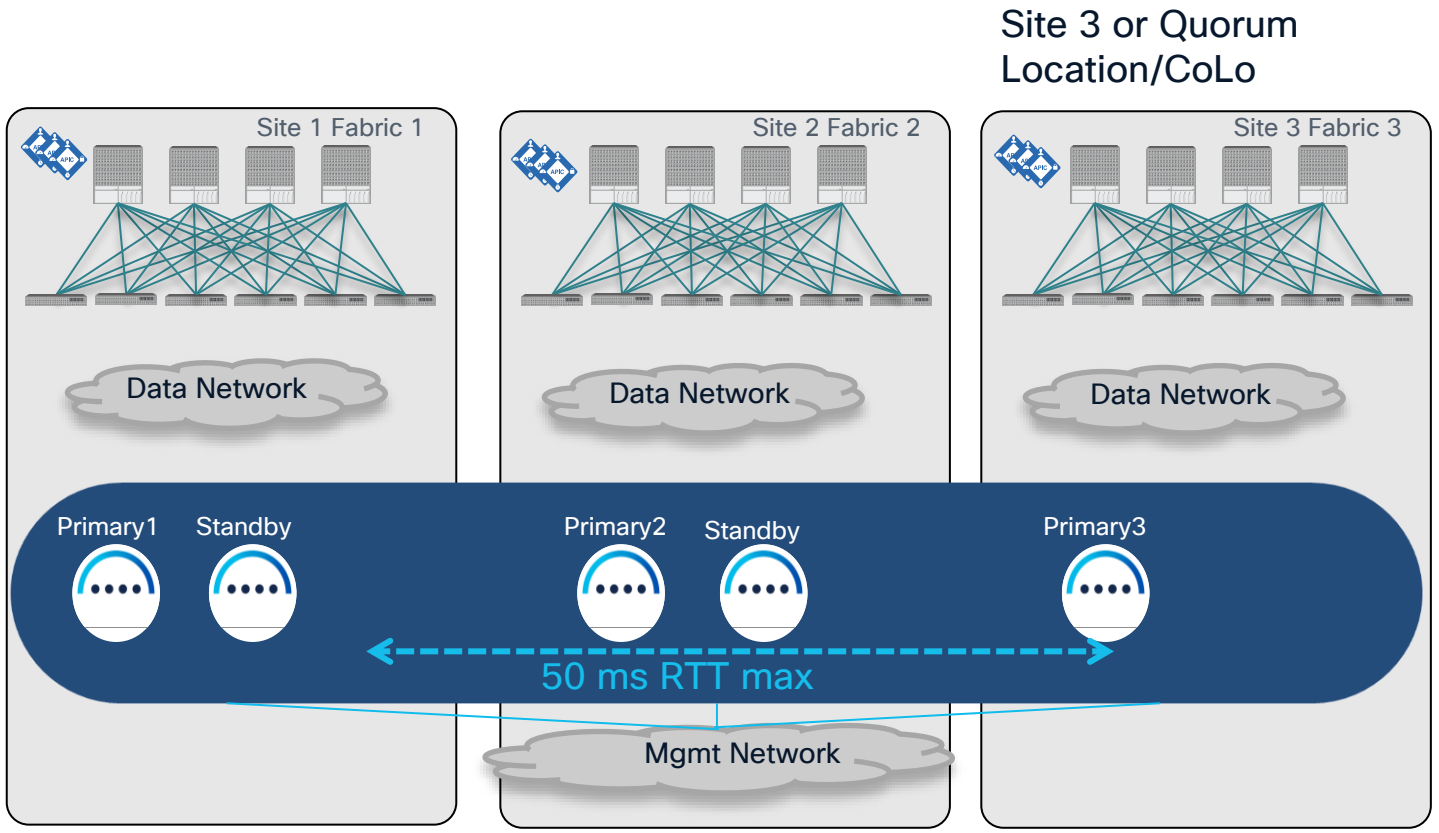


© 2025 Cisco and/or its affiliates. All rights reserved.

Nexus Dashboard Orchestrator High Availability

Distributed ND Cluster Deployment: Three site deployments

- In the case of three or more sites, the ND cluster can be distributed across three sites
- Only one primary node should be deployed at any one site
- Cluster is still recoverable with the loss of one site
- Distributed cluster must be within the 50ms RTT latency requirement

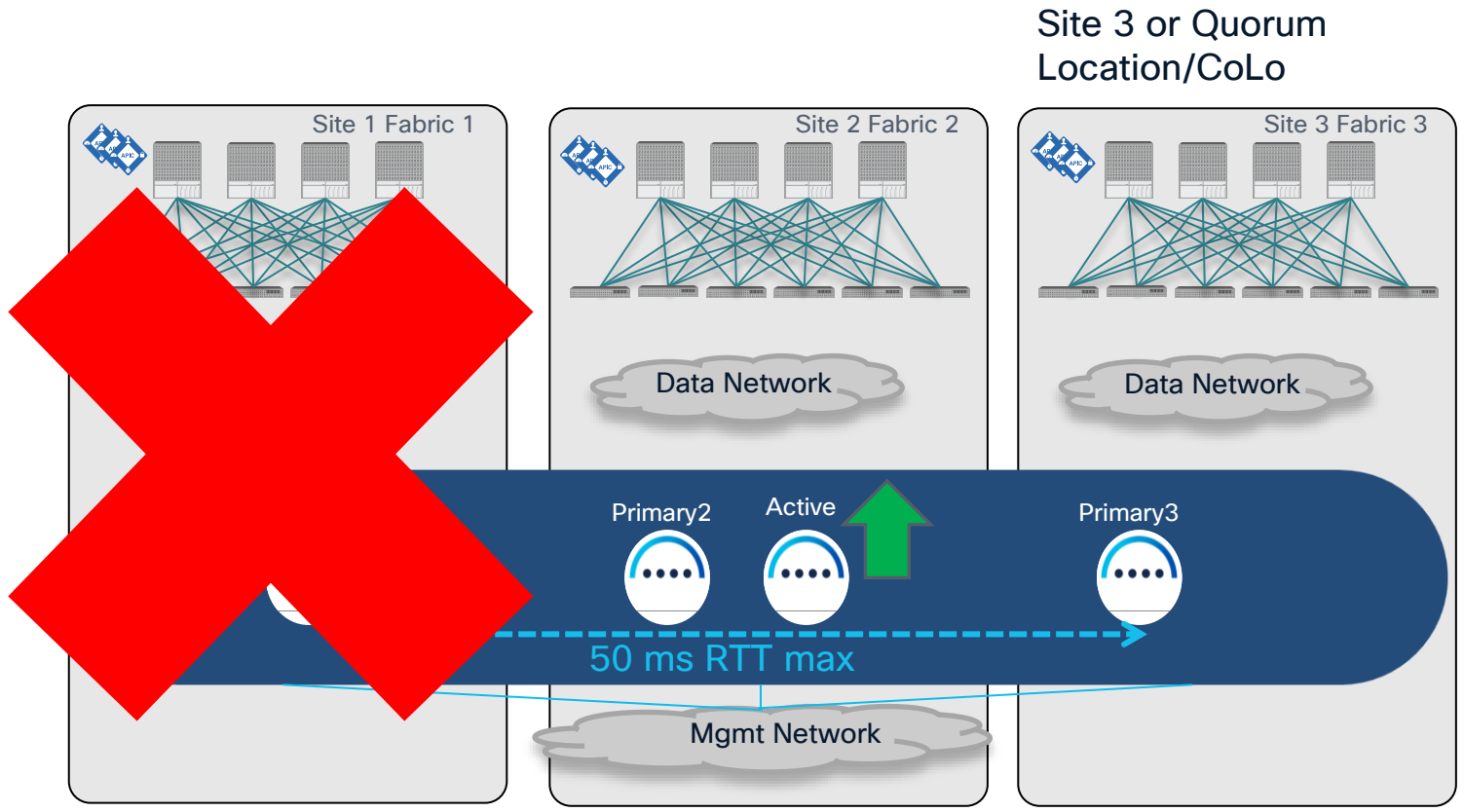


© 2025 Cisco and/or its affiliates. All rights reserved.

Nexus Dashboard Orchestrator High Availability

Distributed ND Cluster Deployment: Three site deployments

- In the case of three or more sites, the ND cluster can be distributed across three sites
- Only one primary node should be deployed at any one site
- Cluster is still recoverable with the loss of one site
- Distributed cluster must be within the 50ms RTT latency requirement



No need to perform any backup config restore procedure

© 2025 Cisco and/or its affiliates. All rights reserved.

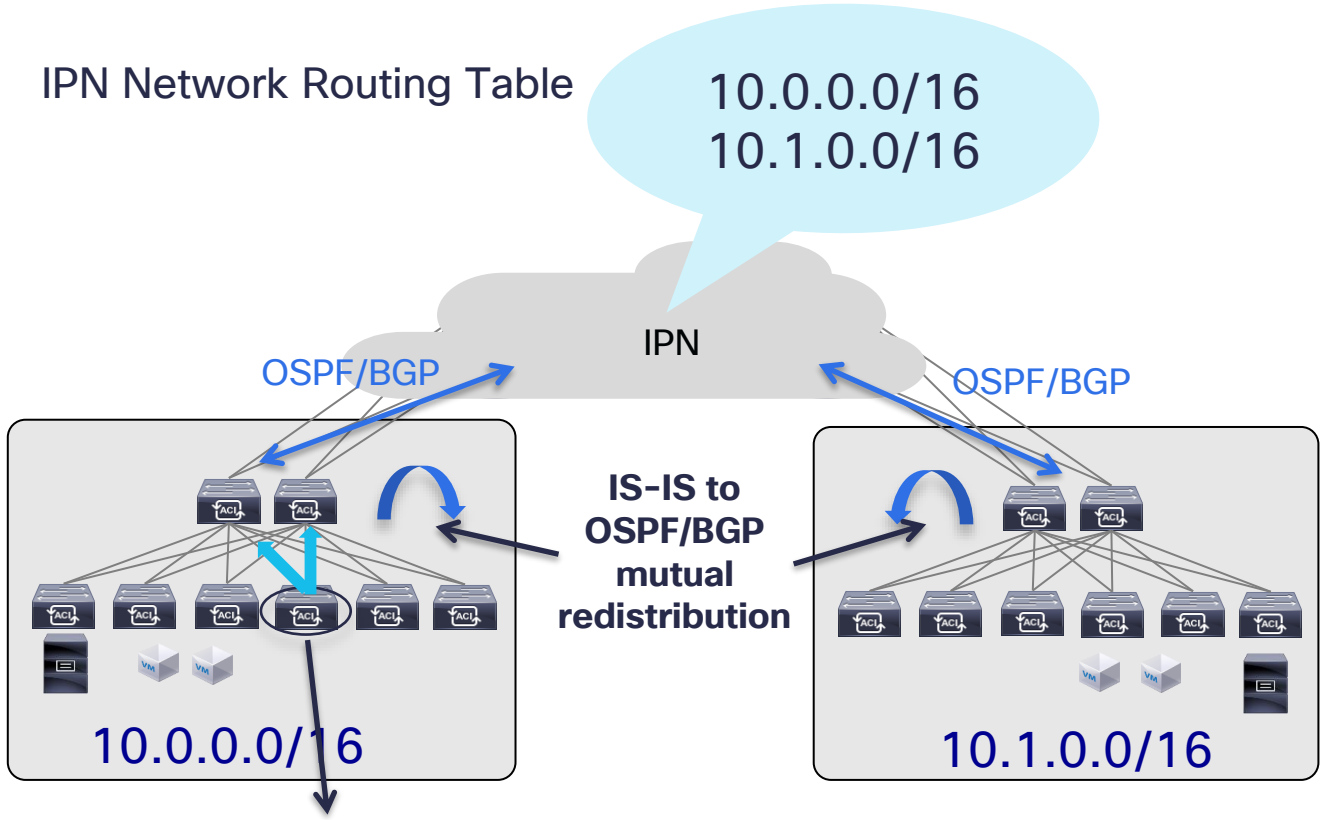
cisco

Multi-Pod Control and Data-Plane

© 2025 Cisco and/or its affiliates. All rights reserved.

Leafs/TEP reachability

- 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 routing table

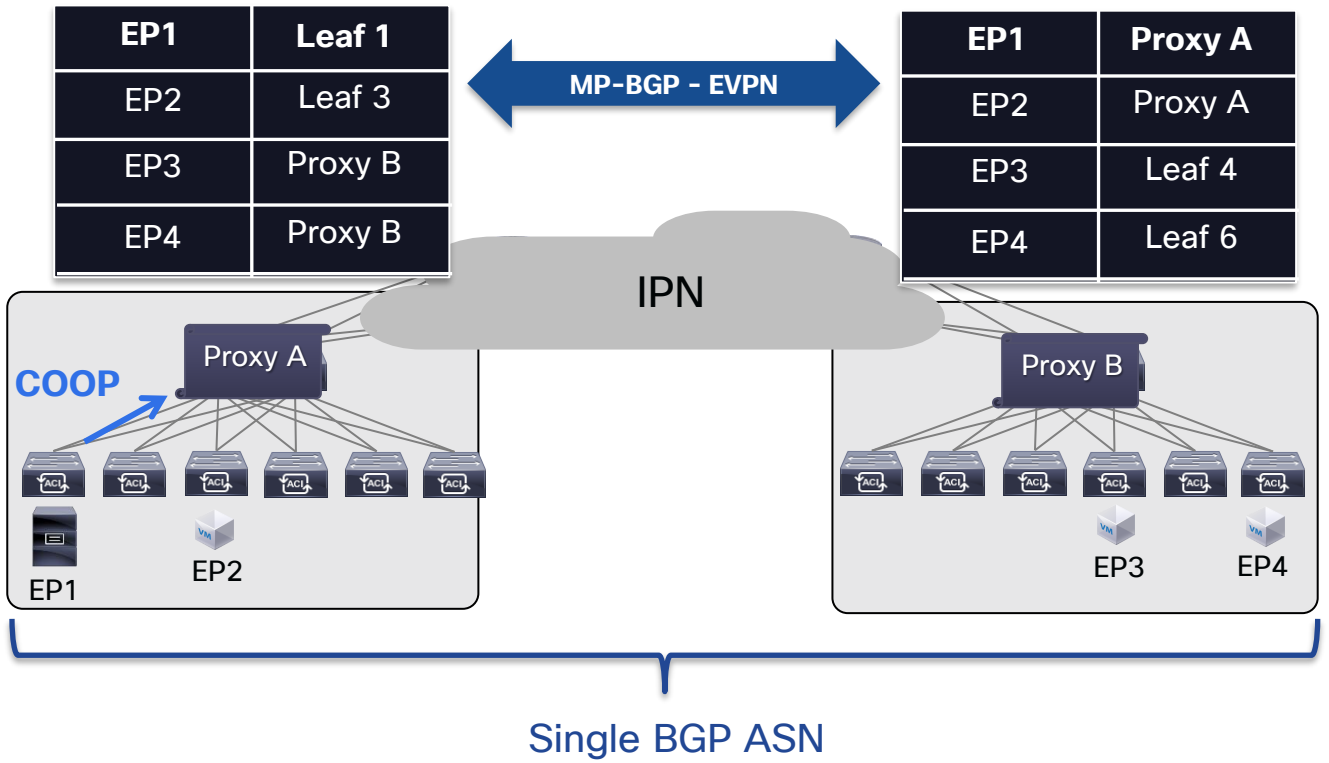
IP Prefix	Next-Hop
10.1.0.0/16	Pod1-S1, Pod1-S2

© 2025 Cisco and/or its affiliates. All rights reserved.

Endpoint Reachability Information

Inter-Pod MP-BGP EVPN Control Plane

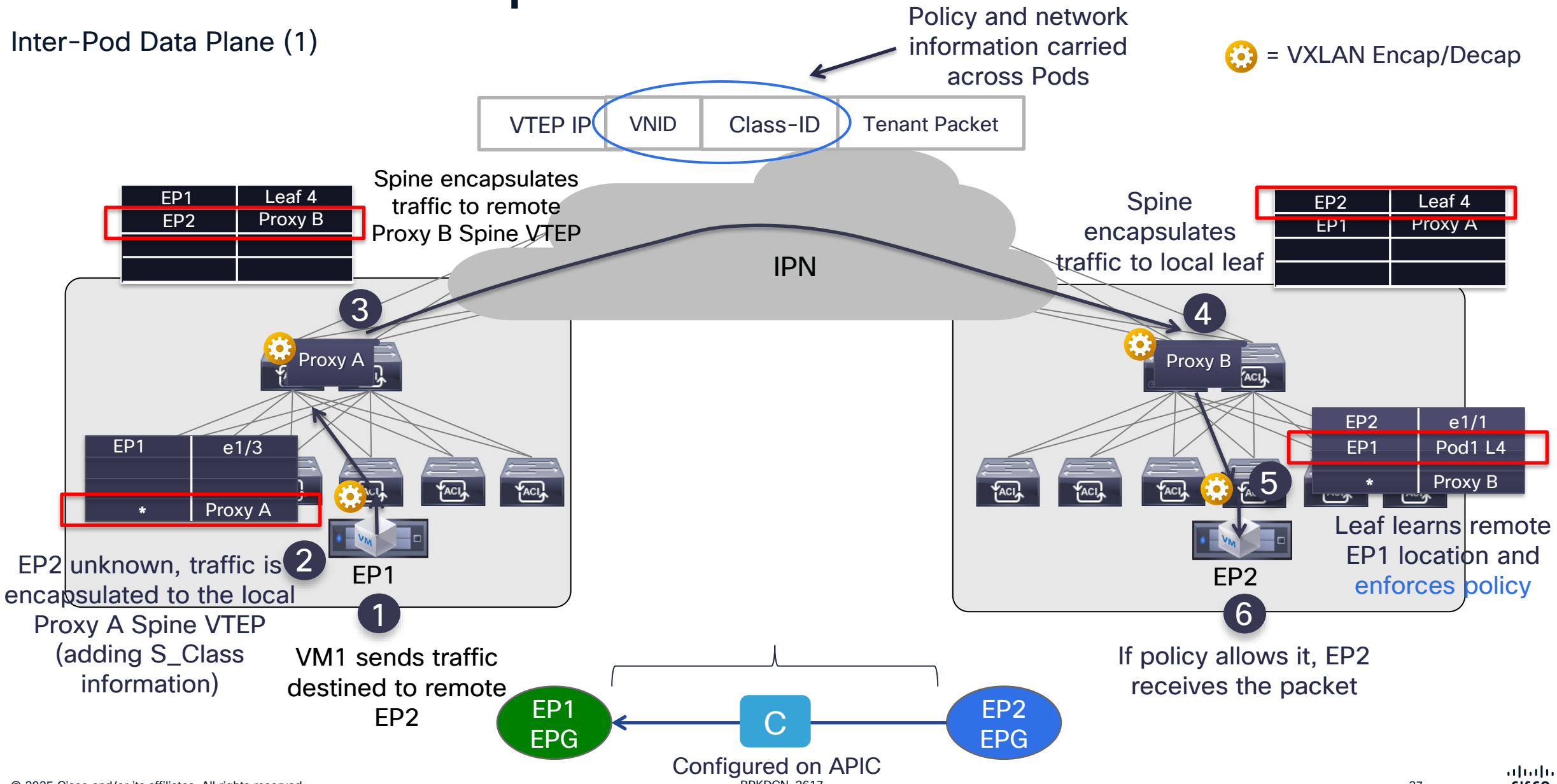
- MP-BGP EVPN to sync Endpoint (EP) and Multicast Group information
  - All remote Pod entries associated to a Proxy VTEP next-hop address (not part of local TEP Pool)
  - Same BGP AS across all the Pods
- iBGP EVPN sessions between spines in separate Pods
  - Full mesh MP-iBGP EVPN sessions between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



© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Pod Dataplane Packet Walk

Inter-Pod Data Plane (1)

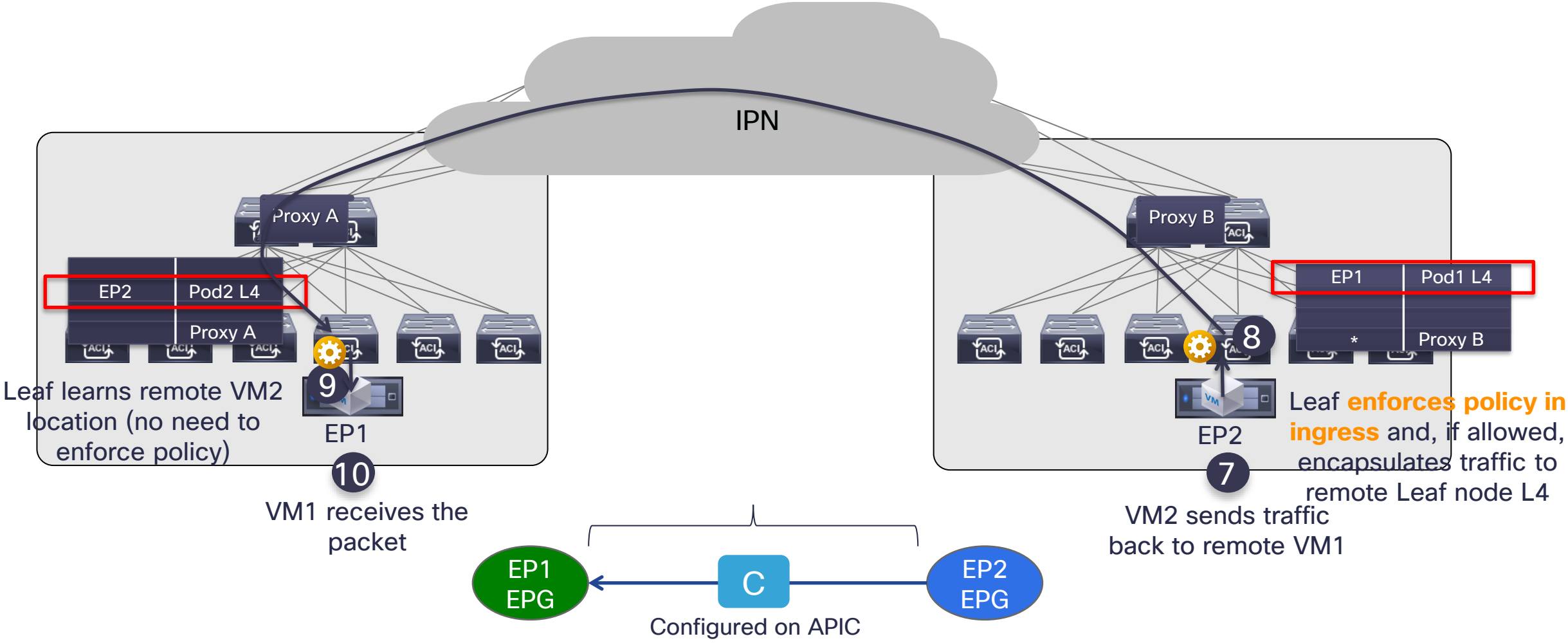


© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Pod Dataplane Packet Walk

Inter-Pod Data Plane (2)

= VXLAN Encap/Decap

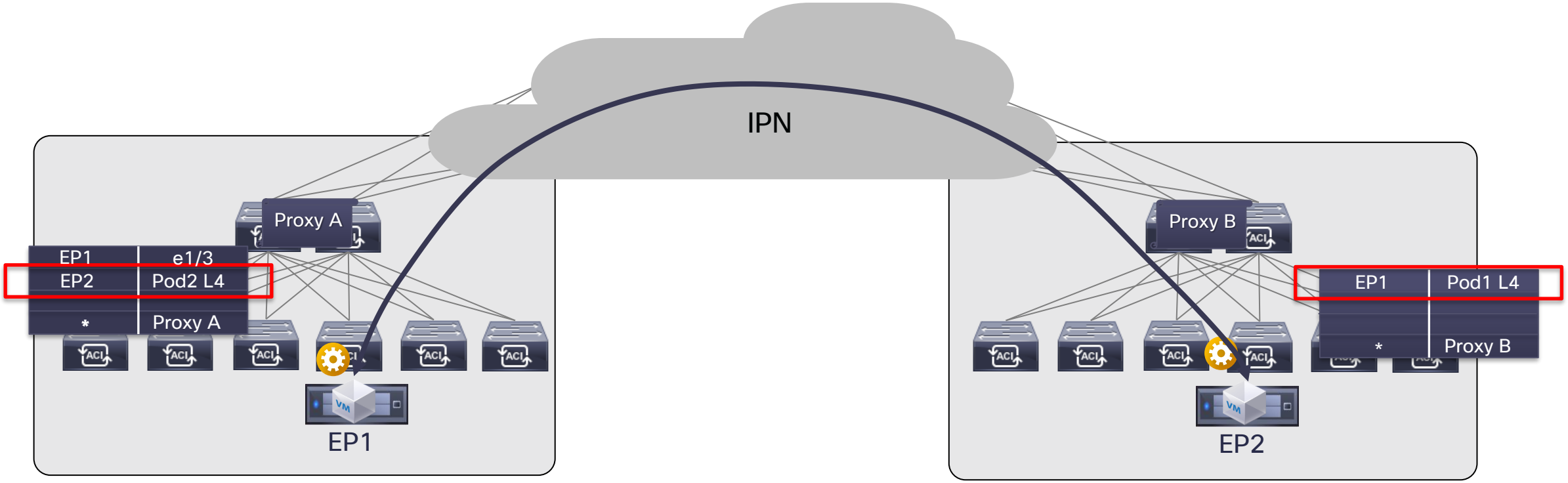


© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Pod Dataplane Packet Walk

Inter-Pod Data Plane (3)

