

CISCO *Live!*

GO BEYOND

#CiscoLiveAPJC



Cisco ACI Design and Deployment

ACI Multi-Pod and Multi-Site

John Weston

Technical Marketing Engineer, Data Center Networking

BRKDCN-2617

CISCO *Live!*

#CiscoLiveAPJC



Cisco Webex App

Questions?

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

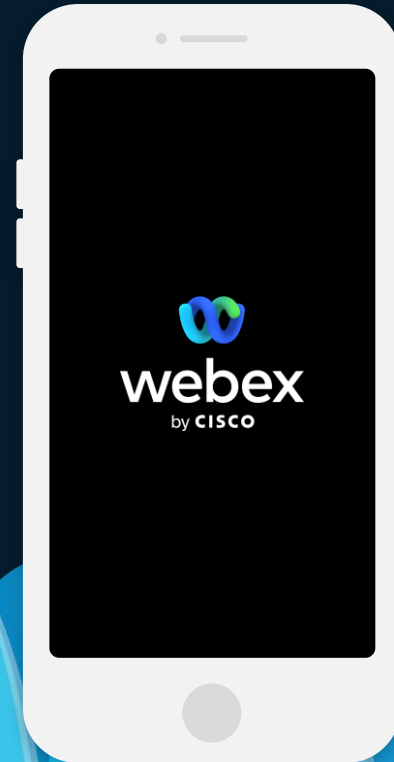
How

- 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 November 15, 2024.