= VXLAN Encap/Decap



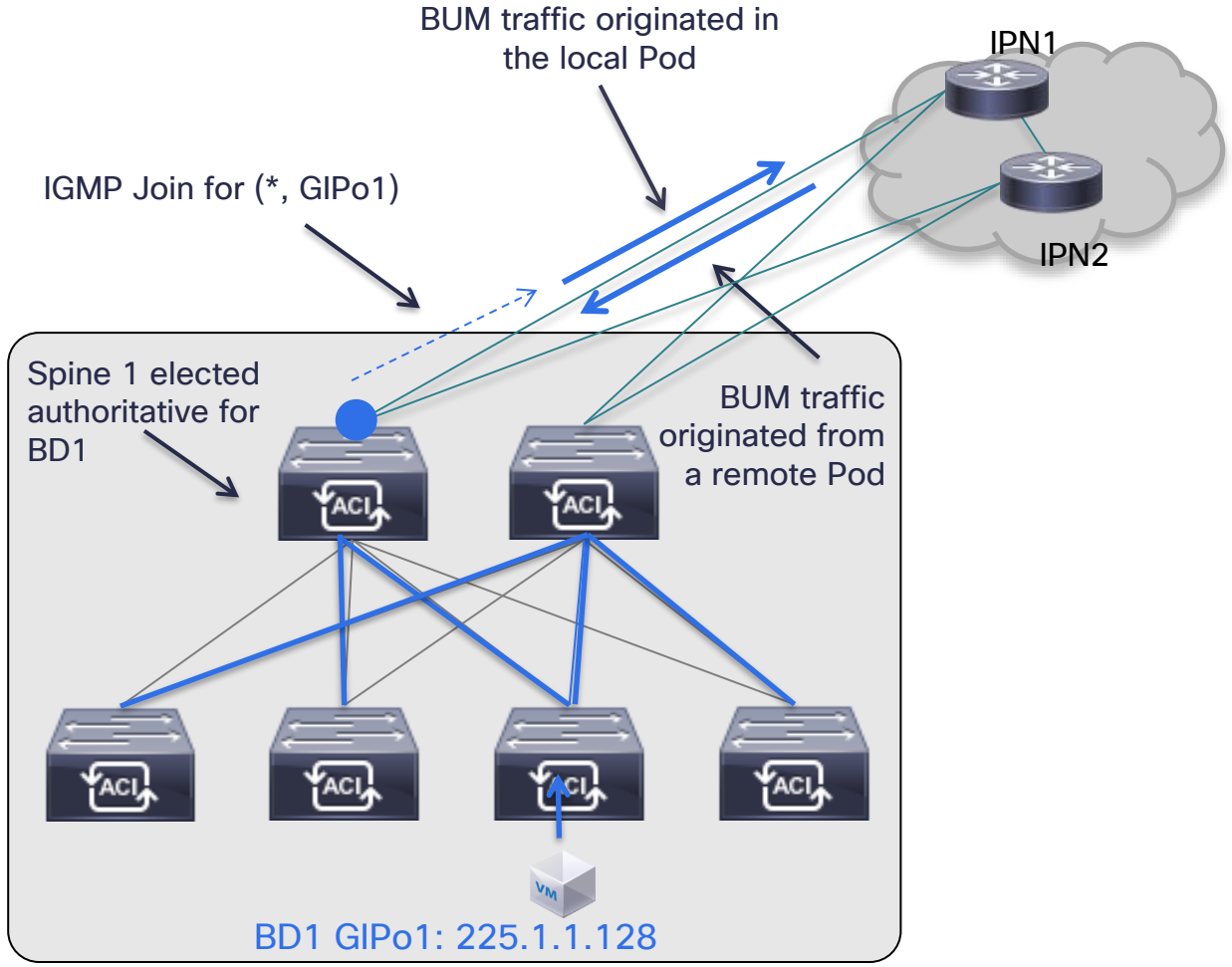
From this point EP1 to EP2 communication is encapsulated Leaf to Leaf (VTEP to VTEP) and **policy always applied at the ingress leaf** (applies to both L2 and L3 communication)

11

© 2025 Cisco and/or its affiliates. All rights reserved.

But how about ARP and other BUM (*) traffic?

Use of Multicast for Inter-Pod Layer 2 BUM Traffic



- BUM traffic is forwarded via Multicast VXLAN Packet called GIPO (Group IP Outer Address)
- Each Bridge Domain is associated to a Multicast Group / GIPO
- PIM Bidir is required in the IPN
- 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

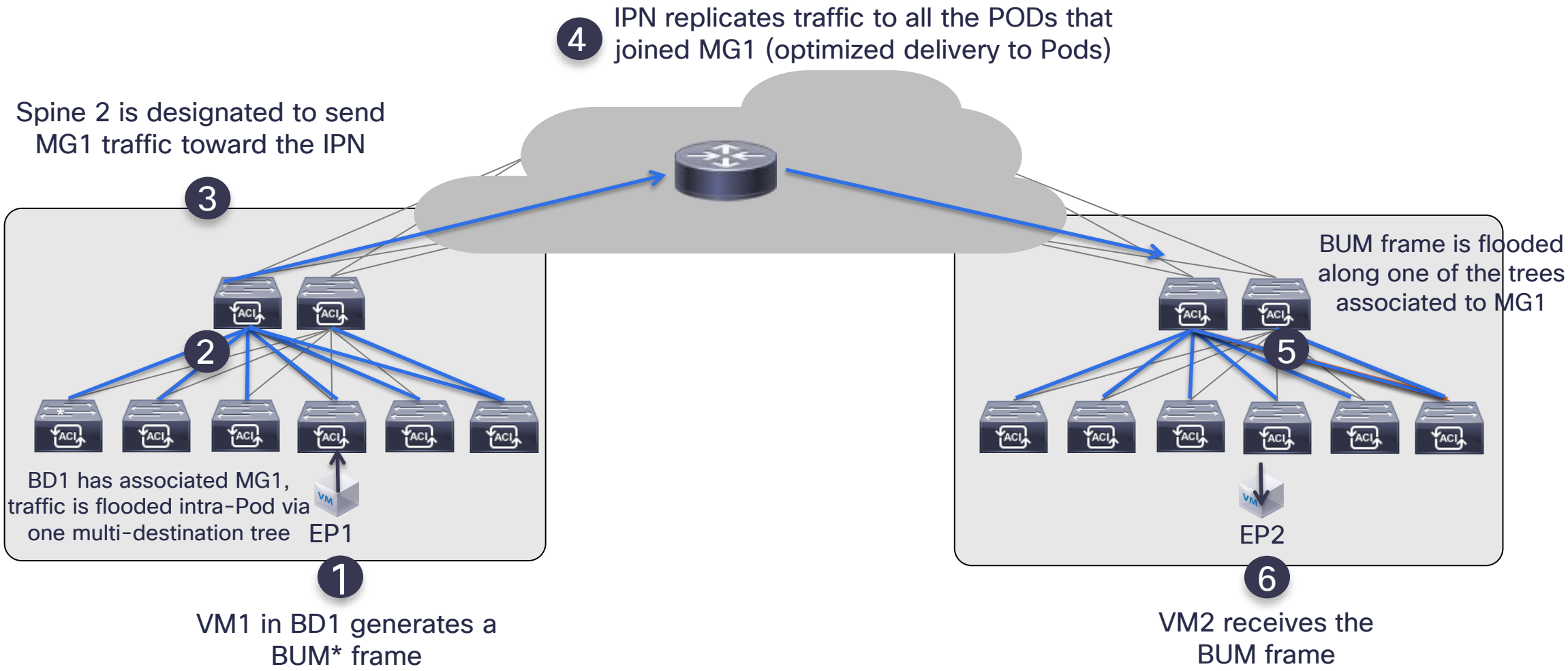
(*) BUM: Broadcast, Unknown Unicast, Multicast

© 2025 Cisco and/or its affiliates. All rights reserved.

cisco

Example with some BUM Traffic

Use of Multicast for Inter-Pod BUM Traffic



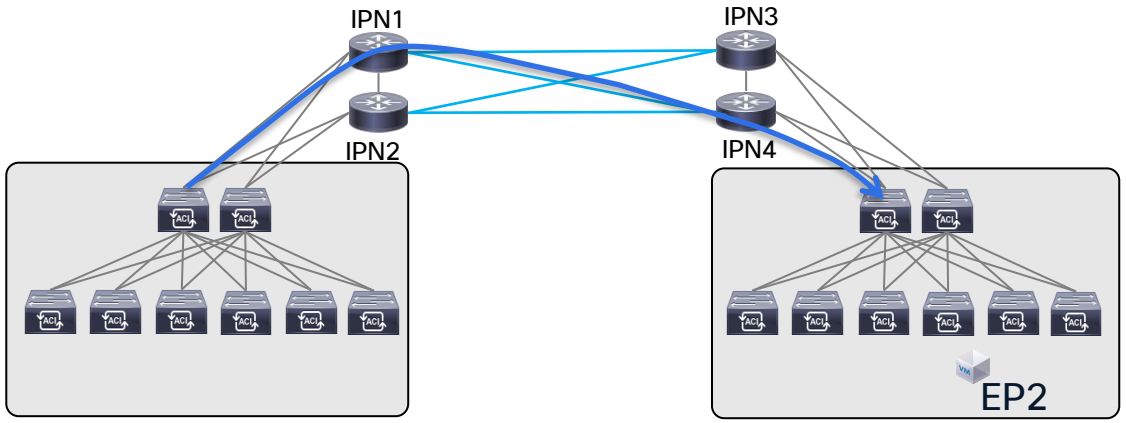
(*) BUM: Layer 2 Broadcast, Unknown Unicast, Multicast

© 2025 Cisco and/or its affiliates. All rights reserved.

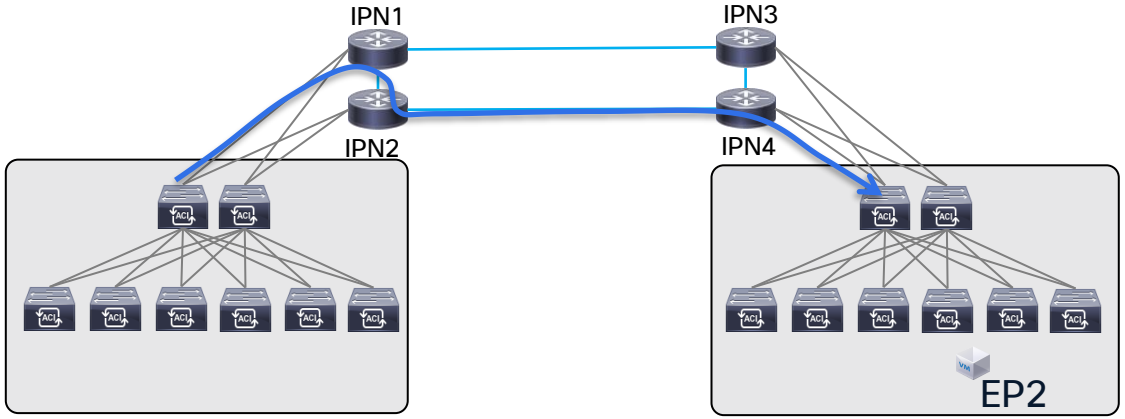
IPN Best Practices

PIM Bidir for BUM – Supported Topologies

Full Mesh between remote IPN devices



Directly connect local IPN devices



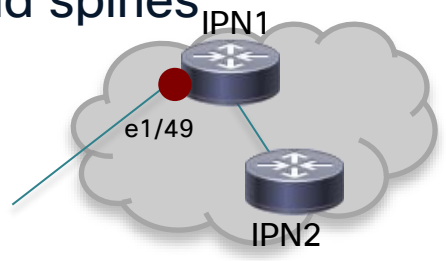
- **Only one link** from the Spine is a Designated Forwarder link per POD
- Because of this:

  Ideally create full-mesh connections between IPN devices

  Alternatively, connect local IPN devices with a port-channel interface (for resiliency)

- 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 [illegible] 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
```



© 2025 Cisco and/or its affiliates. All rights reserved.

Multi-Site Control and Dataplane

© 2025 Cisco and/or its affiliates. All rights reserved.

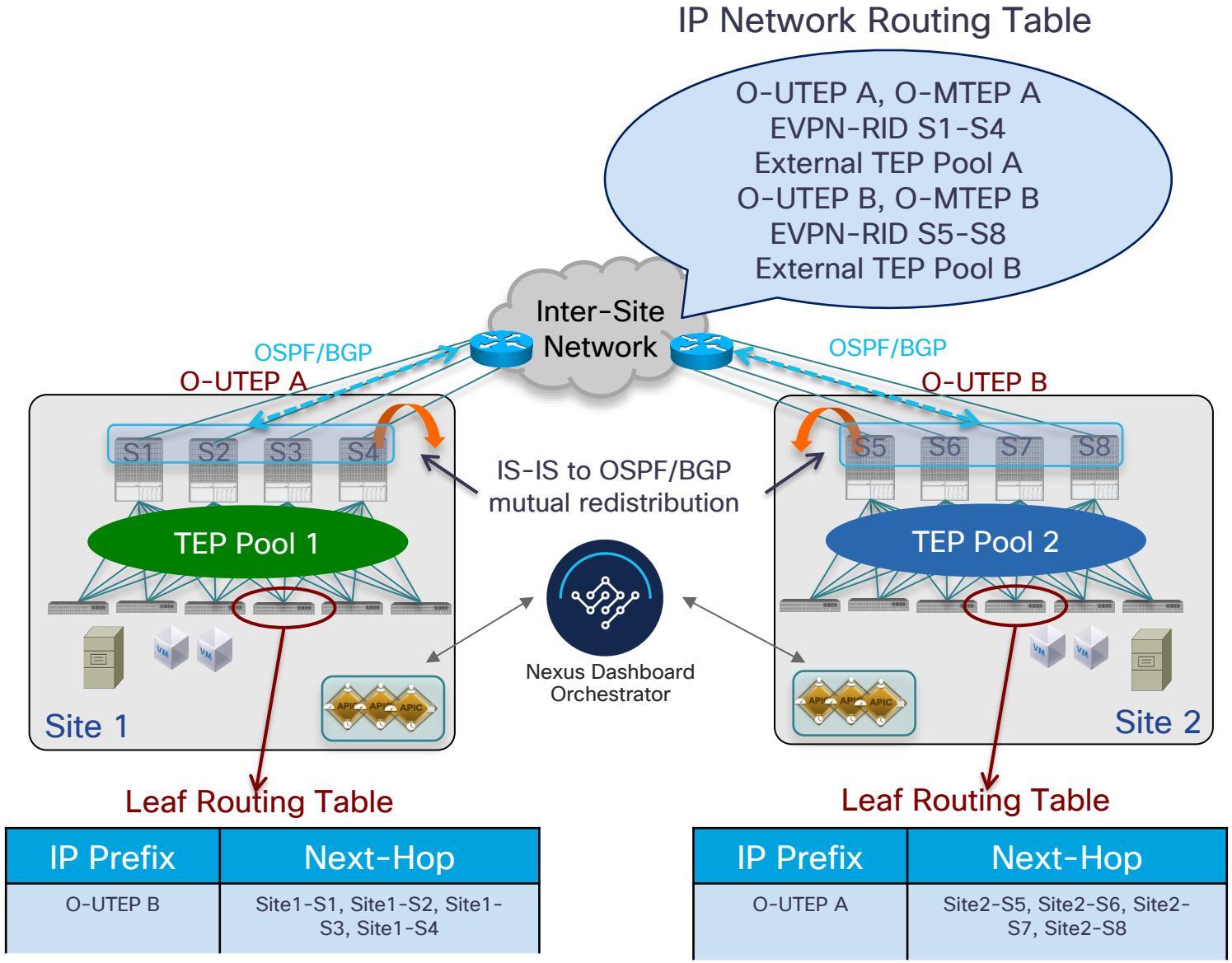
Leafs/TEPs reachability in Multi-Site is different than Multi-Pod

TEP pools may overlap between fabrics

Forwarding of inter-sites data-plane traffic leverages these addresses instead:

- Anycast Overlay Unicast TEP (O-UTEP): assigned to all the spines connected to the ISN and used to source and receive L2/L3 unicast traffic
- Anycast Overlay Multicast TEP (O-MTEP): assigned to all the spines connected to the ISN and used to receive L2 BUM traffic
- External TEP Pool subnet: TEPs from external TEP pool assigned to border leaf switches for intersite L3Out

- OSPF or BGP peering between spines and Inter-Site network
  - Mandates the use of L3 sub-interfaces (with VLAN 4 tag) between the spines and the ISN
- Exchange of External Spine TEP addresses (EVPN-RID, O-UTEP, O-MTEP, and External TEP Pool) across sites



IP Prefix	Next-Hop
O-UTEP B	Site1-S1, Site1-S2, Site1-S3, Site1-S4

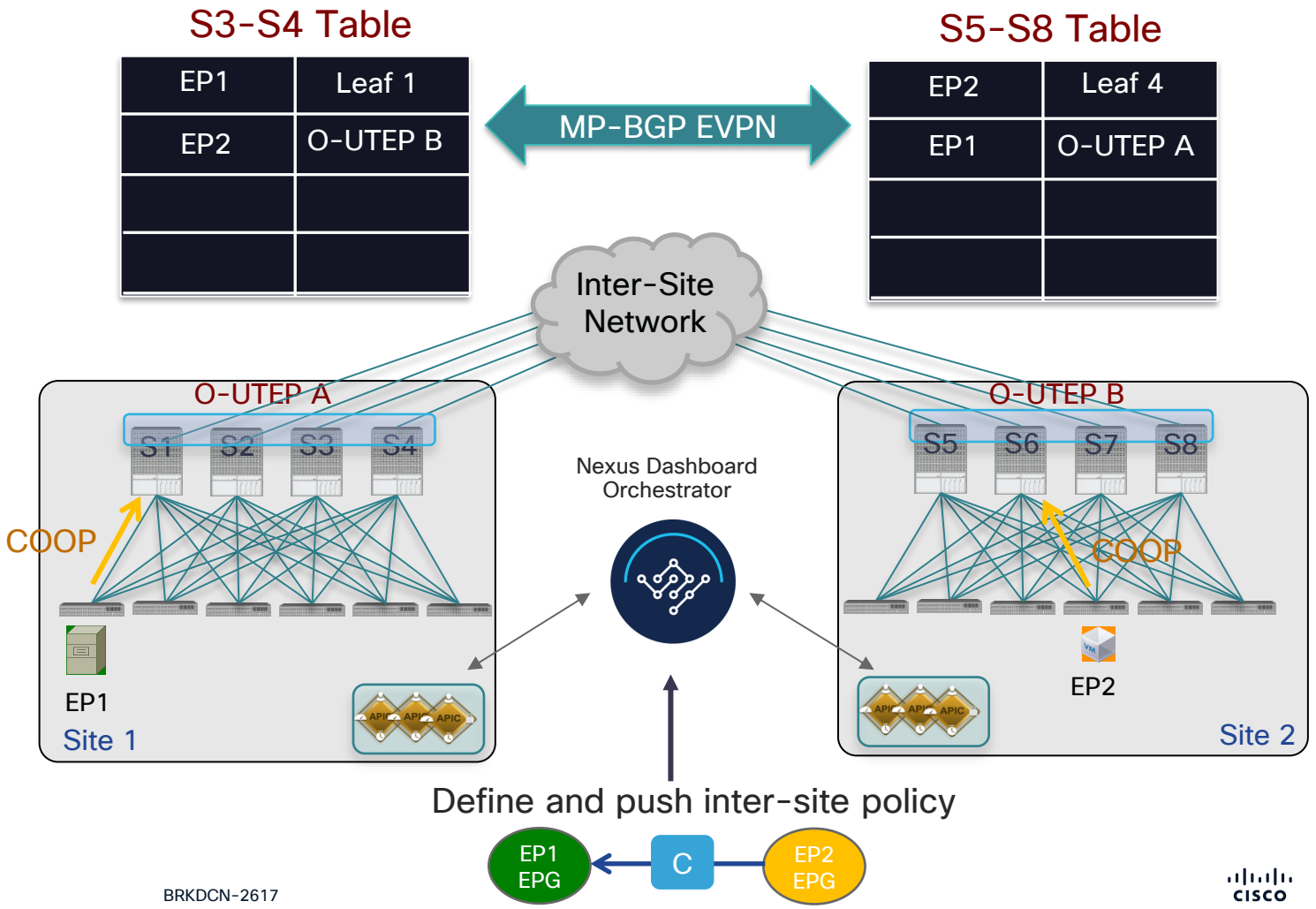
IP Prefix	Next-Hop
O-UTEP A	Site2-S5, Site2-S6, Site2-S7, Site2-S8

© 2025 Cisco and/or its affiliates. All rights reserved.

How is endpoint reachability information propagated?

Inter-Site MP-BGP EVPN Control Plane

- MP-BGP EVPN used to communicate Endpoint (EP) information across Sites
  - MP-iBGP or MP-EBGP peering options supported
  - Required MP-BGP configuration fully automated via NDO
  - Remote host route entries (EVPN Type-2) are associated to the remote site Anycast O-UTEP address
- There is a big difference with Multi-Pod: Host routes are exchanged across sites **only if there is a cross-site contract requiring communication between endpoints**



© 2025 Cisco and/or its affiliates. All rights reserved.

Multi-Site Dataplane => "Translation" of pcTags and VNIDs

© 2025 Cisco and/or its affiliates. All rights reserved.

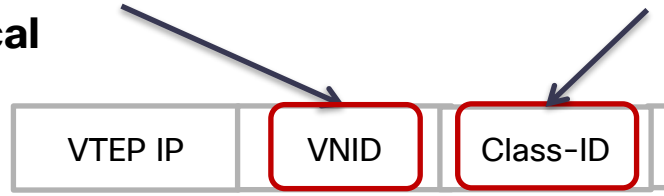
ACI Multi-Site Dataplane Forwarding

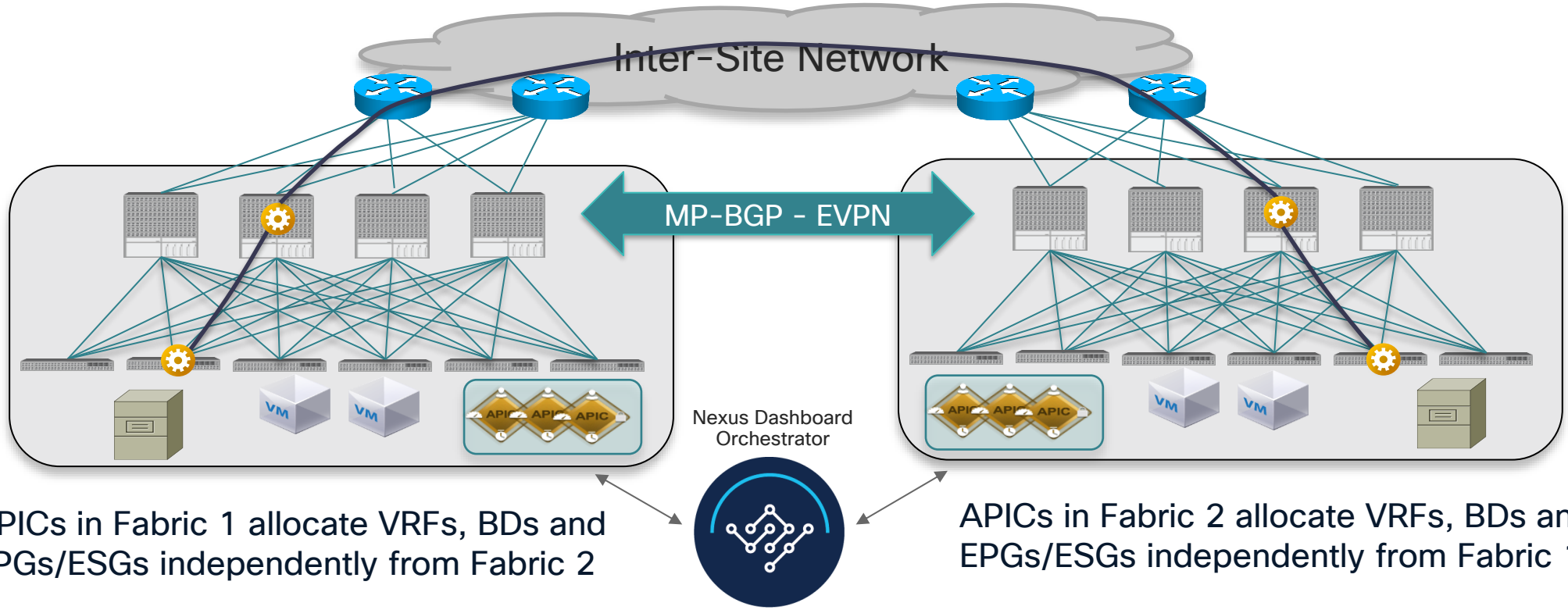
VXLAN packet format to Enable L2 extension and Polcy End-to-End

Deployment Mode
- (•) Multi-Site
- ( ) Autonomous

Identifier for VRFs / BDs
=> gets translated to the local namespace

Identifier for EPGs/ESGs
=> gets translated to the local namespace





APICs in Fabric 1 allocate VRFs, BDs and EPGs/ESGs independently from Fabric 2

APICs in Fabric 2 allocate VRFs, BDs and EPGs/ESGs independently from Fabric 1

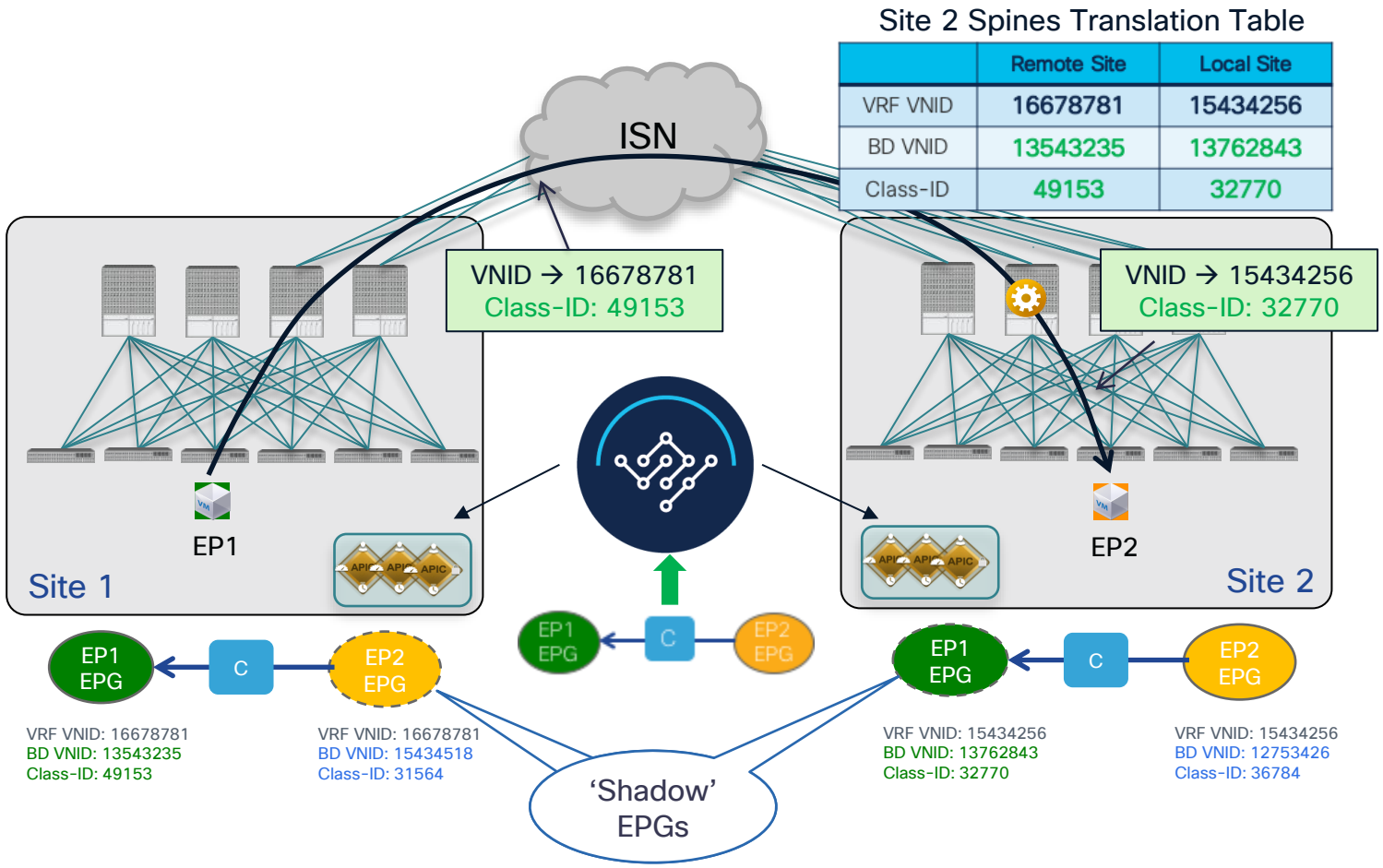
© 2025 Cisco and/or its affiliates. All rights reserved.

cisco

How can a Contract between EPGs of different Site work?

Inter-Site Policies and Spines' Translation Tables

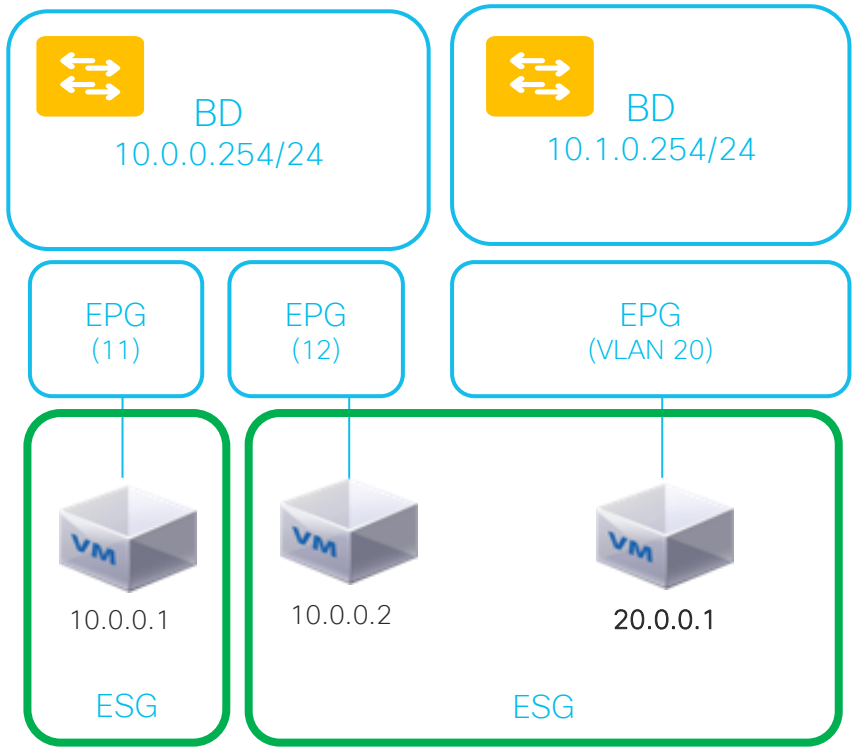
- Inter-Site policies defined on the ACI Nexus Dashboard Orchestrator are pushed to the respective APIC domains
  - End-to-end policy consistency
  - Creation of 'Shadow' objects to locally recreate the policies in each APIC domain
- Inter-site communication requires the installation of translation table entries on the spines (namespace normalization)
- Translation entries are created in different cases:
  - Stretched EPGs/BDs
  - Creation of a contract between site-local (not stretched) EPGs
  - Preferred Group or vzAny deployments



© 2025 Cisco and/or its affiliates. All rights reserved.

Short Digression on ESGs

Do you know what are ESGs?



ND 4.1 adds ESGs support

© 2025 Cisco and/or its affiliates. All rights reserved.

ESGs are a way to classify endpoints

These are the Selectors supported in ND 4.1

Supported Selectors

IP Subnet Selector

Tag Selector

EPG Selector

External EPG

External Subnet

Selectors not supported

Service EPG Selector

VXLAN BD Selector

Supported Policy Tags

BD Subnet Policy Tag

Endpoint IP Policy Tag

Endpoint MAC Policy Tag

VMM Policy Tags

Policy Tags not supported

Static Endpoint Policy Tags

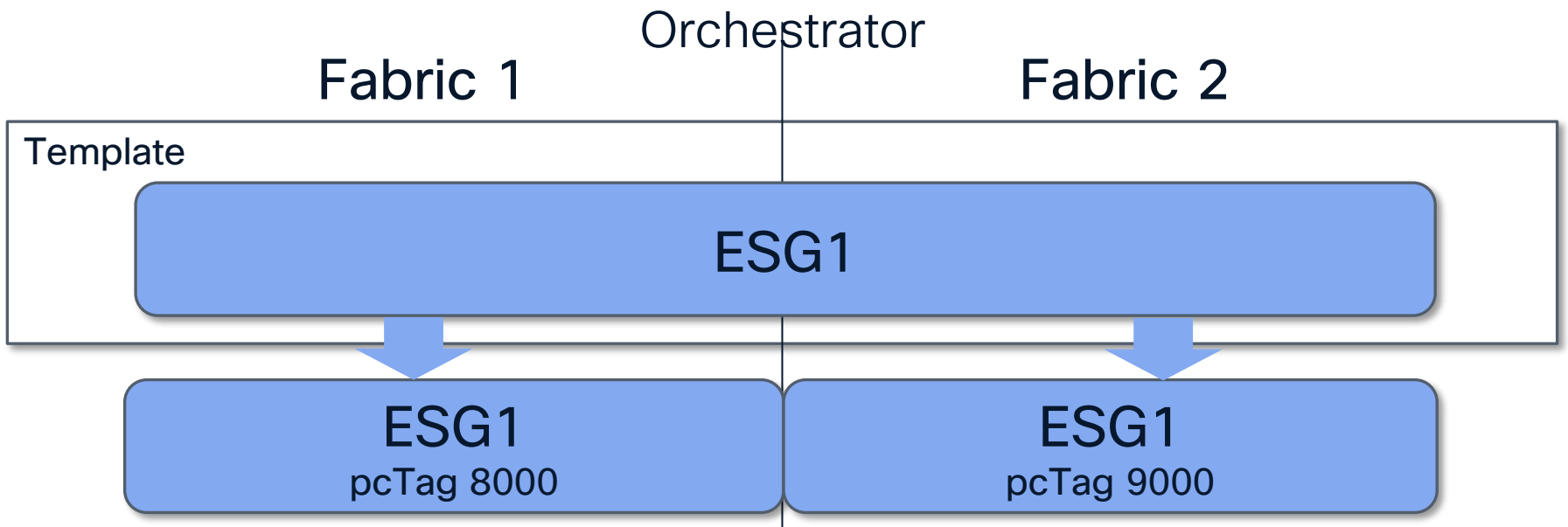
External EPG Subnet Tags

© 2025 Cisco and/or its affiliates. All rights reserved.

cisco

How can I apply the same ESG "Classification" to multiple Sites?

Multi-Site translates the ESG pcTag between the local and the remote Sites



Fabric 1 Spines Translation Table

	Remote Site	Local Site
pcTag	9000	8000

Fabric 2 Spines Translation Table

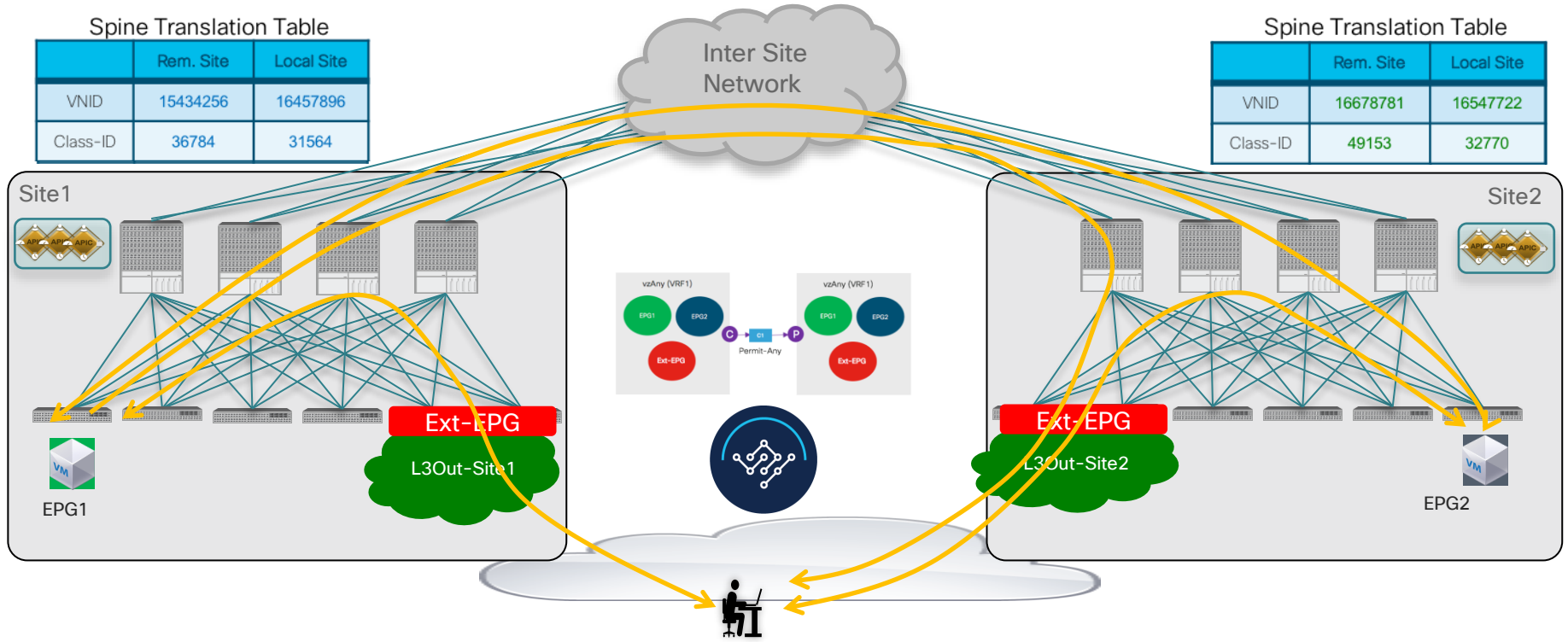
	Remote Site	Local Site
pcTag	8000	9000

© 2025 Cisco and/or its affiliates. All rights reserved.

cisco

Translations are also needed when using vzAny-to-vzAny permit

Enabling Free Communication inside a VRF



- Proper translation entries are created on the spines of both fabrics to enable east-west and north-south communication. It creates translations and shadow objects automatically.
- Supported also for connecting to the external Layer 3 domain

© 2025 Cisco and/or its affiliates. All rights reserved.

Data-Plane Communication across Sites

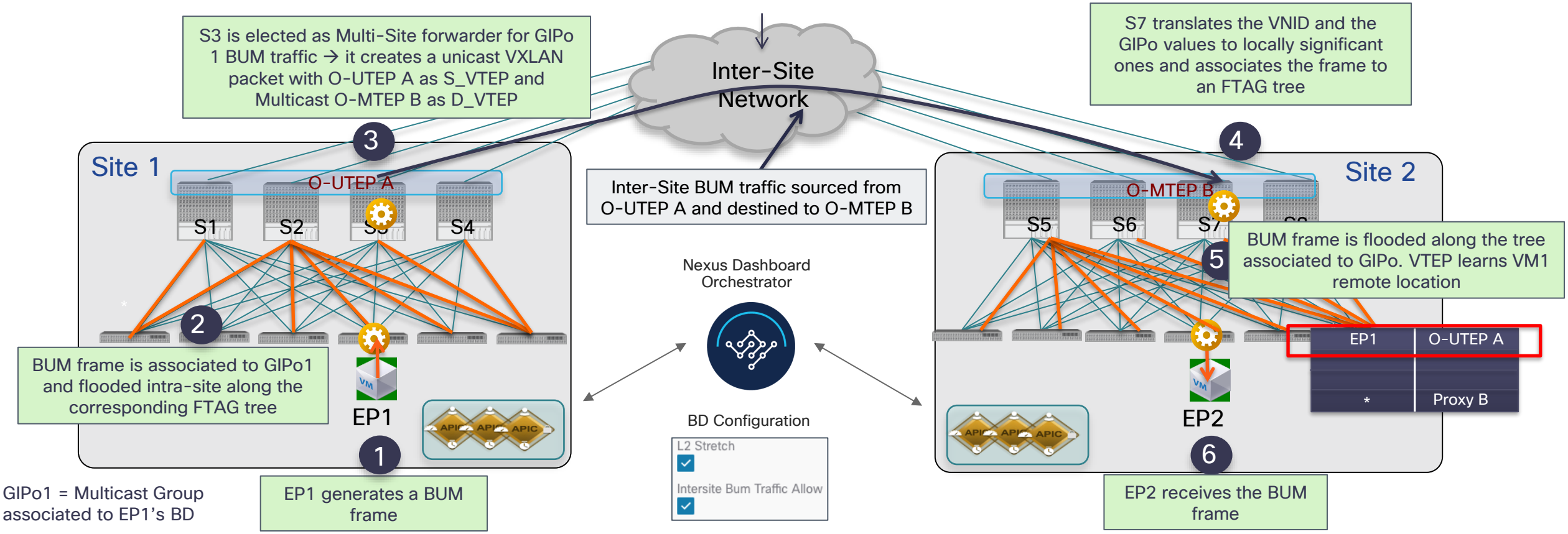
© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Site: how is BUM Traffic forwarded?

Inter-Site Layer 2 BUM* Forwarding

No Multicast Requirement in Backbone, Head-End Replication (HER) for any Layer 2 BUM traffic)

*BUM – Broadcast, Unknown Unicast, Multicast

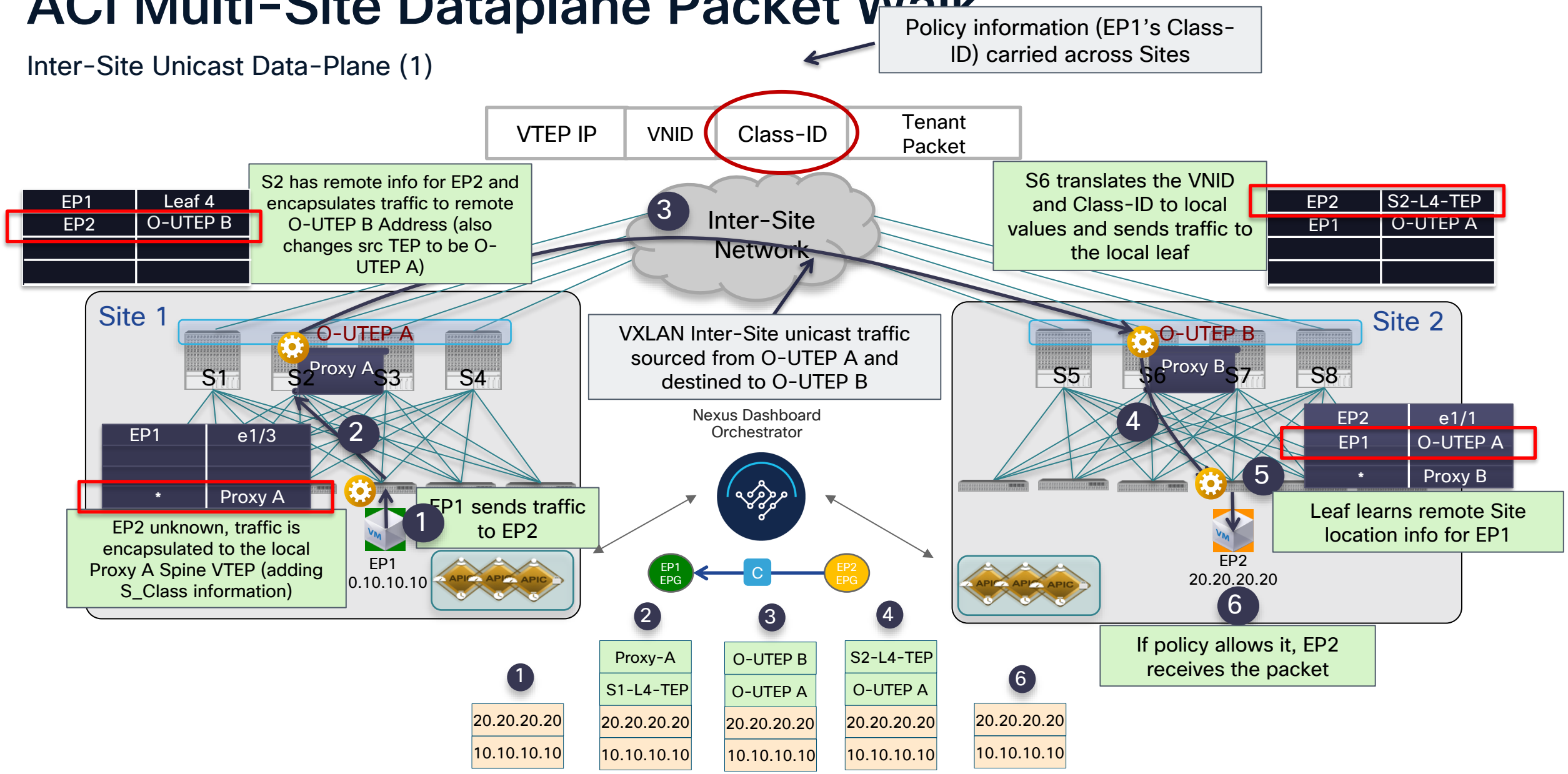


© 2025 Cisco and/or its affiliates. All rights reserved.

cisco

ACI Multi-Site Dataplane Packet Walk

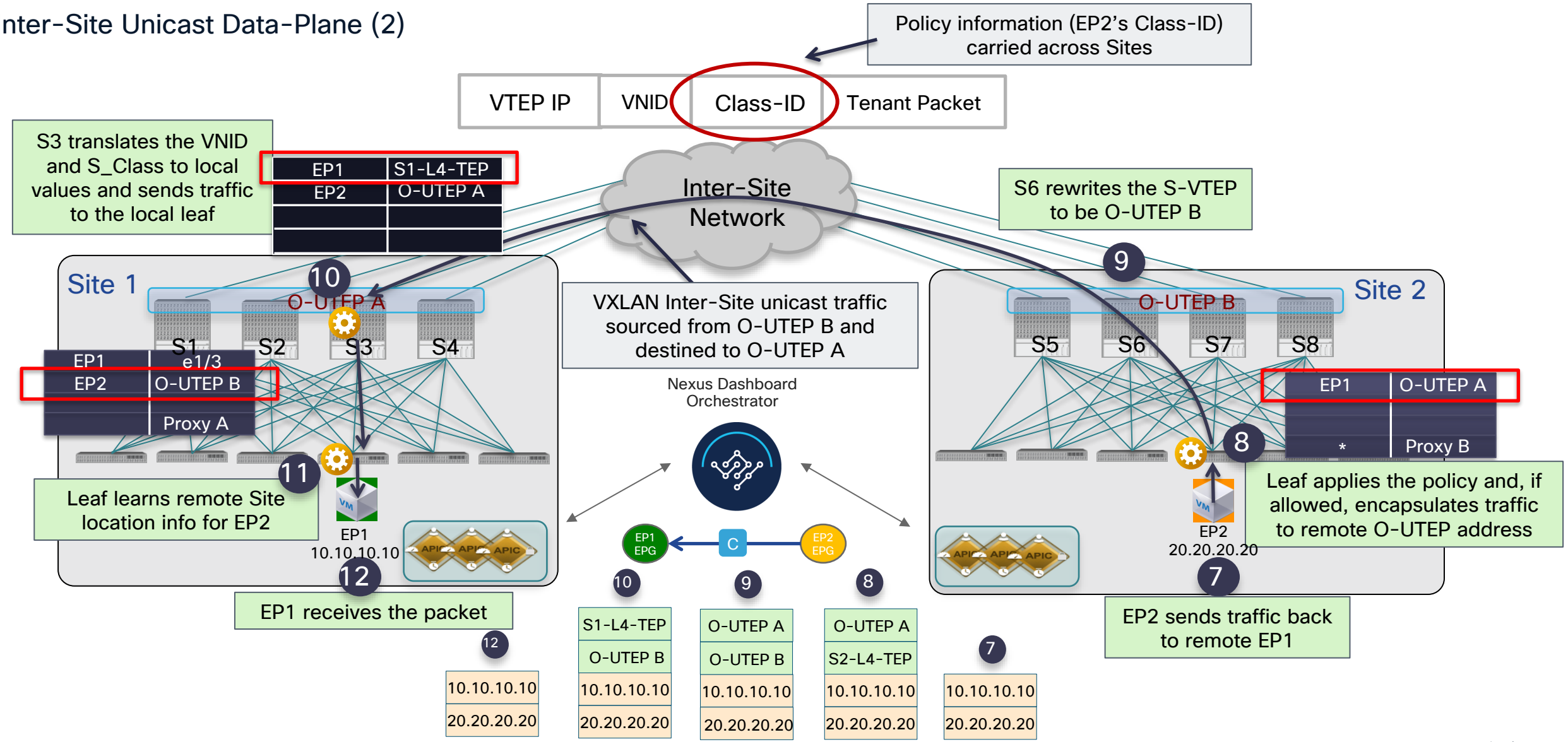
Inter-Site Unicast Data-Plane (1)



© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Site Dataplane Packet Walk

Inter-Site Unicast Data-Plane (2)

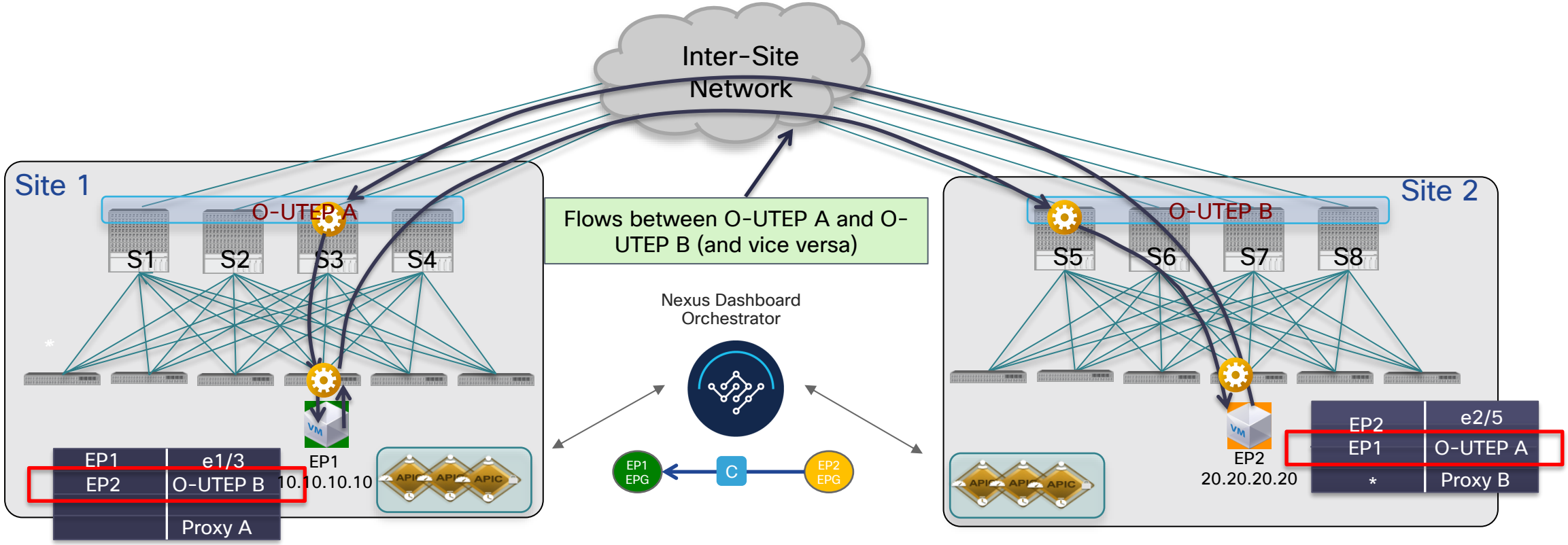


© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Site Dataplane Packet Walk

Inter-Site Unicast Data-Plane (3)

From this point EP1 to EP2 communication is encapsulated Leaf to Remote Spine O-UTEPs in both directions



© 2025 Cisco and/or its affiliates. All rights reserved.

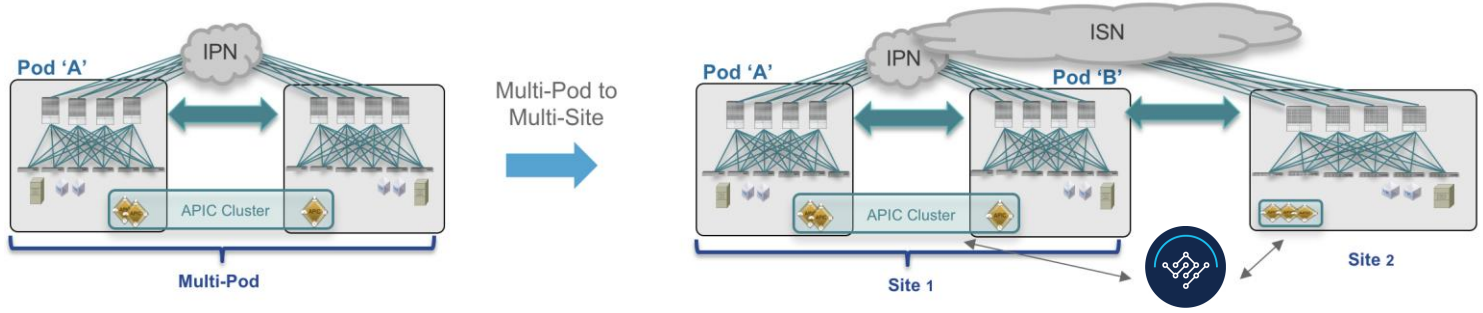
ACI Multi-Pod and ACI Multi-Site Integration

© 2025 Cisco and/or its affiliates. All rights reserved.

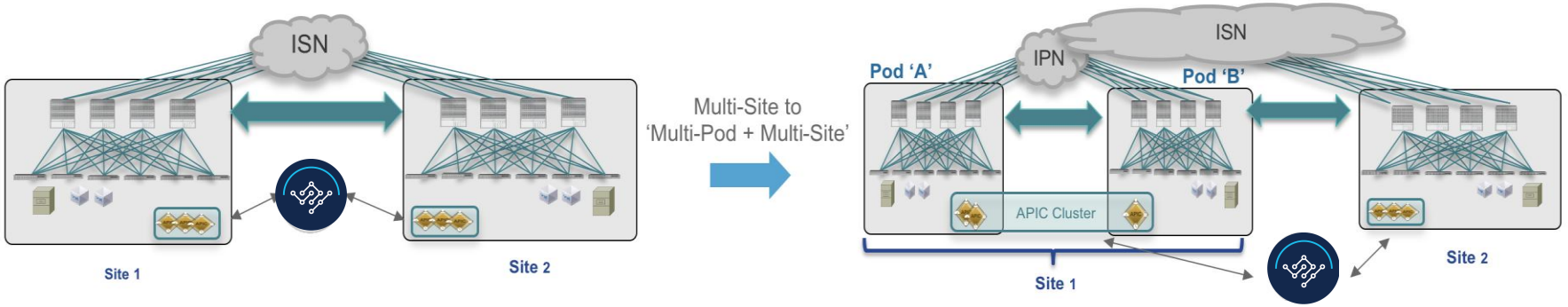
ACI Multi-Pod and Multi-Site Can be Used Together

Main Use Cases

- Adding a Single Pod Fabric as a ‘DR Site’ to the Multi-Site domain



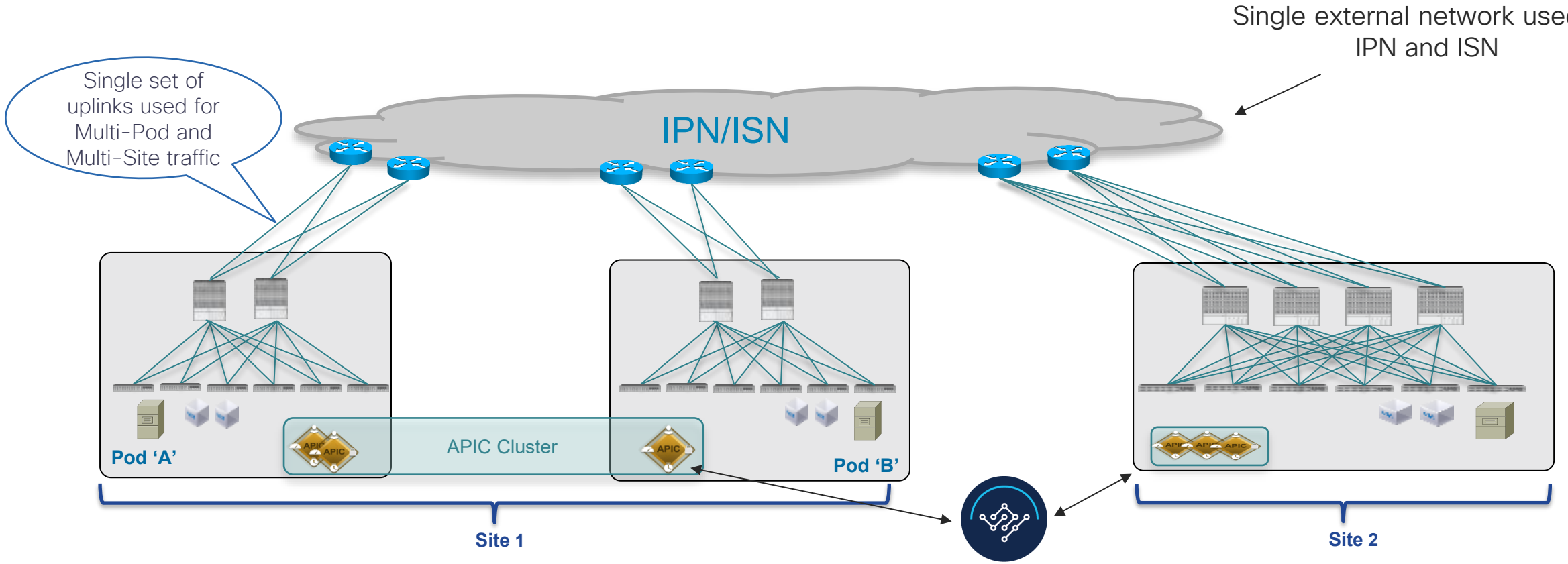
- Converting a single Pod Fabric (already added to NDO) to a Multi-Pod fabric



© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Pod and Multi-Site Integration

Connectivity between Pods and Sites

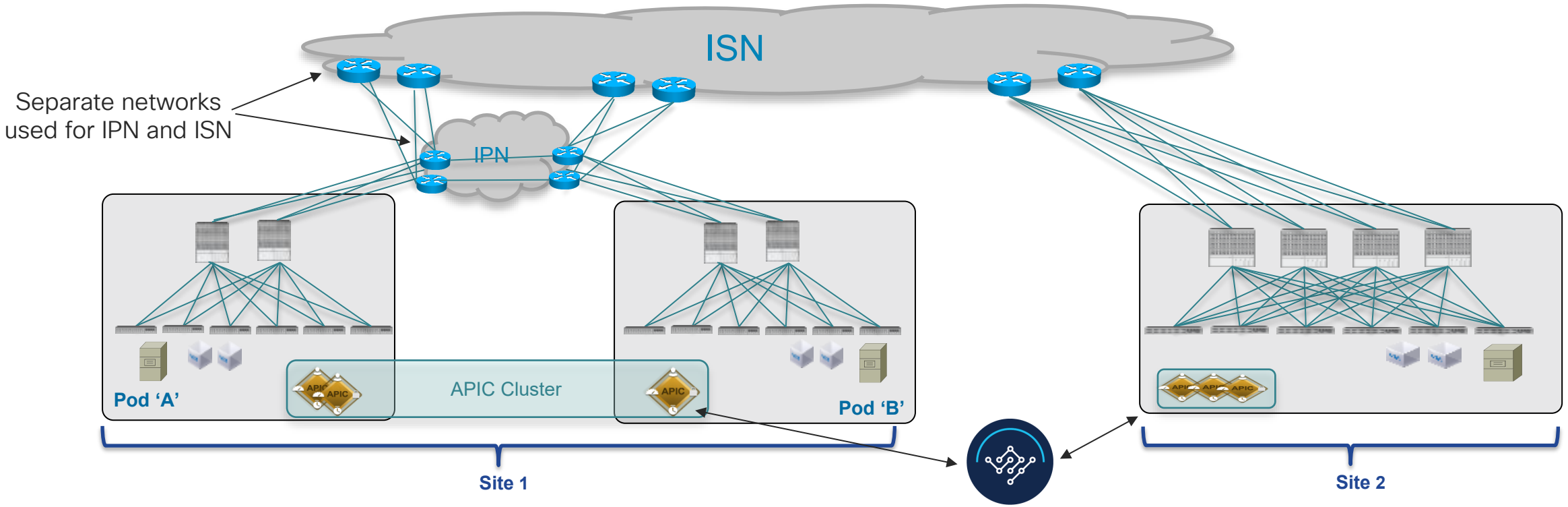


- The same IPN/ISN infrastructure can be used to carry Multi-Pod and Multi-Site traffic
  - Could be a dedicated VRF deployed in a shared Layer 3 network
- A single set of uplinks must be used between each spine and the IPN/ISN for both types of traffic

© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Pod and Multi-Site Integration

Connectivity between Pods and Sites

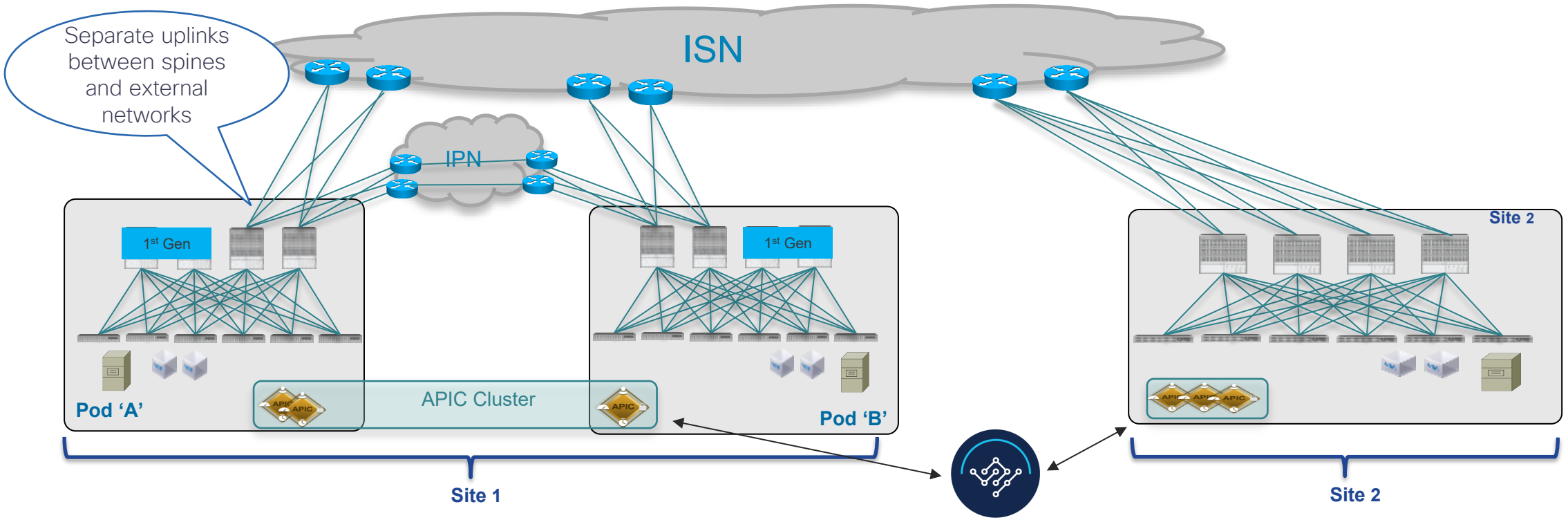


- Separate IPN and ISN infrastructures can also be used
- The two networks must be connected to allow for inter-site communication

© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Pod and Multi-Site Integration

Connectivity between Pods and Sites – Not Supported Topology

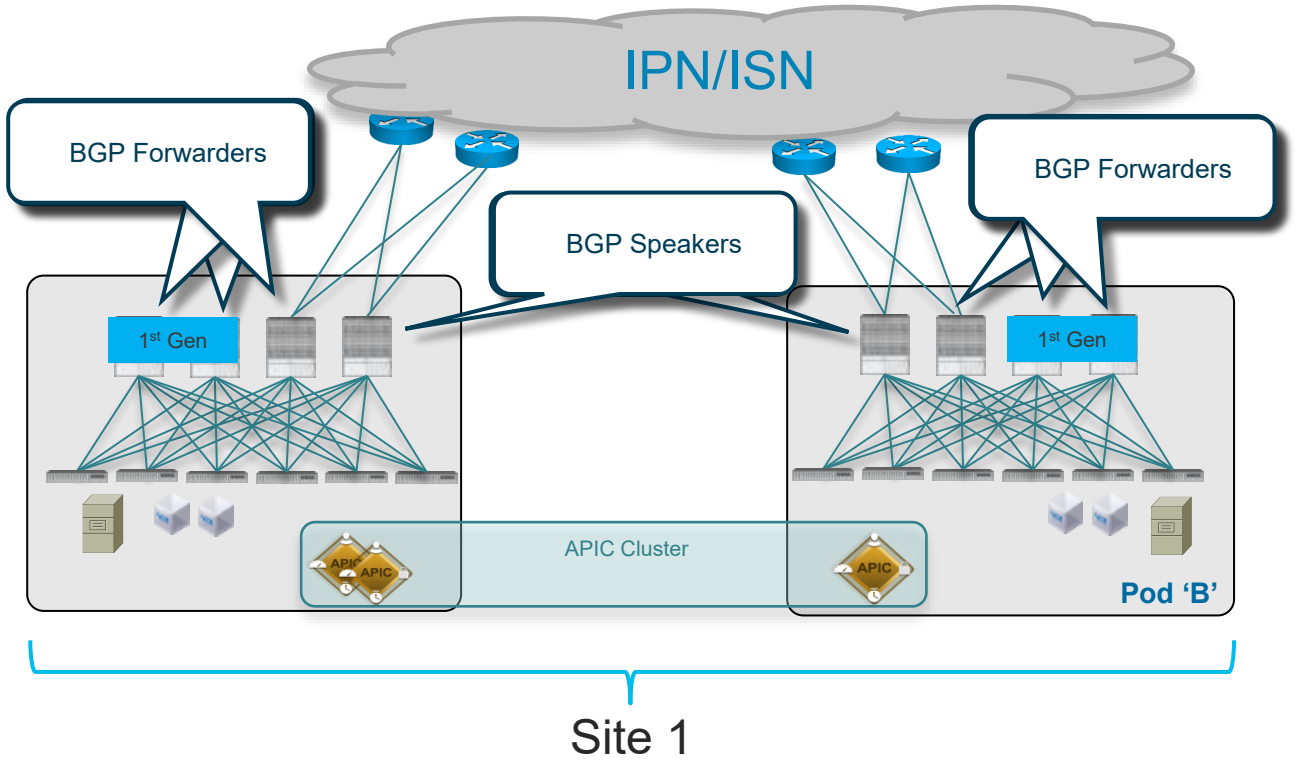


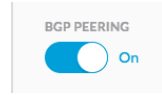
- The use of separate set of uplinks between spines and IPN and ISN is not a supported option
- The election of the DF link (for each BD) between spines of the same Pod is done across all the available uplinks
  - BUM forwarding between Pods would break if one of the link to the ISN became the DF for a given BD

© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Pod and Multi-Site Integration

BGP Spine Roles



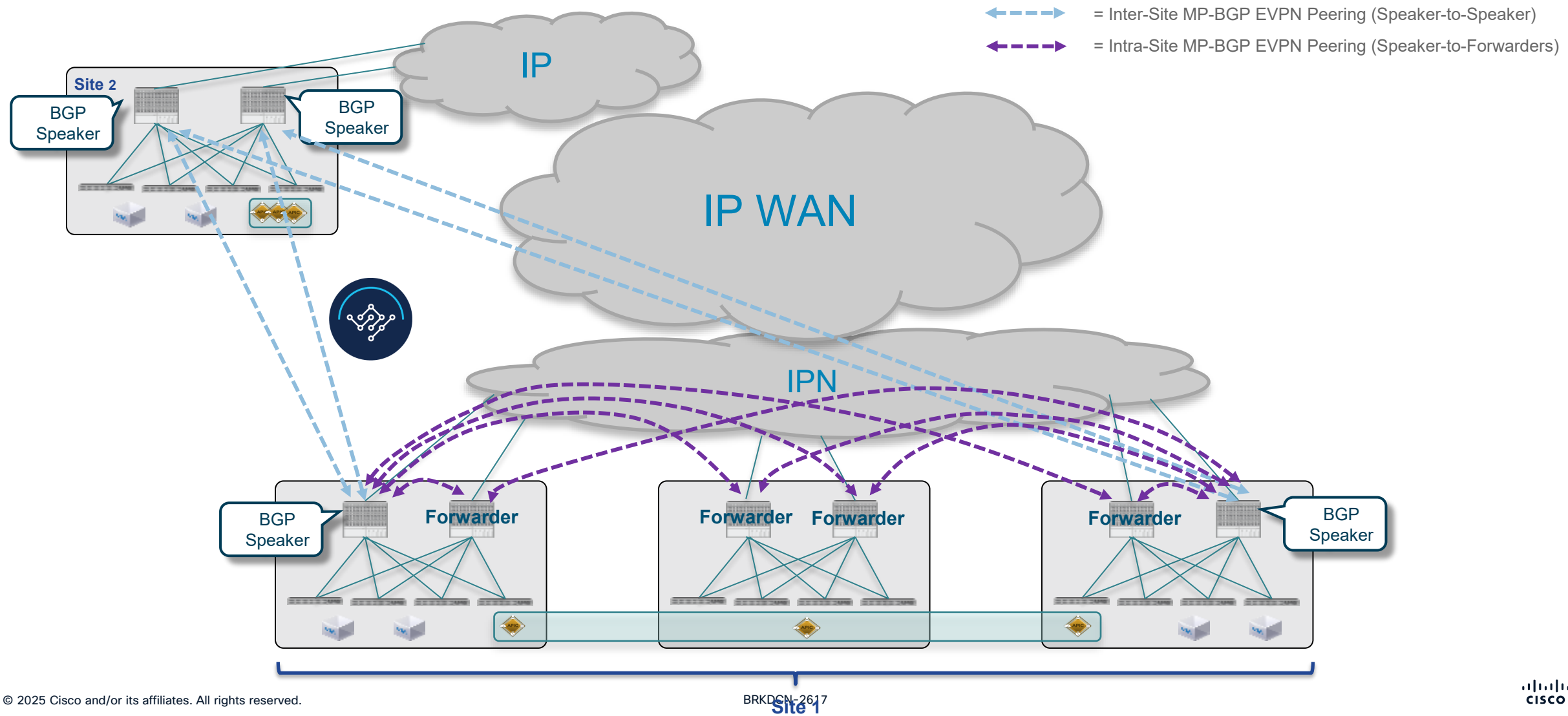
- Spines in each Multi-Pod fabric can have one of those two roles:
  1. BGP Speakers: establish EVPN peerings with BGP speakers in **remote sites** and with BGP Forwarders in the **local site** (intra- and inter- Pods)
     - Recommended to deploy two speakers per Multi-Pod fabric (in separate Pods)
     - Explicitly configured on MSO 

     - Multi-Site Speaker must be a Multi-Pod spine as well
  2. BGP Forwarders: establish BGP EVPN peerings with BGP speakers in the **local site**
     - All the spines that are not speakers implicitly become forwarders

© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Pod and Multi-Site Integration

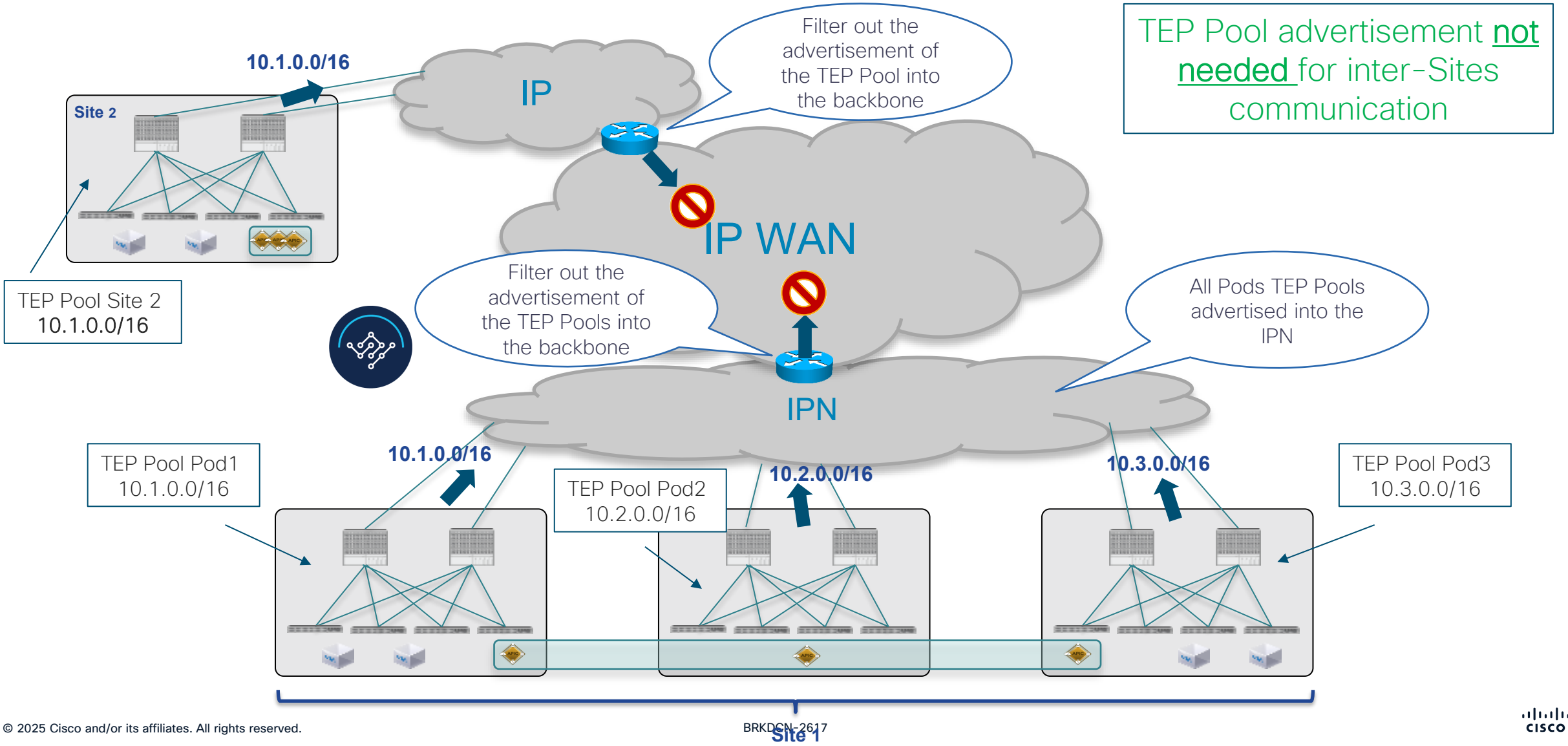
Inter-Site and Intra-Site EVPN Sessions



© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Pod and Multi-Site Integration

Internal TEP Pools Filtering



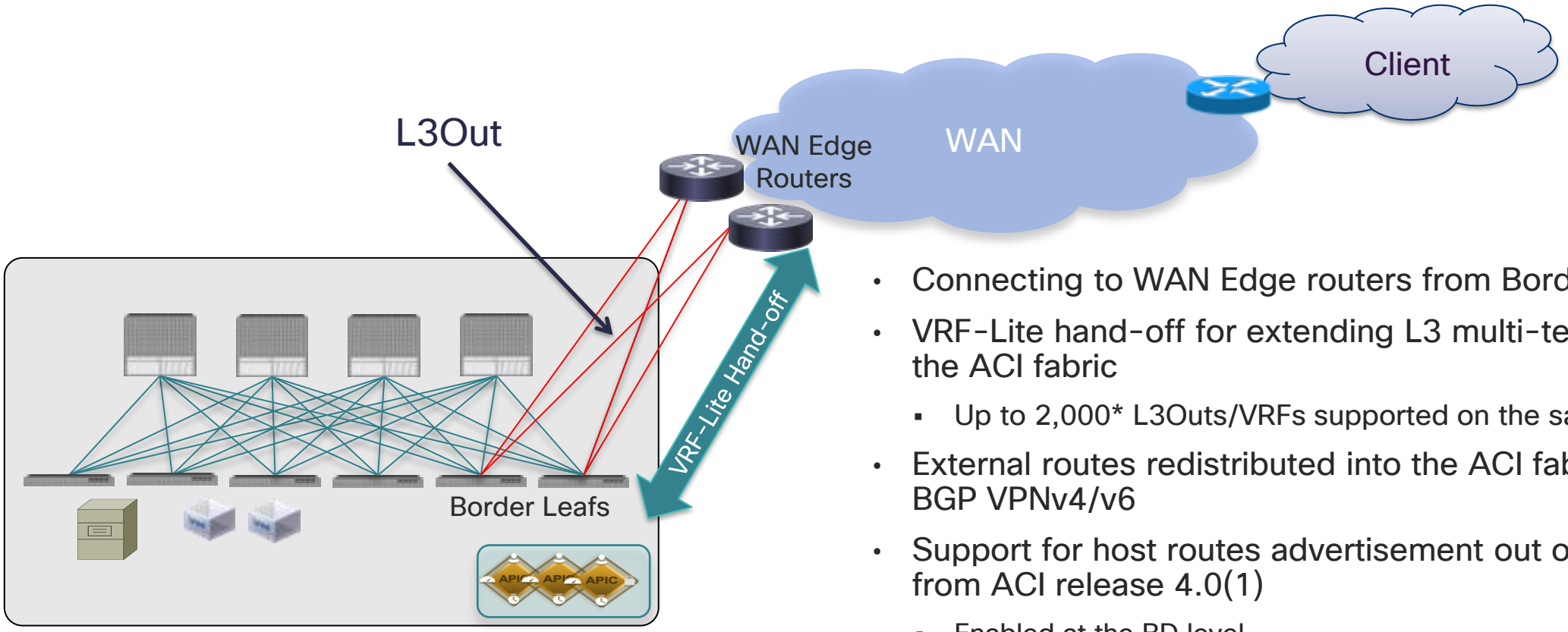
© 2025 Cisco and/or its affiliates. All rights reserved.

Connecting to the External L3 Domain with Multi-Pod

© 2025 Cisco and/or its affiliates. All rights reserved.

Connecting to the External Layer 3 Domain

VRF-lite or SR-MPLS



- Connecting to WAN Edge routers from Border Leaf nodes
- VRF-Lite hand-off for extending L3 multi-tenancy outside the ACI fabric
  - Up to 2,000* L3Outs/VRFs supported on the same BL nodes pair
- External routes redistributed into the ACI fabric with MP-BGP VPNv4/v6
- Support for host routes advertisement out of the ACI Fabric from ACI release 4.0(1)
  - Enabled at the BD level
- Support for L3 Multicast and Shared L3Out

* From release 6.0(3), 800 L3Outs 6.0(2) and earlier

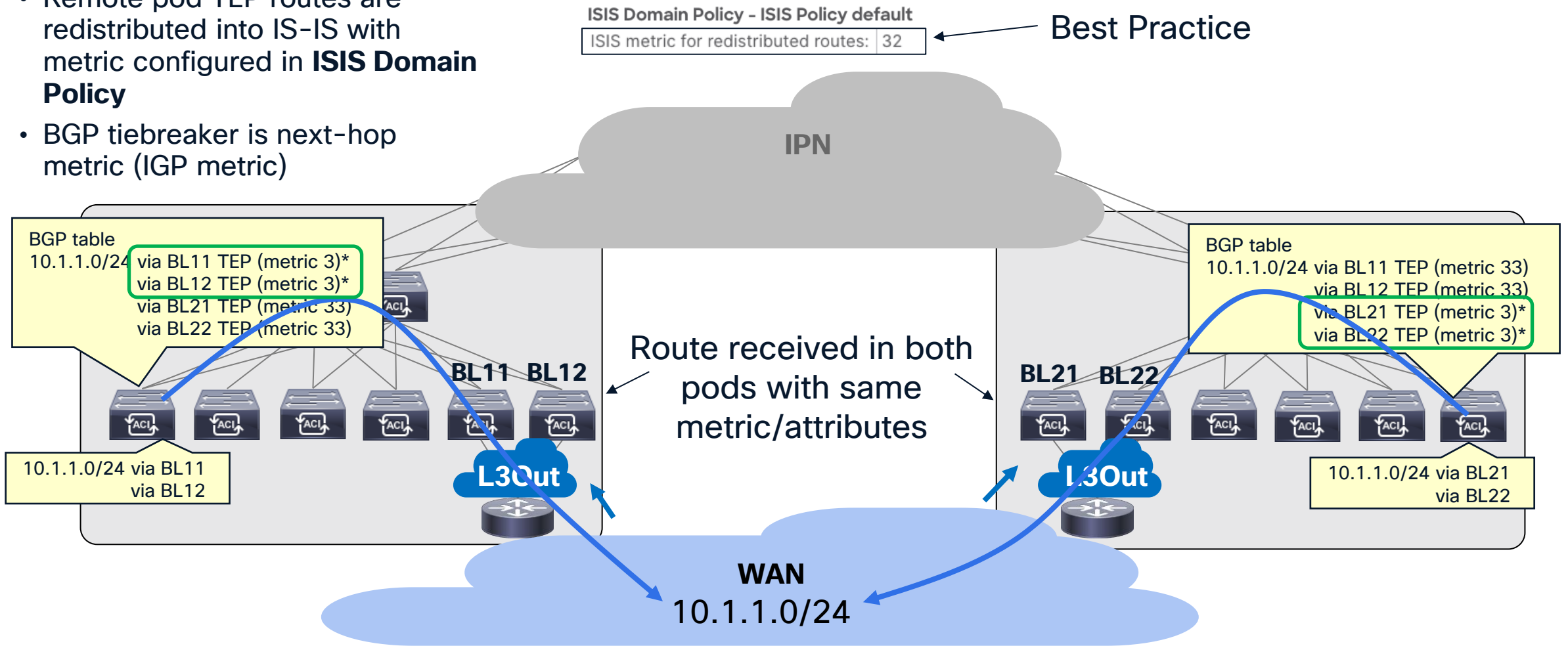
© 2025 Cisco and/or its affiliates. All rights reserved.

cisco

Route Preference with Multiple L3Outs in different PODs

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)

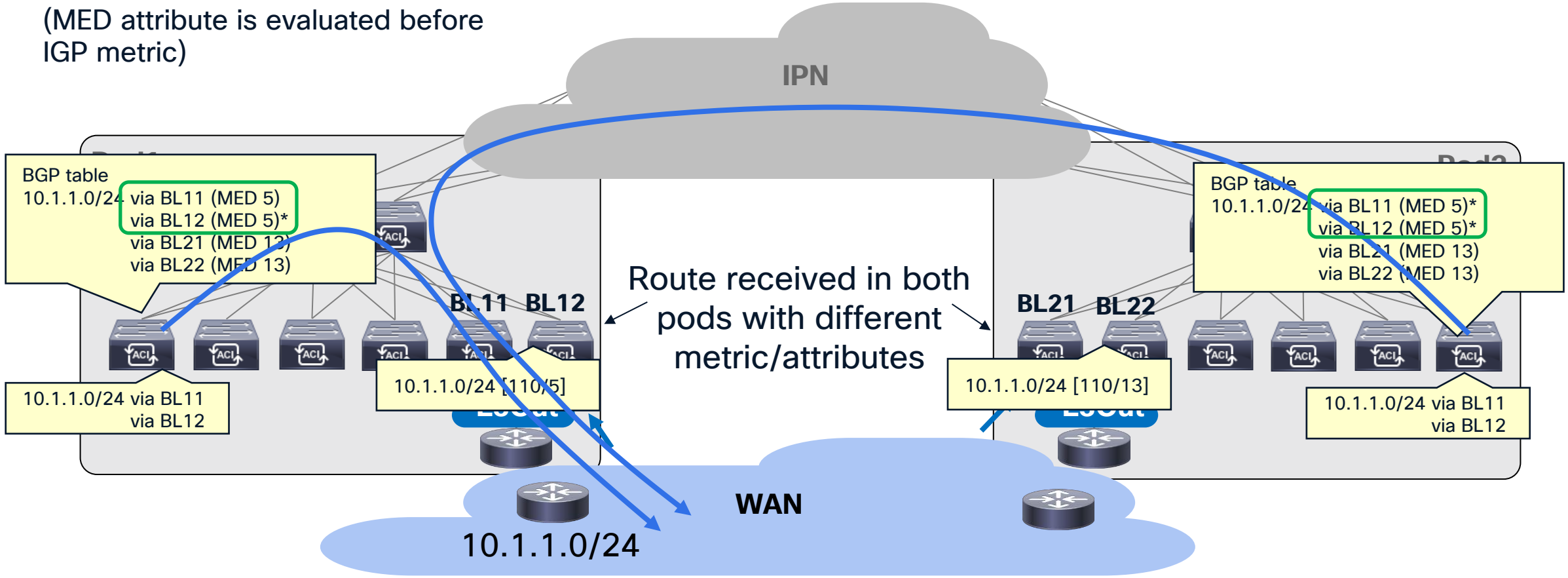


© 2025 Cisco and/or its affiliates. All rights reserved.

Route Preference with Multiple L3Outs in different PODs

Influencing the outbound path: Remote pod L3Out may be used if it has a better external metric

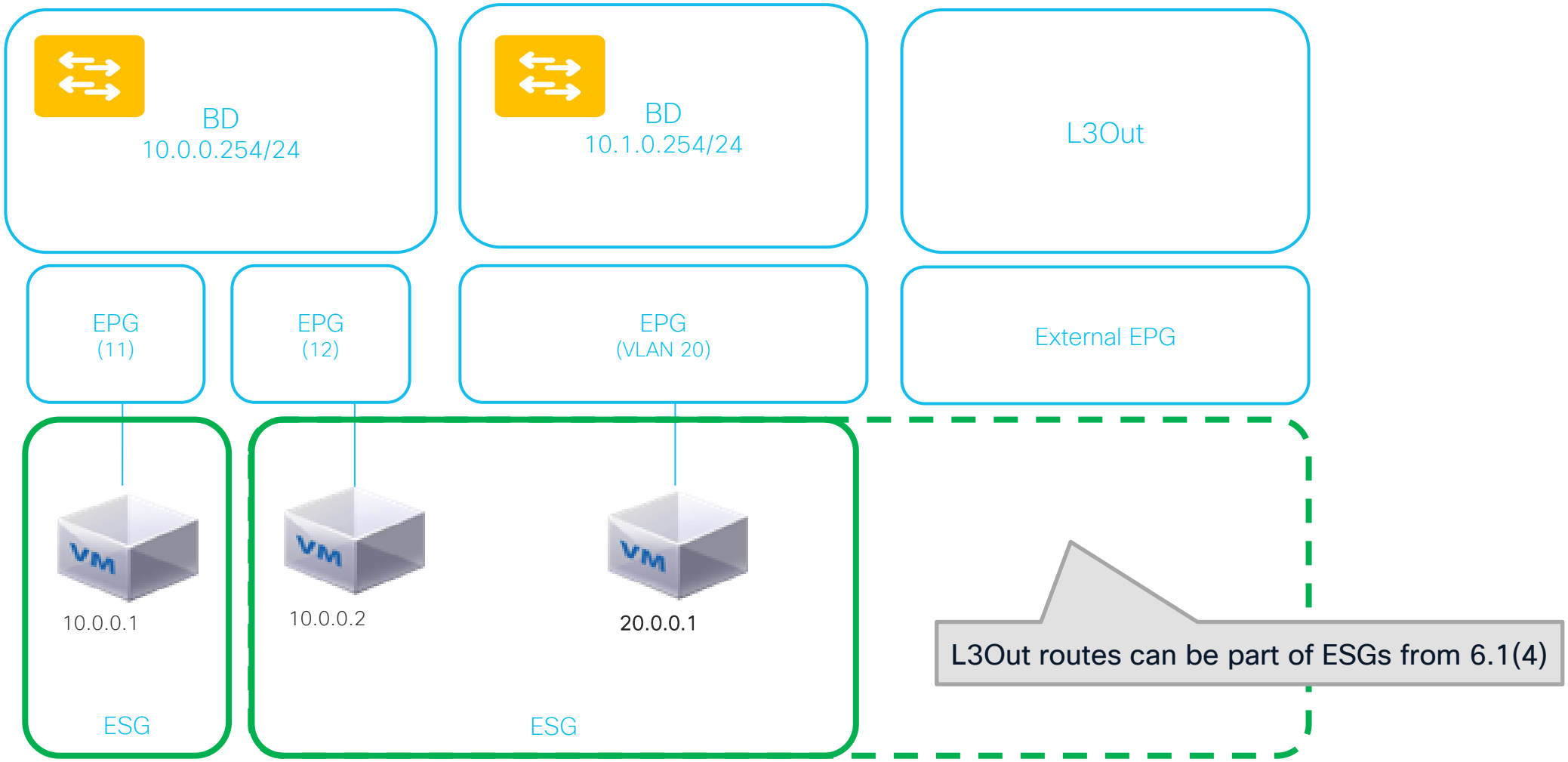
- BGP MED is set to OSPF metric when redistributed into fabric
- Routes with lower MED are selected (MED attribute is evaluated before IGP metric)



© 2025 Cisco and/or its affiliates. All rights reserved.

ESG for L3Out

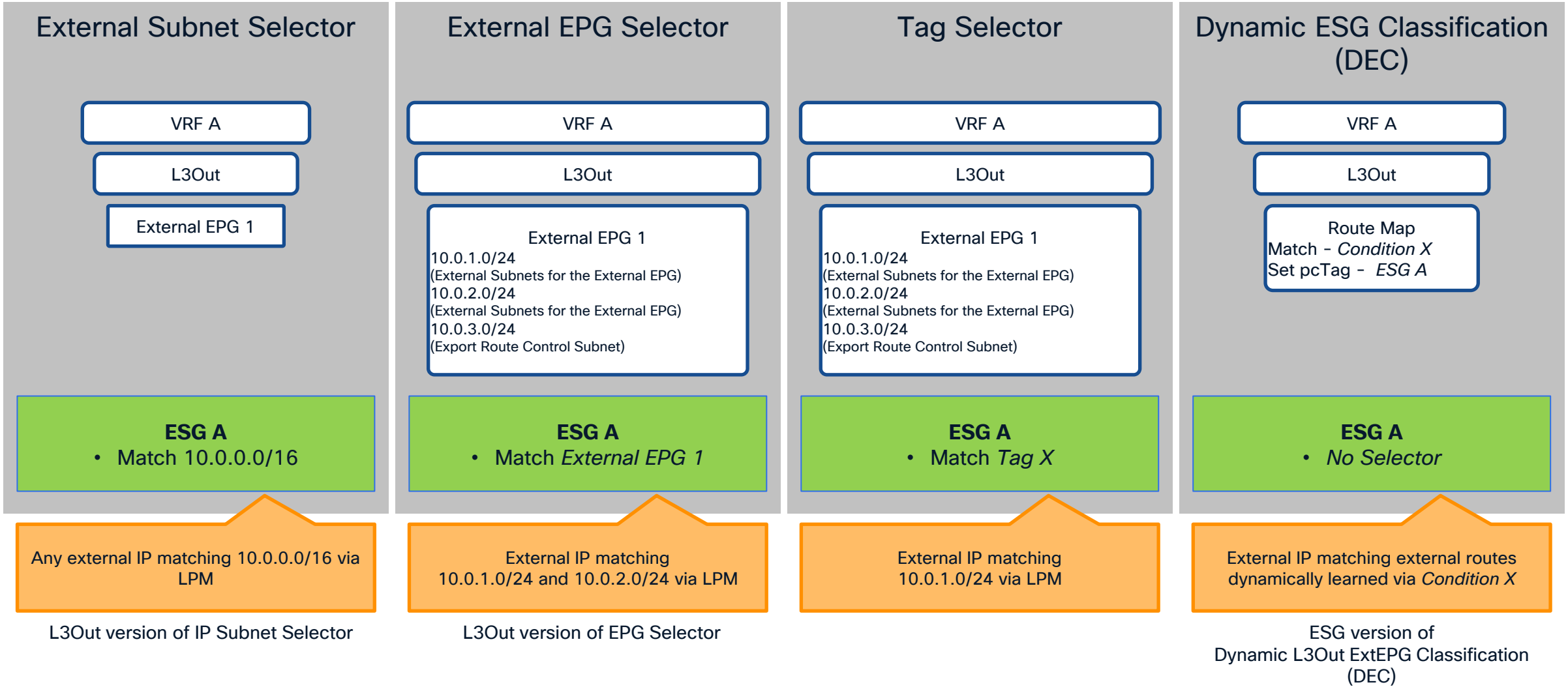
Added in ACI 6.1(4)



© 2025 Cisco and/or its affiliates. All rights reserved.

Options to classify external IP addresses to ESG

MultiPod Only (not Multi-Site)



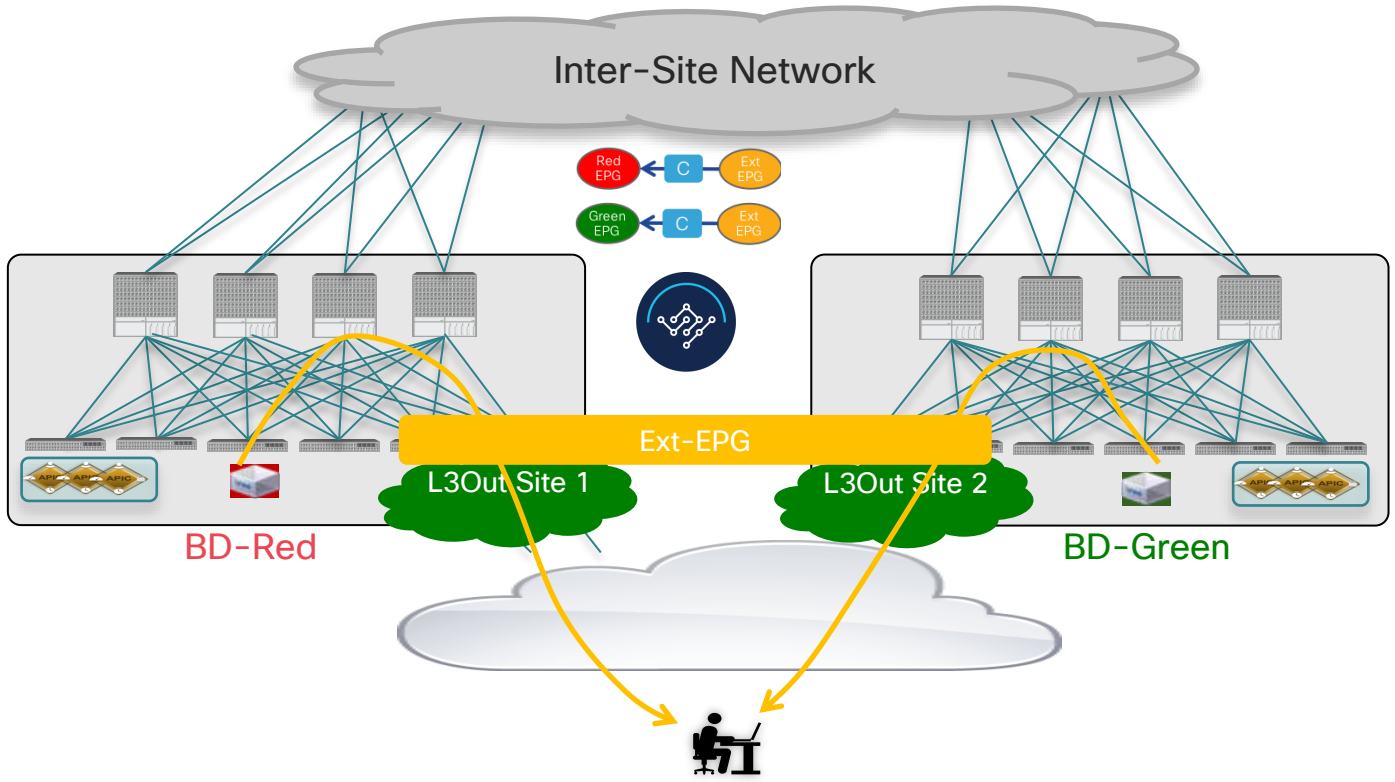
© 2025 Cisco and/or its affiliates. All rights reserved.

Connecting to the External L3 Domain with Multi-Site

© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Site and L3Out

Stretching or Not Stretching the Ext-EPG?

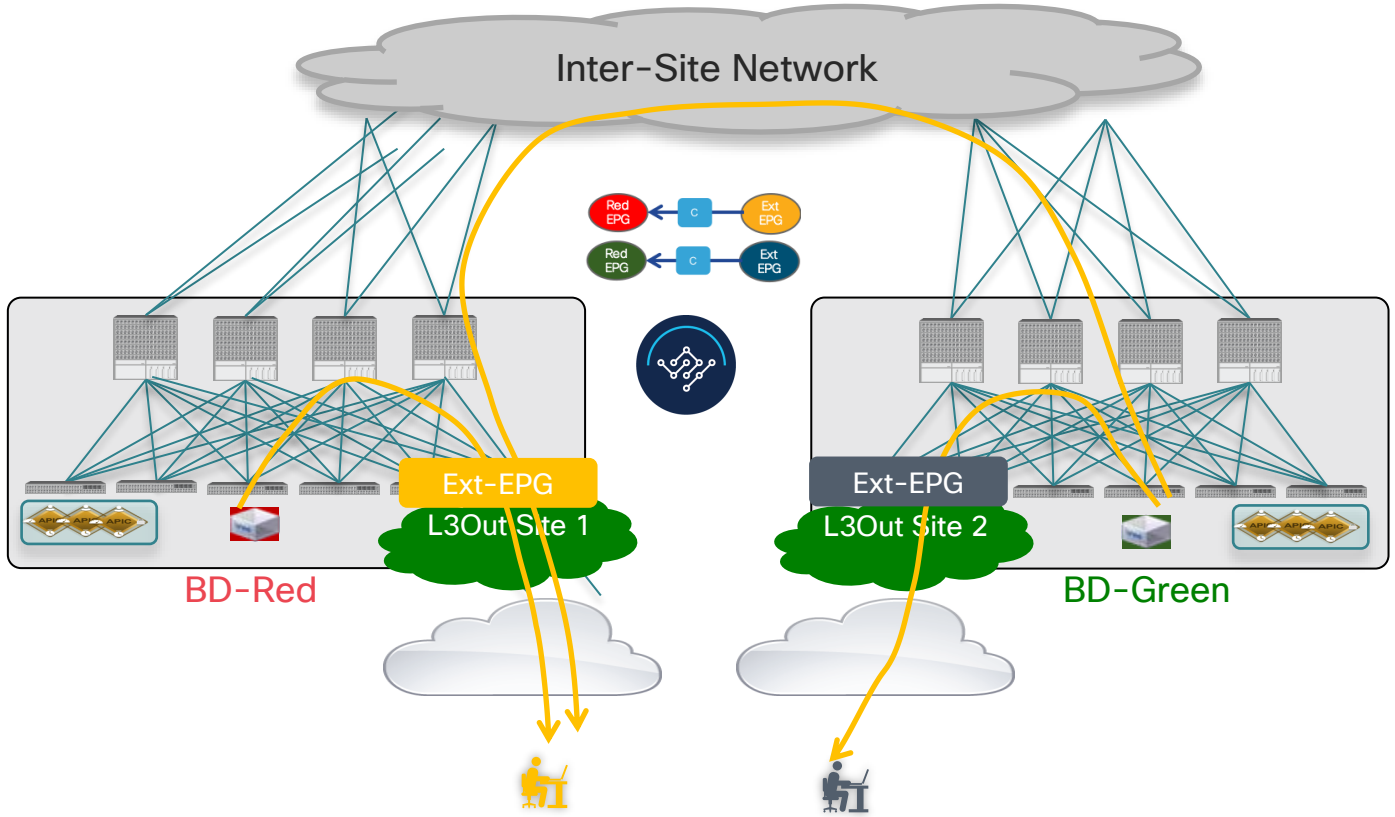


- The Ext-EPG can be defined in a template associated to multiple sites (stretched object)
  - The Ext-EPG must then be mapped to the local L3Outs in the “site level” section of the template configuration
  - L3Outs remain independent objects defined in each site
- Recommended when the L3Outs in the separate sites provide access to a common set of external resources (as the WAN)
  - Simplifies the policy definition and external traffic classification
  - Still allows to apply route-map polices on each L3Out (since we have independent APIC domains)

© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Site and L3Out

Stretching or Not Stretching the Ext-EPG?



- Separate Ext-EPGs can be defined in templates mapped to separate sites (non stretched objects)
  - Each Ext-EPG can be mapped to the local L3Out in the “global” or “site level” section of the template configuration
- Allows to apply different policies to each Ext-EPGs at different time

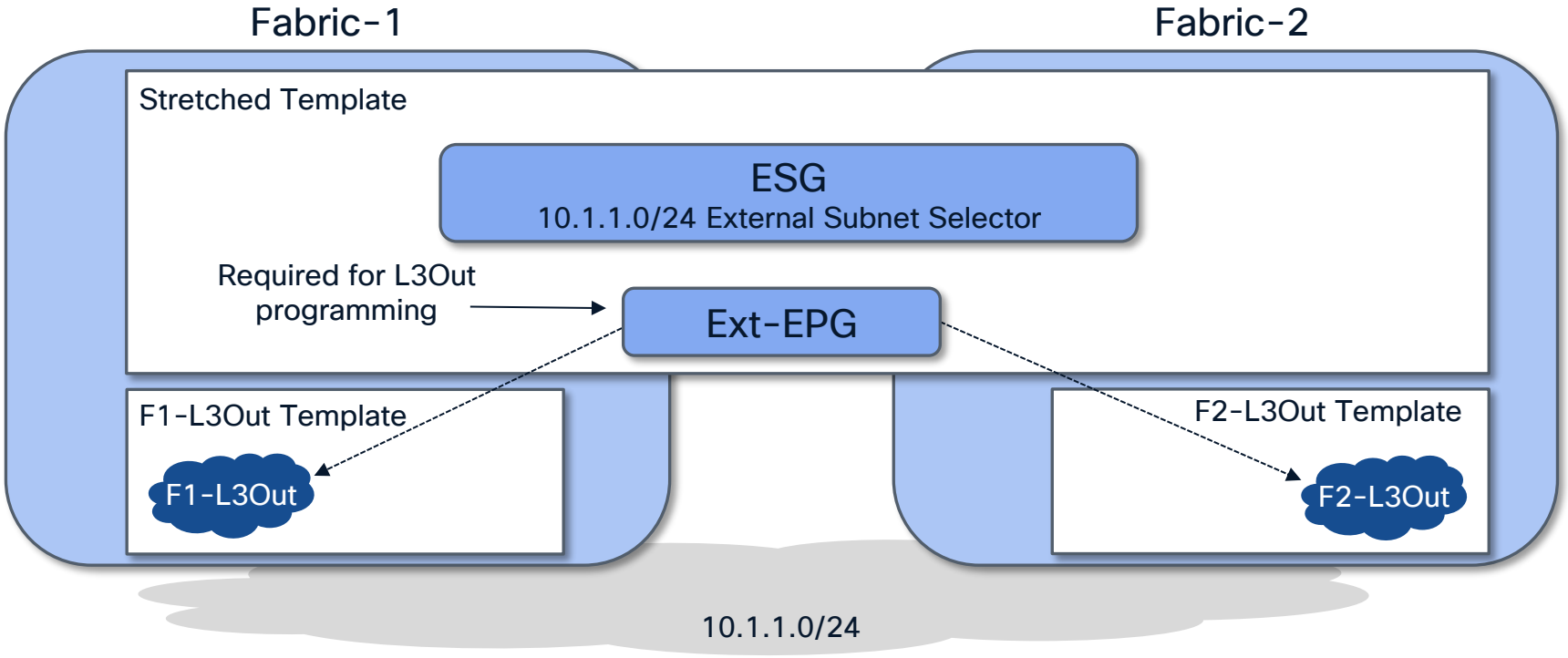
© 2025 Cisco and/or its affiliates. All rights reserved.

Using ESGs that Match L3Outs

© 2025 Cisco and/or its affiliates. All rights reserved.

with ND 4.1 and ACI 6.1(4) you can use ESGs for External Prefixes

ESG selectors for External Prefixes

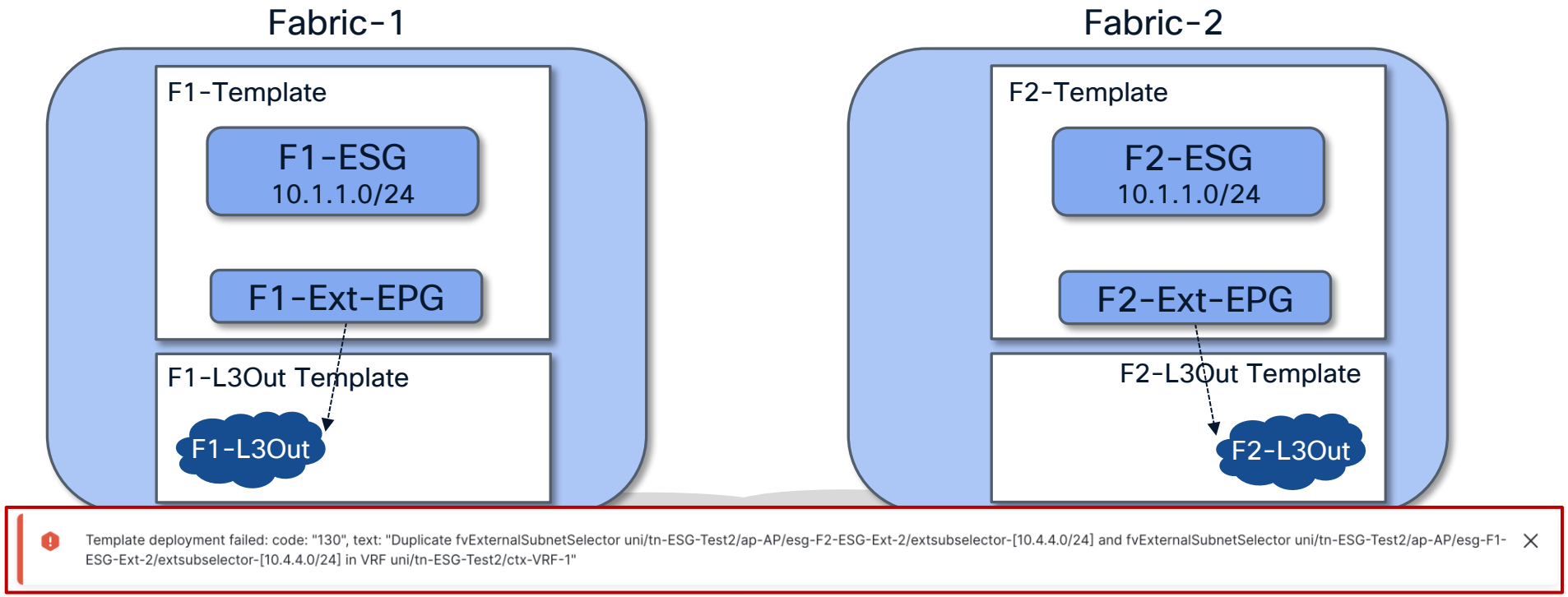


- The ESG should be stretched to the fabrics where the external prefix is reachable via L3Outs
- If using ESGs with External Subnet Selectors you must still configure and Ext-EPG. The Ext-EPG is required for programming the L3Out on the border leafs. Ext-EPG can be stretched or non-stretched since it should not have any classification prefixes configured

© 2025 Cisco and/or its affiliates. All rights reserved.

Multi-Site ESGs Has Validations to Prevent Overlapping ESGs

Non-stretched ESGs with same prefix Blocked if shadowed



- It is Not Recommended to configure ESGs in multiple fabrics with the same External Subnet selector prefixes. These selectors are shadowed if the ESG is shadowed
- Configuration creating the shadow will be blocked.
- Always Stretch ESGs configured to match the same External Prefixes (same as External EPG recommendation)

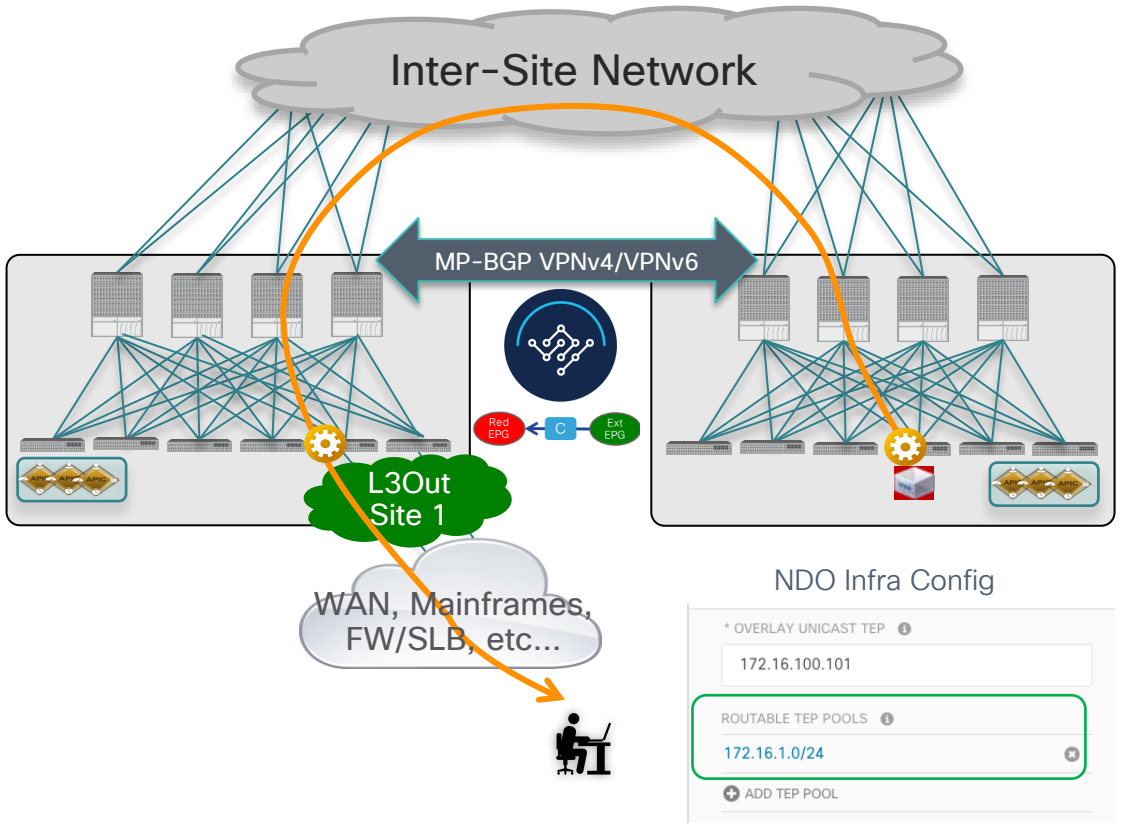
© 2025 Cisco and/or its affiliates. All rights reserved.

Intersite L3Out Support

© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Site and L3Out

Support of Intersite L3Out



- Multi-Site has been designed assuming that each site would use its local L3Out: in the original Multi-Site designs sites where just exchanging endpoint reachability via EVPN
- Since ACI 4.2(1) external prefixes can be exchanged across sites via MP-BGP VPNV4/VPNv6 sessions between spines
- Traffic is directly encapsulated to the TEP of the remote BL nodes
  - The BL nodes will get assigned an address part of an additional (configurable) prefix that must be routable across the ISN
- Same solution also supports transit routing across sites (L3Out to L3Out)

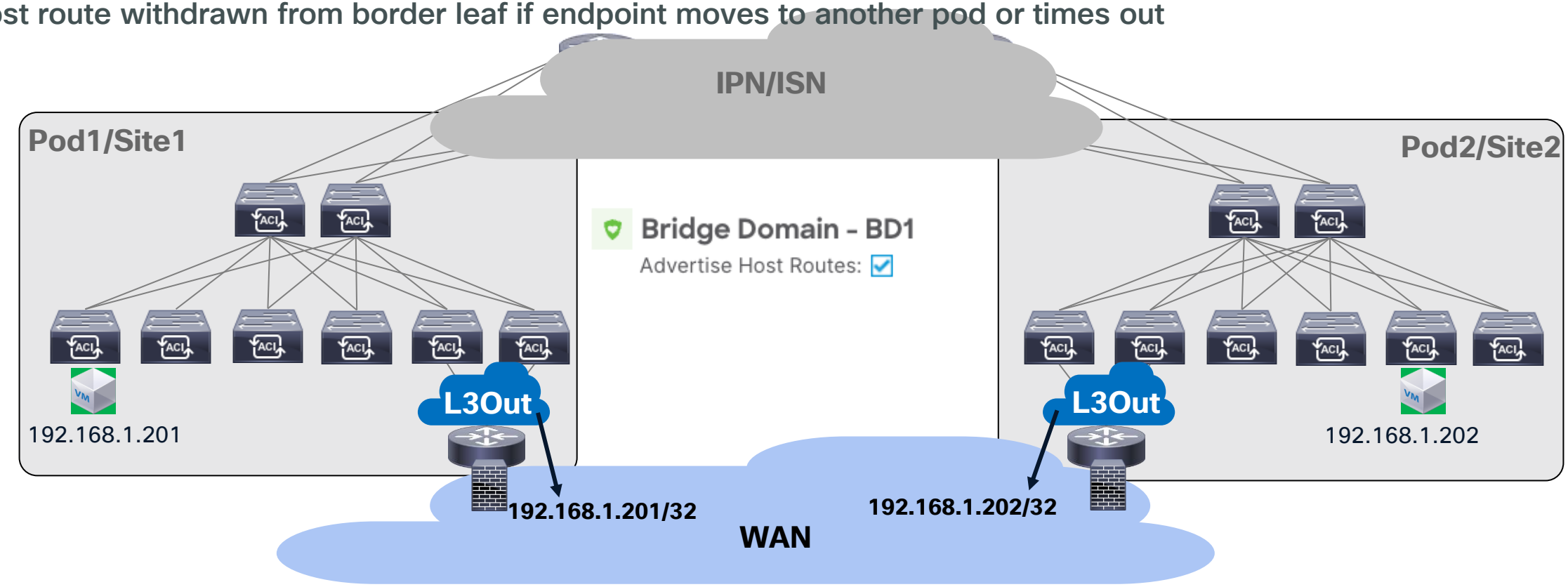
© 2025 Cisco and/or its affiliates. All rights reserved.

Influencing the inbound traffic

© 2025 Cisco and/or its affiliates. All rights reserved.

Influencing inbound path: Host route advertisement

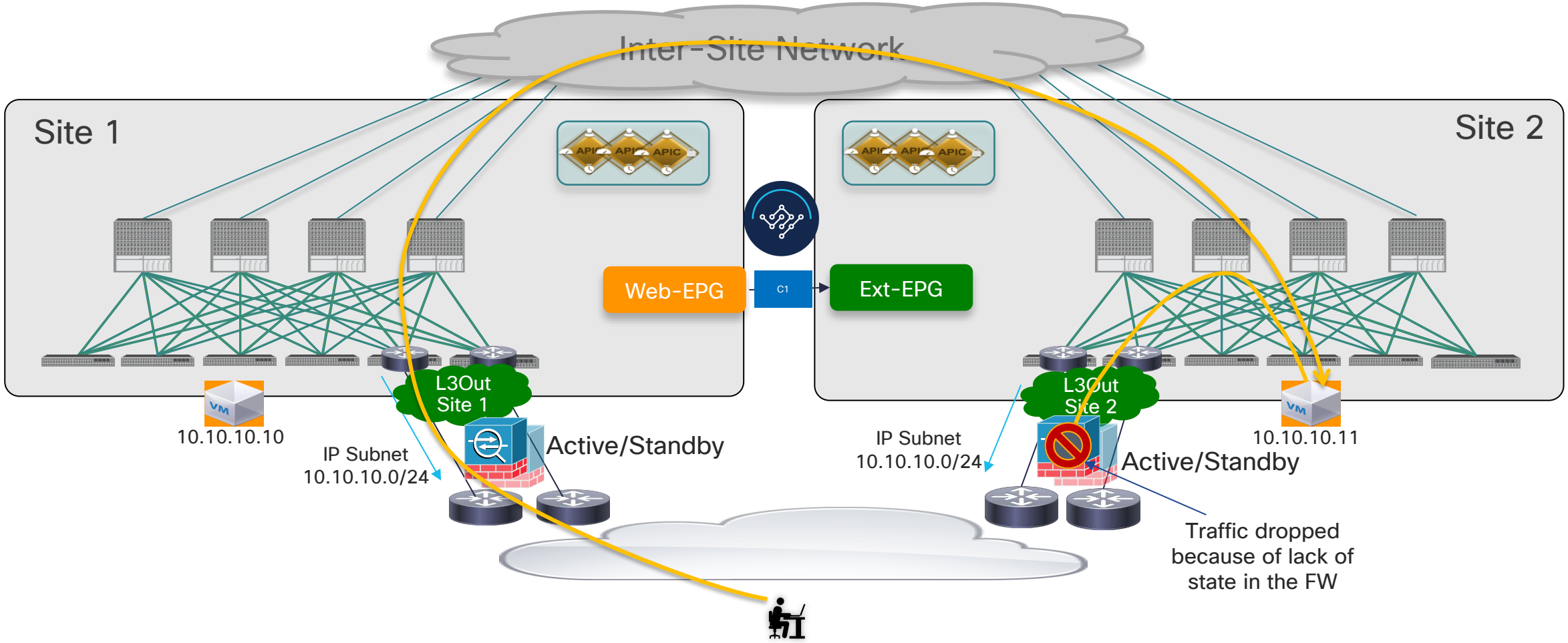
- Host route advertisement can be enabled per BD, and in Multi-Site you can enable this on the same BD per Site
- Spines know which endpoint is local or remote
- Spine downloads endpoints local to the Pod/Site to local BLs as /32 host routes
- Host route withdrawn from border leaf if endpoint moves to another pod or times out



© 2025 Cisco and/or its affiliates. All rights reserved.

Example where Host Routes are useful

Multi-Site with Perimeter Firewalls

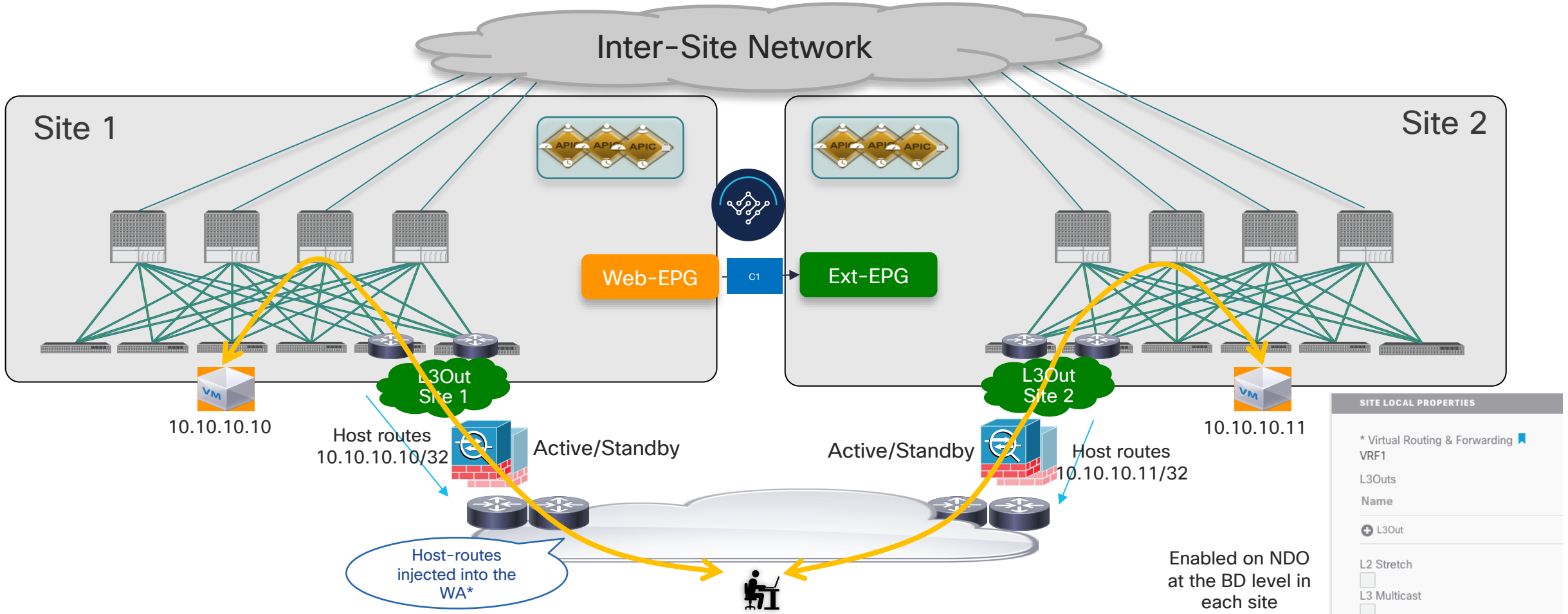


© 2025 Cisco and/or its affiliates. All rights reserved.

cisco

Solving Asymmetric Routing Issues

Use of Host-Routes Advertisement



- Ingress optimization requires <u>host-routes advertisement</u> on the L3Out
  - Native support on ACI Border Leaf nodes available from ACI release 4.0(1)
  - Not currently supported for SR-MPLS L3Outs

© 2025 Cisco and/or its affiliates. All rights reserved.

Network Services Integration with Multi-Pod

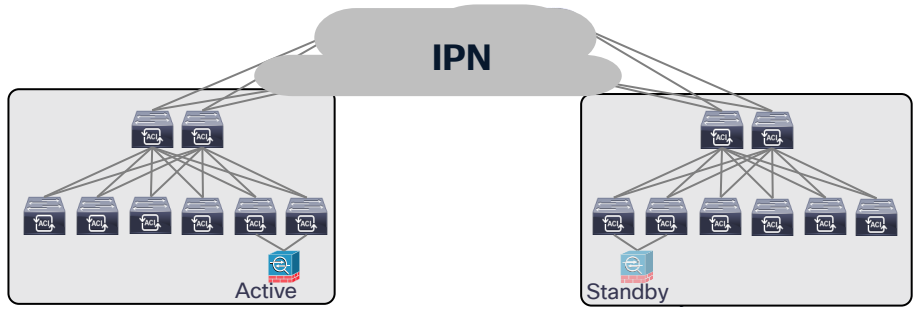
© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Pod

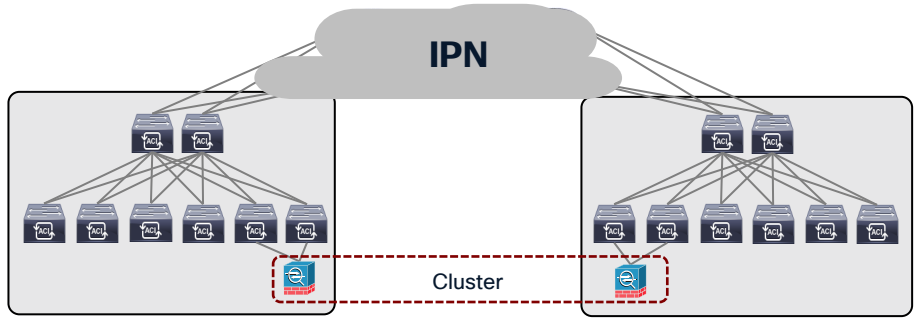
Design options

For Your Reference

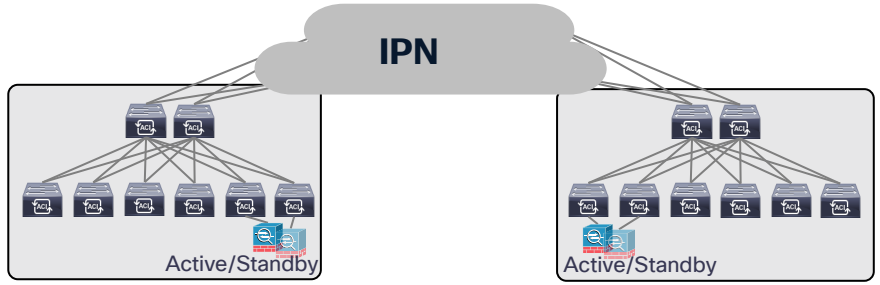
Ideal for Multi-Pod



- 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

© 2025 Cisco and/or its affiliates. All rights reserved.

cisco

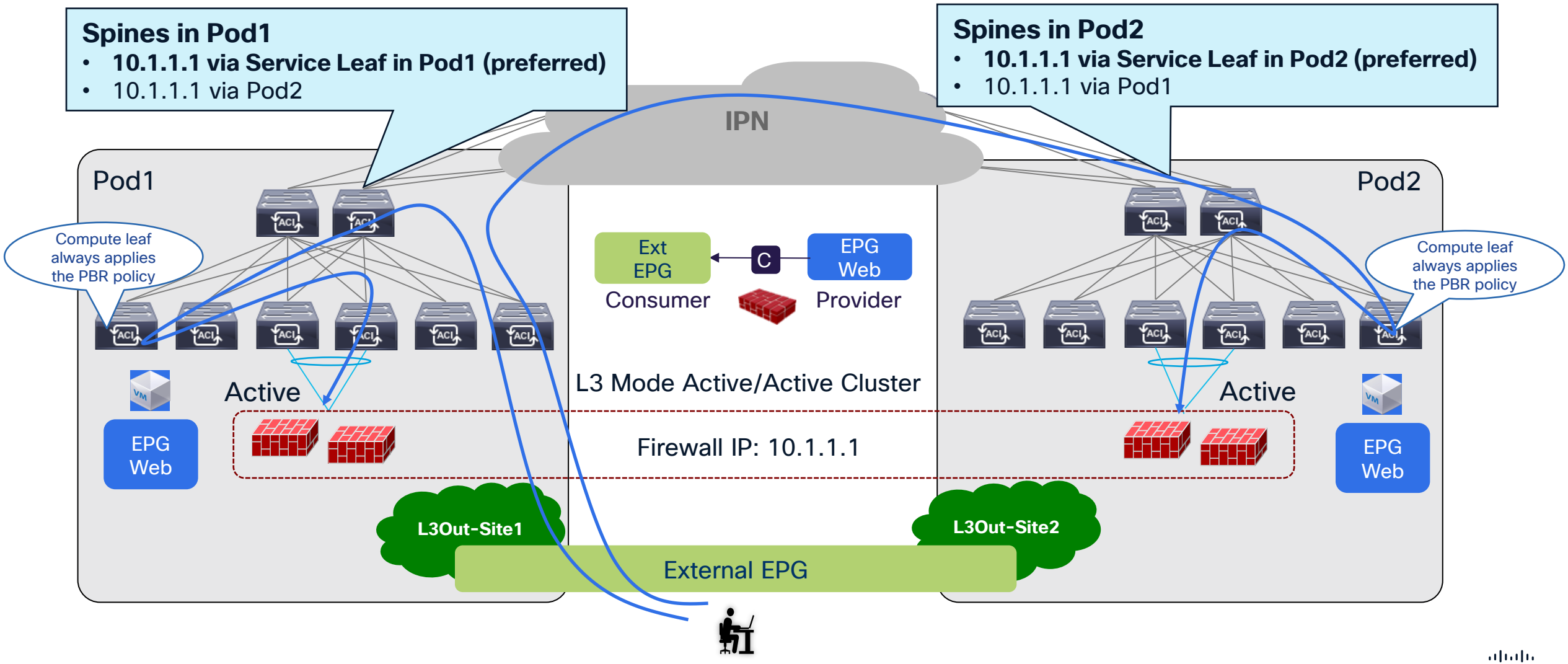
What is the Single Most Important Design Consideration when Deploying a Firewall in a Network Design?

Ensuring that the firewall sees both directions of the same flow

© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Pod: Active/Active cluster across pods

North-South Traffic Flow



© 2025 Cisco and/or its affiliates. All rights reserved.

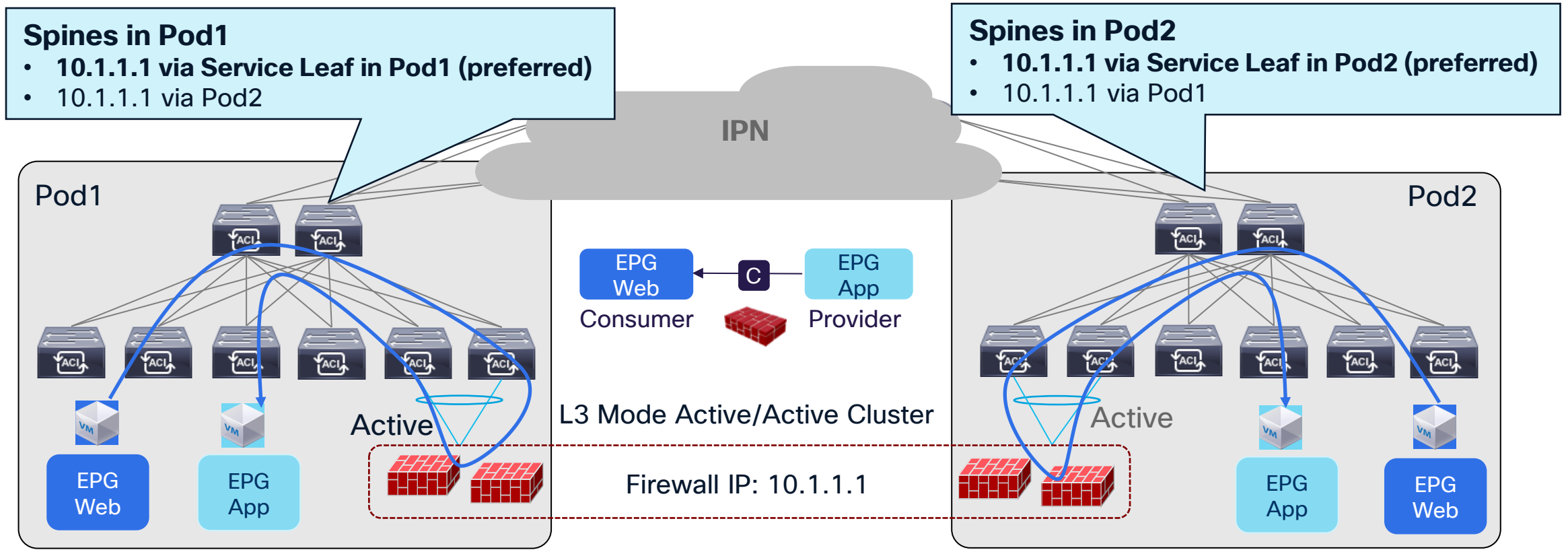
How was symmetricity ensured in this design

By doing PBR on the Compute Leaf

© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Pod: Active/Active cluster across pods

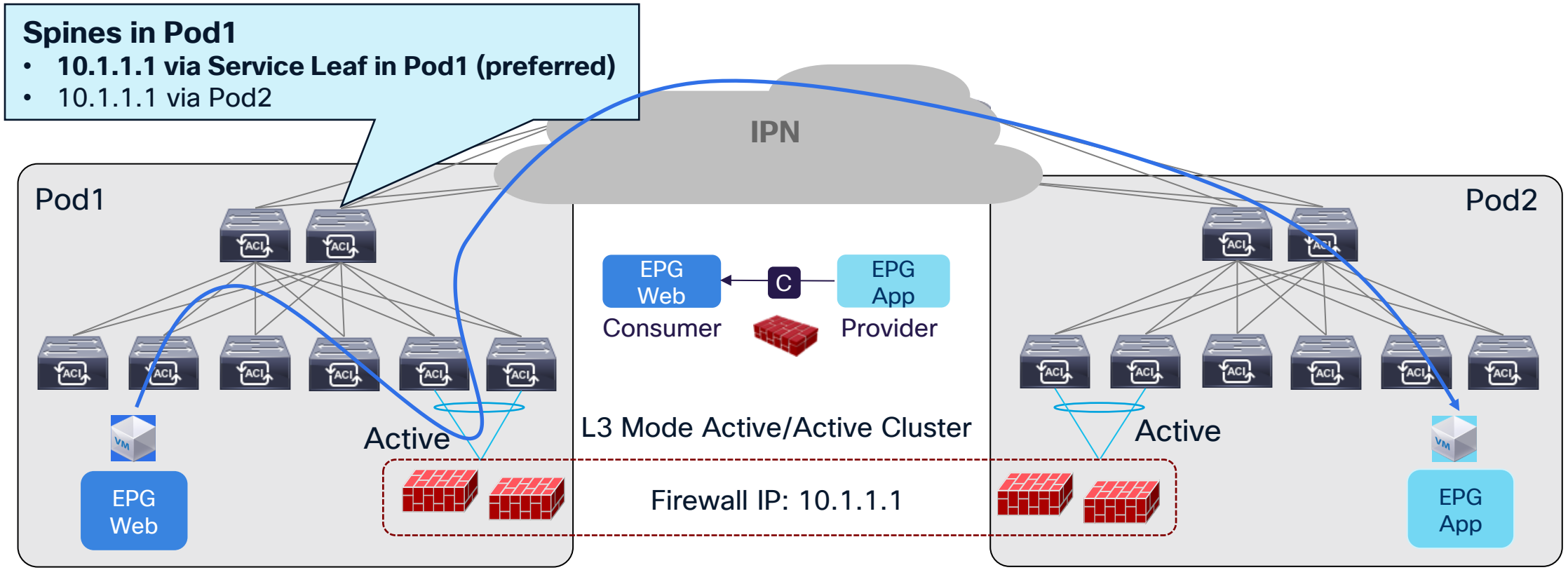
East-West Traffic Flow INTRA-POD



© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Pod: Active/Active cluster across pods

East-West Traffic Flow **INTERPOD**

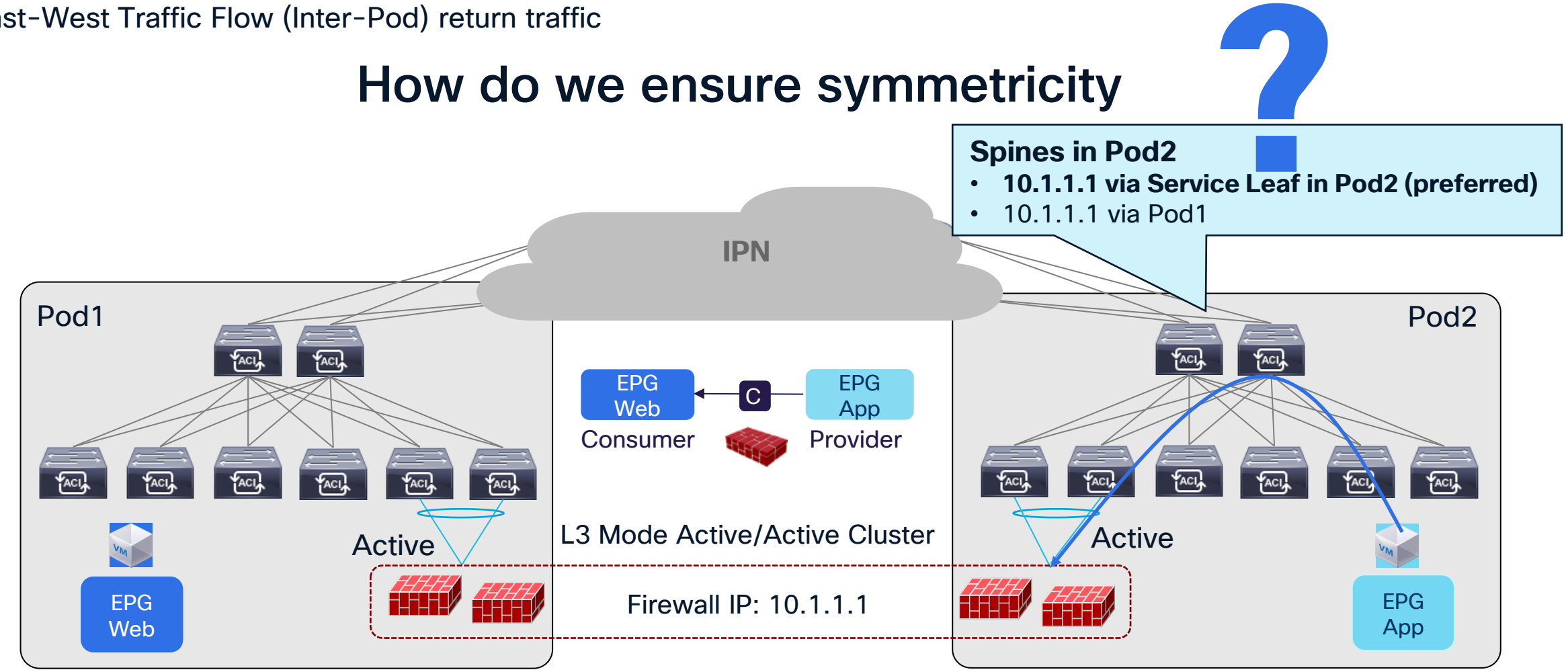


© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Pod: Active/Active cluster across pods

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

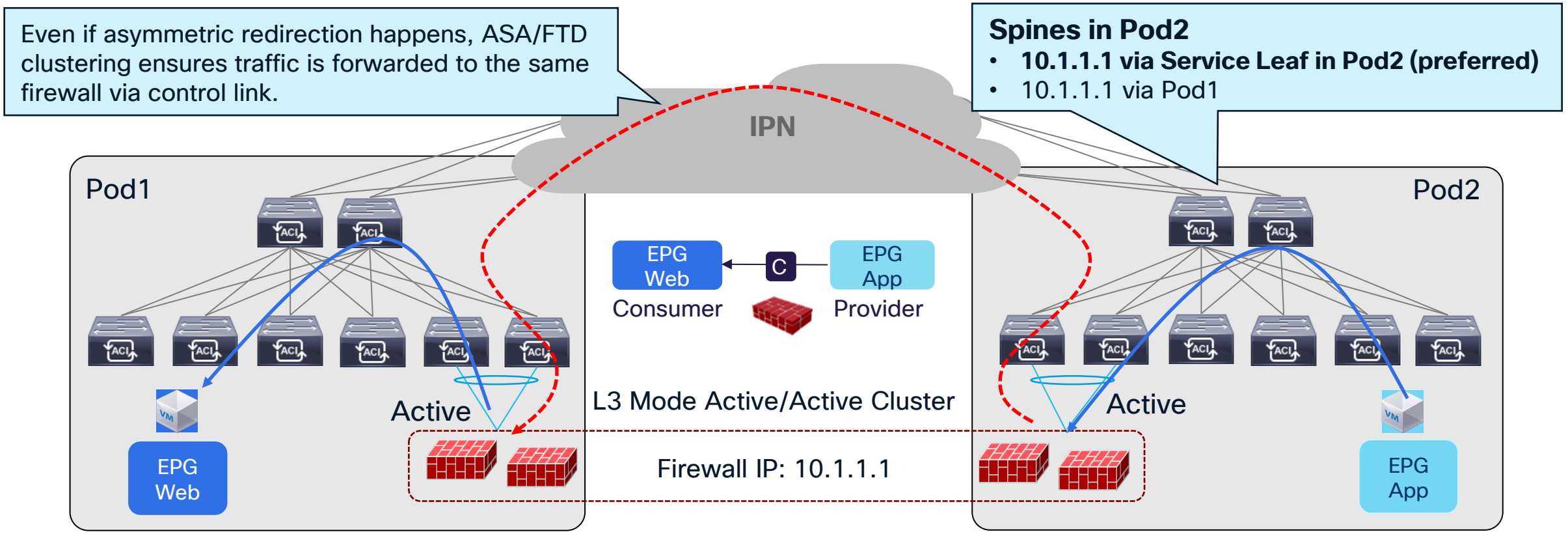
How do we ensure symmetricity



© 2025 Cisco and/or its affiliates. All rights reserved.

ACI Multi-Pod: Active/Active cluster across pods

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



© 2025 Cisco and/or its affiliates. All rights reserved.

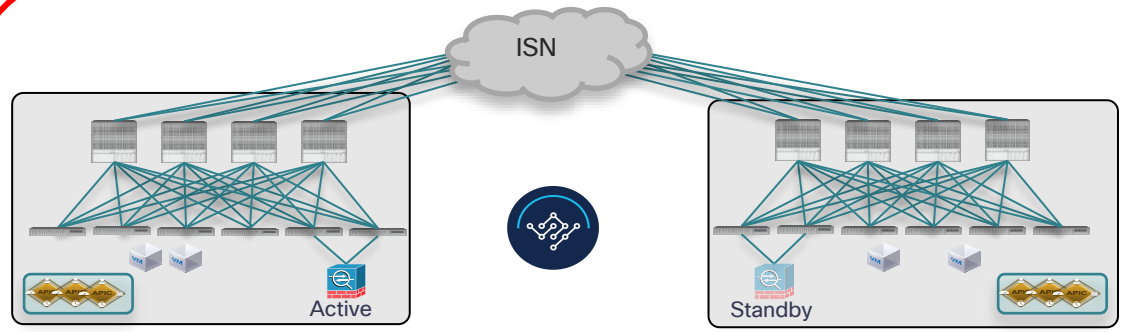
Network Services Integration with Multi-Site

© 2025 Cisco and/or its affiliates. All rights reserved.

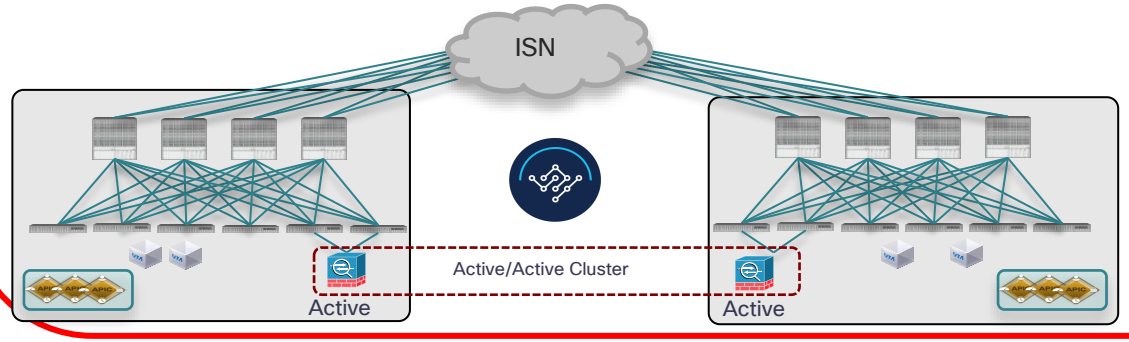
ACI Multi-Site and Network Services

Deployment options fully supported with ACI Multi-Pod

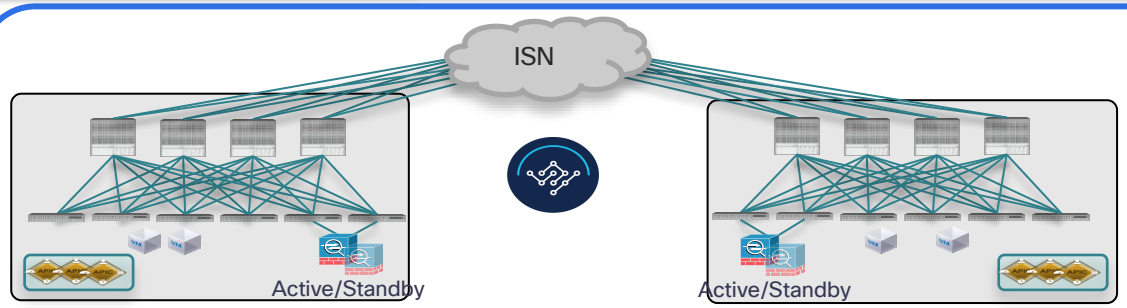
Integration Models



- Active and Standby pair deployed across Pods
- Limited supported options



- Active/Active FW cluster nodes stretched across Sites (single logical FW)
- Limited supported options



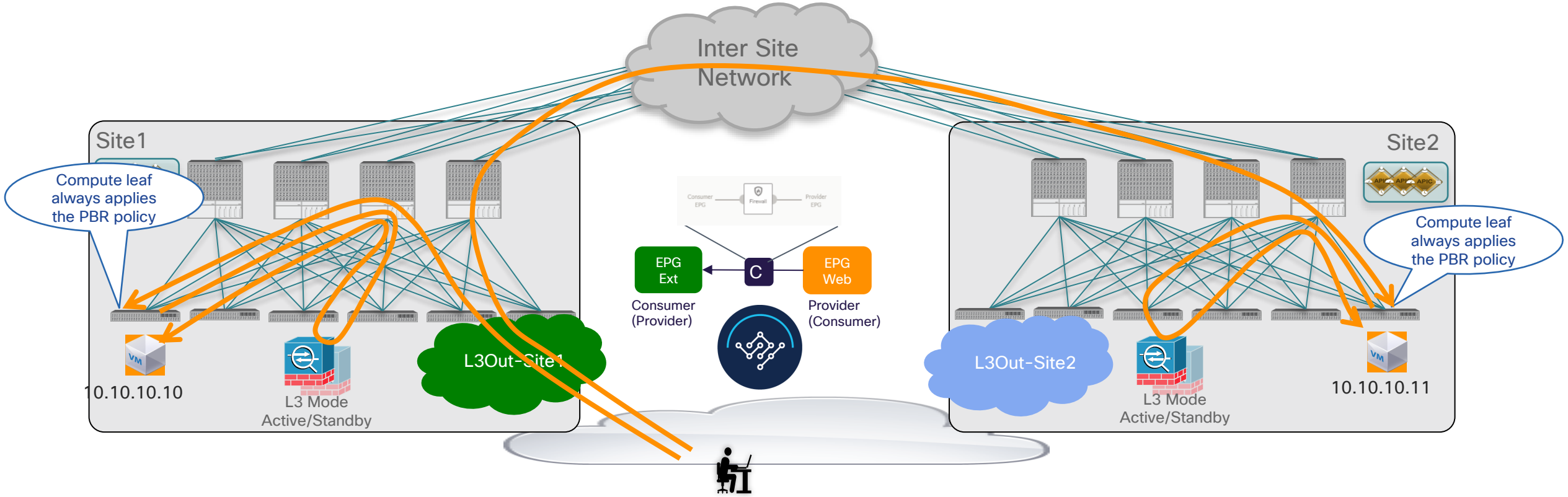
- Typical deployment model for ACI Multi-Site, each fabric leverages a dedicated service node function
- Use of PBR to avoid creating asymmetric paths through stateful devices (FWs, LBs, etc.) for both North-South and East-West communication

© 2025 Cisco and/or its affiliates. All rights reserved.

cisco

North-South Communication

Inbound Traffic

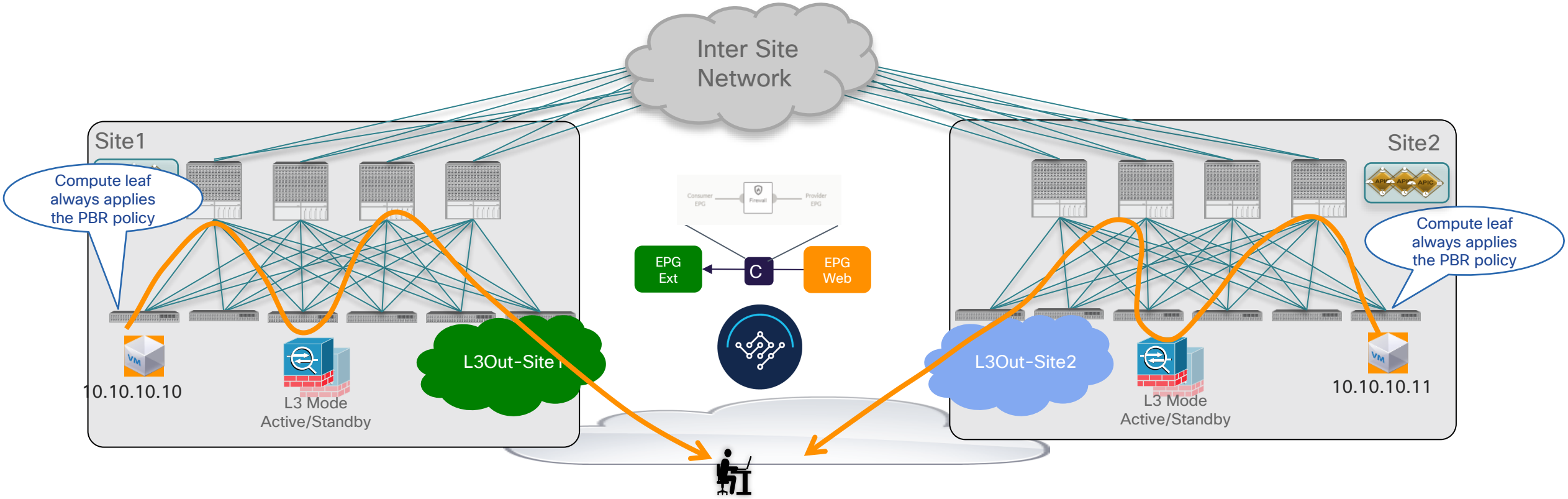


- Inbound traffic can enter any site when destined to a stretched subnet (if ingress optimization is not deployed or possible)
- HOW IS SYMMETRICITY ENSURED?
- PBR policy is always applied on the compute leaf node where the destination endpoint is connected
  - Requires the VRF to have the default policies for enforcement preference and direction
  - Ext-EPG and Web EPG can indifferently be provider or consumer of the contract

© 2025 Cisco and/or its affiliates. All rights reserved.

North-South Communication

Outbound Traffic

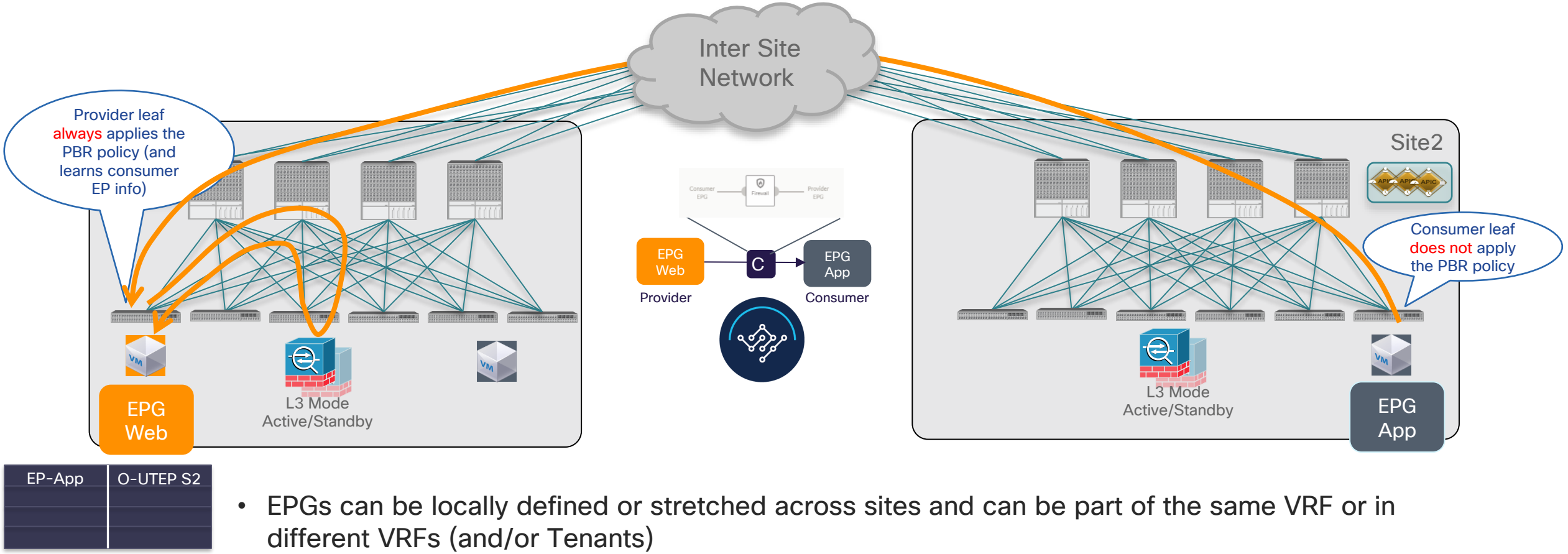


- PBR policy always applied on the same compute leaf where it was applied for inbound traffic
- Ensures the same service node is selected for both legs of the flow
- Different L3Outs can be used for inbound and outbound directions of the same flow

© 2025 Cisco and/or its affiliates. All rights reserved.

East-West Communication

Consumer to Provider Flow



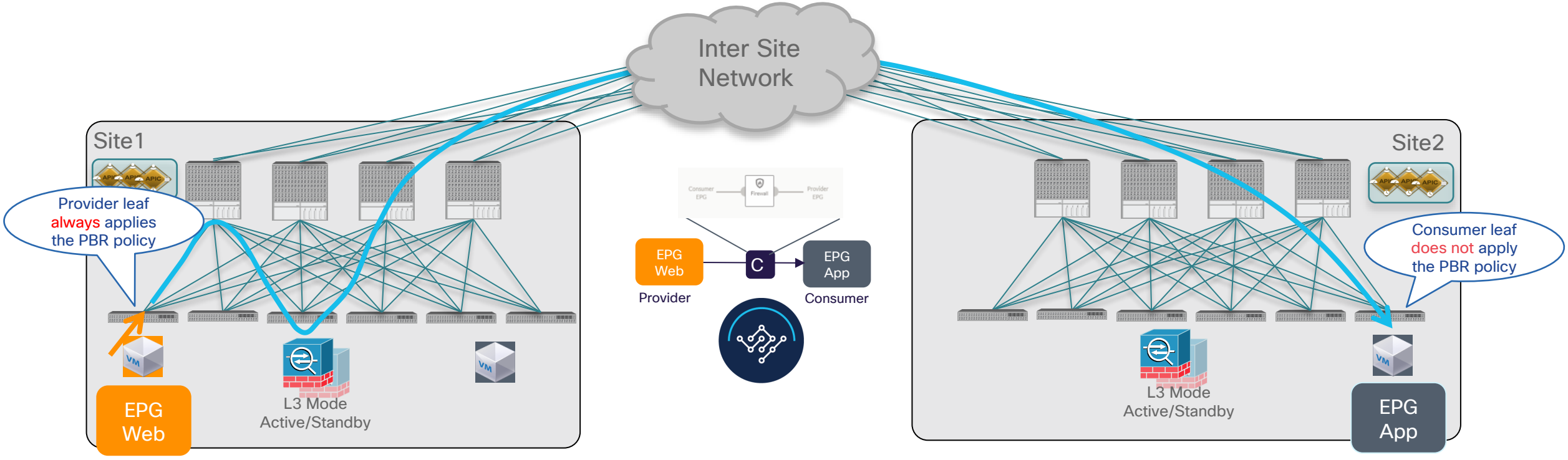
EP-App	O-UTEP S2

- EPGs can be locally defined or stretched across sites and can be part of the same VRF or in different VRFs (and/or Tenants)
- PBR policy is always applied only on the leaf switch where the **Provider** endpoint is connected
  - The Provider leaf always redirects traffic to a local service node

© 2025 Cisco and/or its affiliates. All rights reserved.

East-West Communication

Provider to Consumer Return Flow



- EPGs can be locally defined or stretched across sites and can be part of the same VRF or in different VRFs (and/or Tenants)
- PBR policy is always applied only on the leaf switch where the **Provider** endpoint is connected
  - The Provider leaf always redirects traffic to a local service node

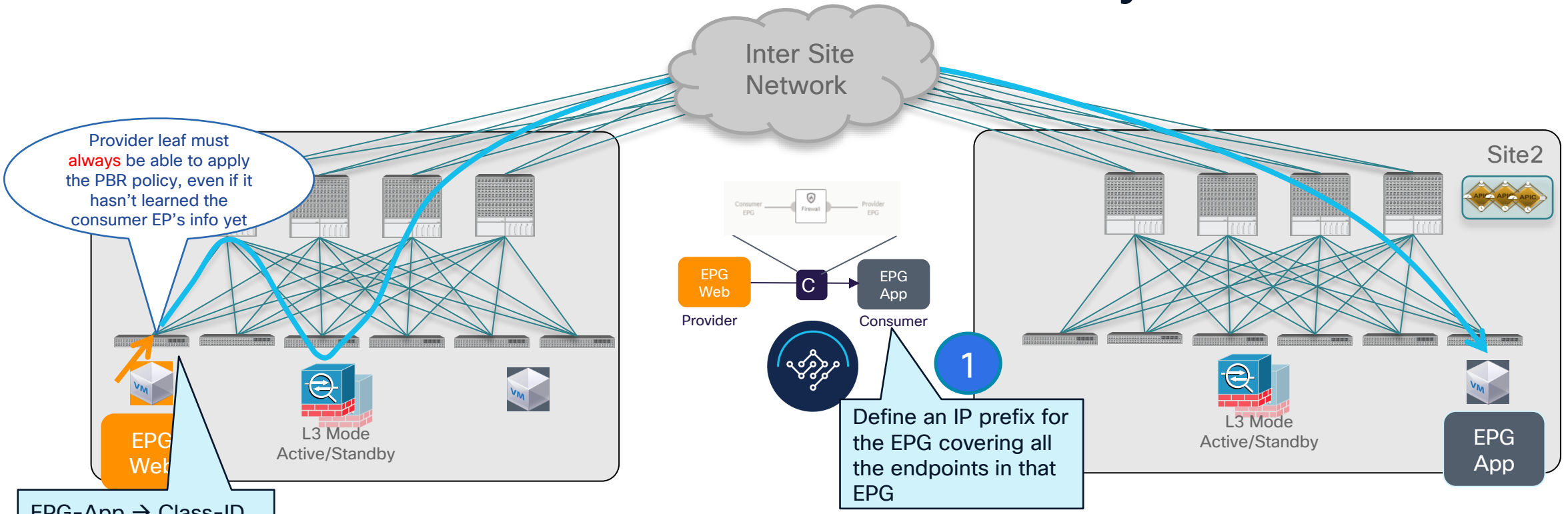
© 2025 Cisco and/or its affiliates. All rights reserved.

How is Symmetricity Ensured for East-to-West Traffic with Multi-Site

By doing PBR on the Provider Leaf

© 2025 Cisco and/or its affiliates. All rights reserved.

East-West Communication: What if the Communication is Initiated by the Provider?



- The Provider leaf must always apply the PBR policy, even if it hasn't learned the EP endpoint yet
- **Mandates to specify the IP prefix under the consumer EPG covering all the endpoints part of that EPG (this configuration is enforced on NDO)**
- **Becomes challenging when multiple EPGs are part of the same BD ("application centric" deployment model), use of /32 prefixes possible from ACI release 6.0(3)F**

© 2025 Cisco and/or its affiliates. All rights reserved.

Can we have service graph without having to enter Subnets under the EPG (i.e. can we have Application Centric designs AND Service Graph)

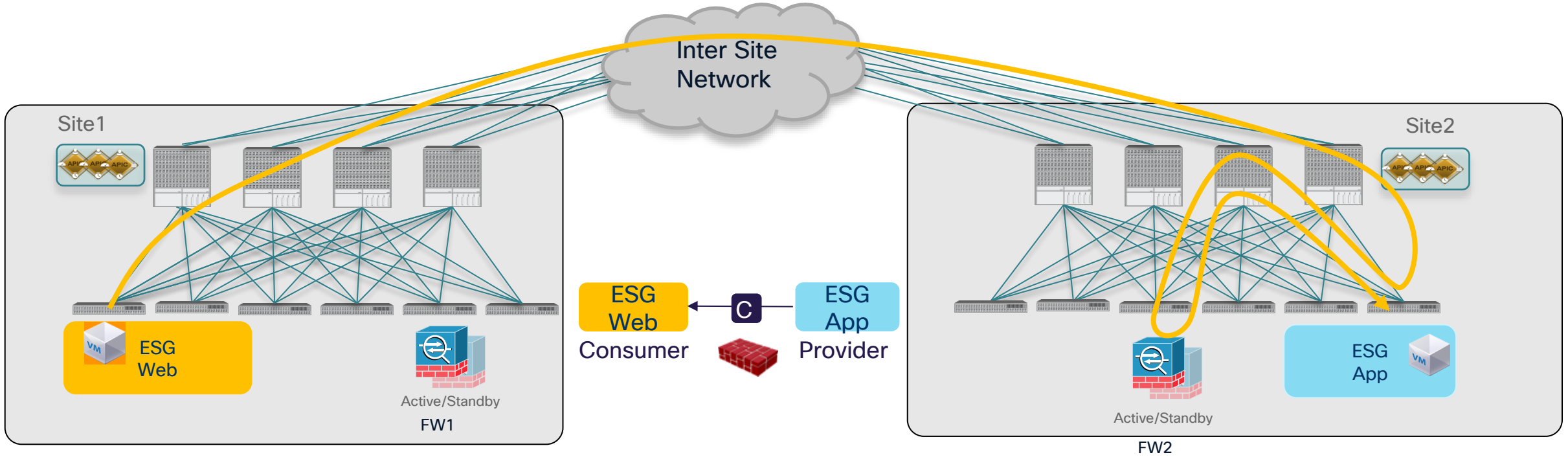
Yes with ESGs and with vzAny

© 2025 Cisco and/or its affiliates. All rights reserved.

Network Services Integration with Multi-Site with ESGs

© 2025 Cisco and/or its affiliates. All rights reserved.

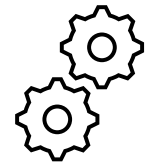
How to Keep Traffic Symmetric with ESGs Enforcement on the provider side only just like EPGs



© 2025 Cisco and/or its affiliates. All rights reserved.

ESG-to-ESG

Under the Hood



- **How to avoid asymmetric traffic paths through different FW nodes?**
  - Redirecting “inter-site” traffic to the Firewall services only on the provider leaf
  - Only redirection to the local Firewall service for intra-fabric flows

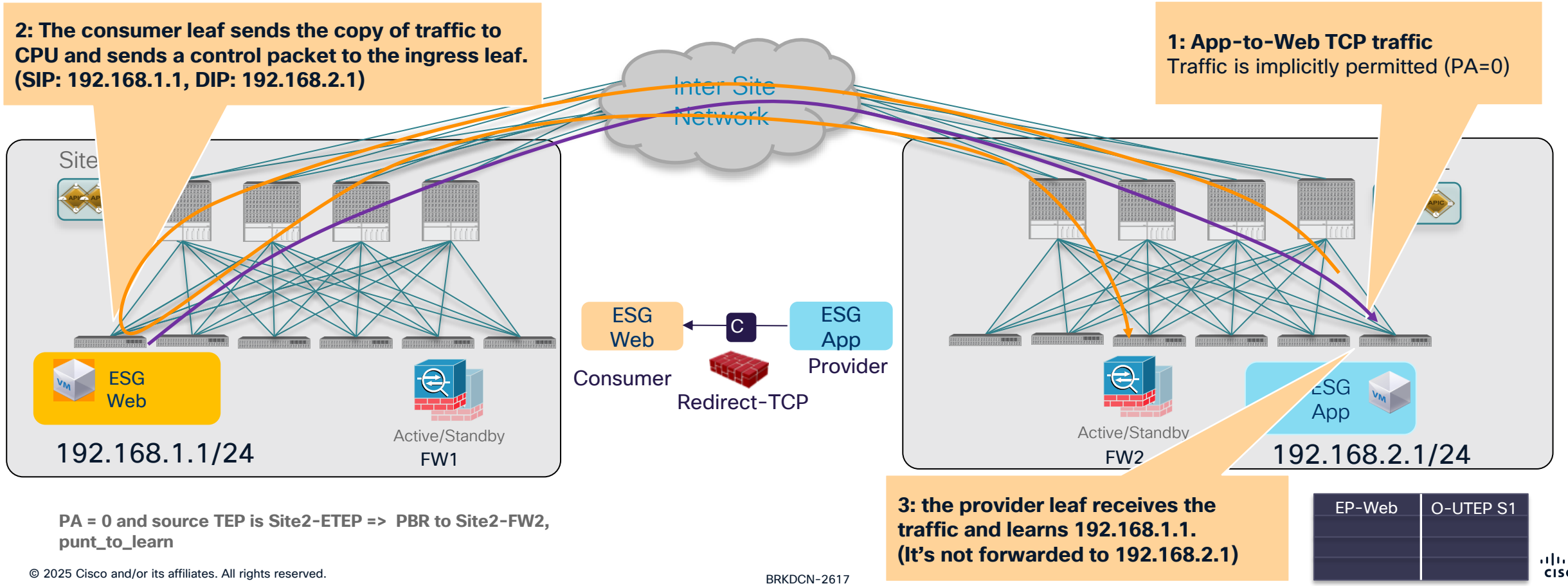
- **How to ensure the ingress leaf nodes can resolve the destination class-ID without requiring any static IP prefix configuration under the ESG/EPG?**

© 2025 Cisco and/or its affiliates. All rights reserved.

cisco

Conversational Learning

- For the Consumer to Provider direction, with enforcement on the provider there is no challenge
- What if the connection is initiated on the provider EPG and the provider leaf doesn't know the consumer endpoint?
- Conversational Learning to get the ingress leaf learn the destination EP.

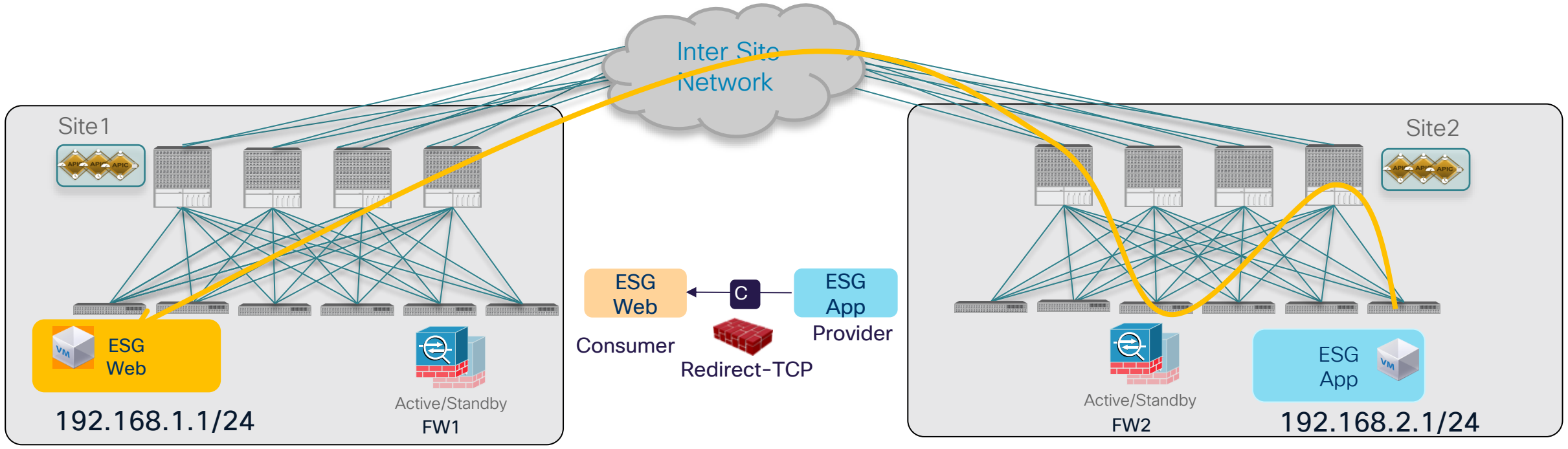


PA = 0 and source TEP is Site2-ETEP => PBR to Site2-FW2, punt_to_learn

© 2025 Cisco and/or its affiliates. All rights reserved.

Subsequent Packets from Provider to Consumer

- With conversational learning the ingress leaf learns the destination EP class-id
- PBR to the local firewall is performed on the provider leaf without any tromboning

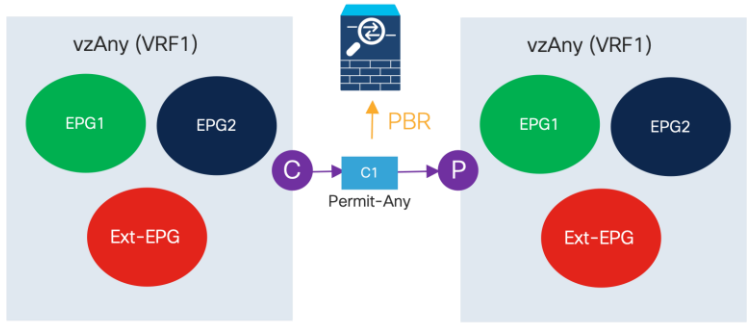


EP-Web	O-UTEP S1

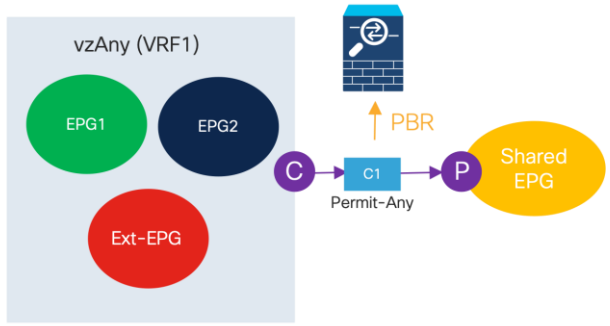
© 2025 Cisco and/or its affiliates. All rights reserved.

Application Centric Deployments now possible with: ESG-to-ESG, vzAny-to-vzAny, vzAny-to-EPG, L3Out-to-L3Out

Any-to-Any

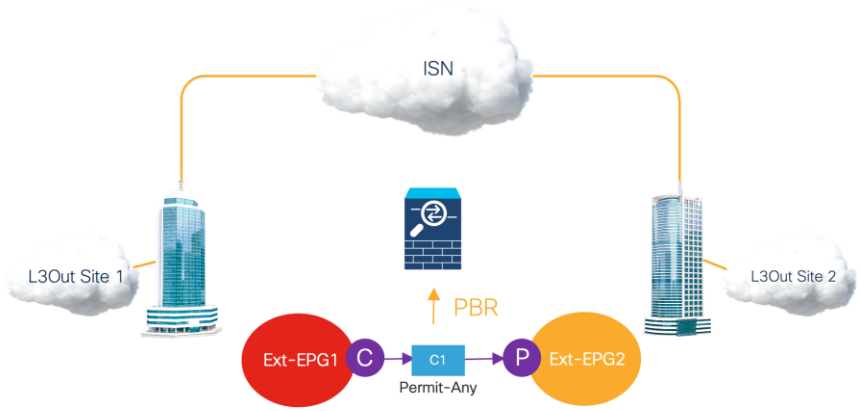


Many-to-One



- Two scenarios:
  1. vzAny-to-EPG
  2. vzAny-to-L3Out

Transit Intersite L3Out



For all these use cases above
(except L3OutEPG-to-L3OutEPG, where it is not needed)
the introduction of “conversational learning” ensures that the ingress leaf or the provider leaf can always resolve the destination class-ID without requiring any static IP prefix configuration under the consumer EPG (fully supporting “app centric” designs)

© 2025 Cisco and/or its affiliates. All rights reserved.

Note: L3Out means external EPG in this table

Starting from ACI release 6.1(4)

What's new/changed

- Starting from ACI release 6.1(4), the following PBR deployments are supported in Multi-Site PBR.

	EPG-to-EPG **ESG-to-ESG**	**EPG-to-L3Out, L3Out-to-EPG** **ESG-to-L3Out, L3Out-to-ESG**	**vzAny-to-vzAny**	**vzAny-to-EPG** **vzAny-to-ESG**	**vzAny-to-L3Out**	**L3Out-to-L3Out**
Redirection	Site of the provider EPG	Site of the EPG	Both sites	Site of the provider EPG	Both sites	Both sites
Service node	One-node/Two-node One-arm/Two-arm	One-node/Two-node One-arm/Two-arm	One-node One-arm	One-node/Two-node One-arm/Two-arm	One-node One-arm	One-node One-arm
VRF	Intra-VRF Inter-VRF	Intra-VRF Inter-VRF ~~(L3Out must be the provider)~~	Intra-VRF	Intra-VRF Inter-VRF	Intra-VRF	Intra-VRF and Inter-VRF

© 2025 Cisco and/or its affiliates. All rights reserved.

Useful Links

Where to Go for More Information

- ✓ 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

- ✓ ACI Multi-Site White Paper

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

- ✓ Cisco Multi-Site Deployment Guide for ACI Fabrics

  https://www.cisco.com/c/en/us/td/docs/dcn/whitepapers/cisco-multi-site-deployment-guide-for-aci-fabrics.html

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

  https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-743107.html

- ✓ ACI Multi-Site Training Sessions

  https://www.cisco.com/c/en/us/solutions/data-center/learning.html#~nexus-dashboard

© 2025 Cisco and/or its affiliates. All rights reserved.

Related Sessions

- BRKDCN-2909 Automating Micro-Segmentation and Deploying Layer 4-7 Services in VXLAN EVPN Fabrics using Group Policy Option (GPO) and Nexus Dashboard
- BRKDCN-2602 Nexus One: An architecture for your Cisco Datacenter of tomorrow
- BRKDCN-2959 Turbo charge your ACI Segmentation and Visibility
- BRKDCN-2663 Kubernetes (K8s) Infrastructure Connectivity for the Modern Data Center

© 2025 Cisco and/or its affiliates. All rights reserved.

cisco

Complete your session evaluations

Complete a minimum of 4 session surveys and the Overall Event Survey to claim a Cisco Live T-Shirt.

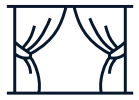
Earn up to 800 points by completing all surveys and climb the Cisco Live Challenge leaderboard.

Level up and earn exclusive prizes!

Complete your surveys in the Cisco Live Events app.

© 2025 Cisco and/or its affiliates. All rights reserved.

Continue your education



Visit the Cisco Stand for related demos

Book your one-on-one Meet the Expert meeting

Attend the interactive education with Capture the Flag, and Walk-in Labs

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

© 2025 Cisco and/or its affiliates. All rights reserved.

Thank you

© 2025 Cisco and/or its affiliates. All rights reserved.

CISCO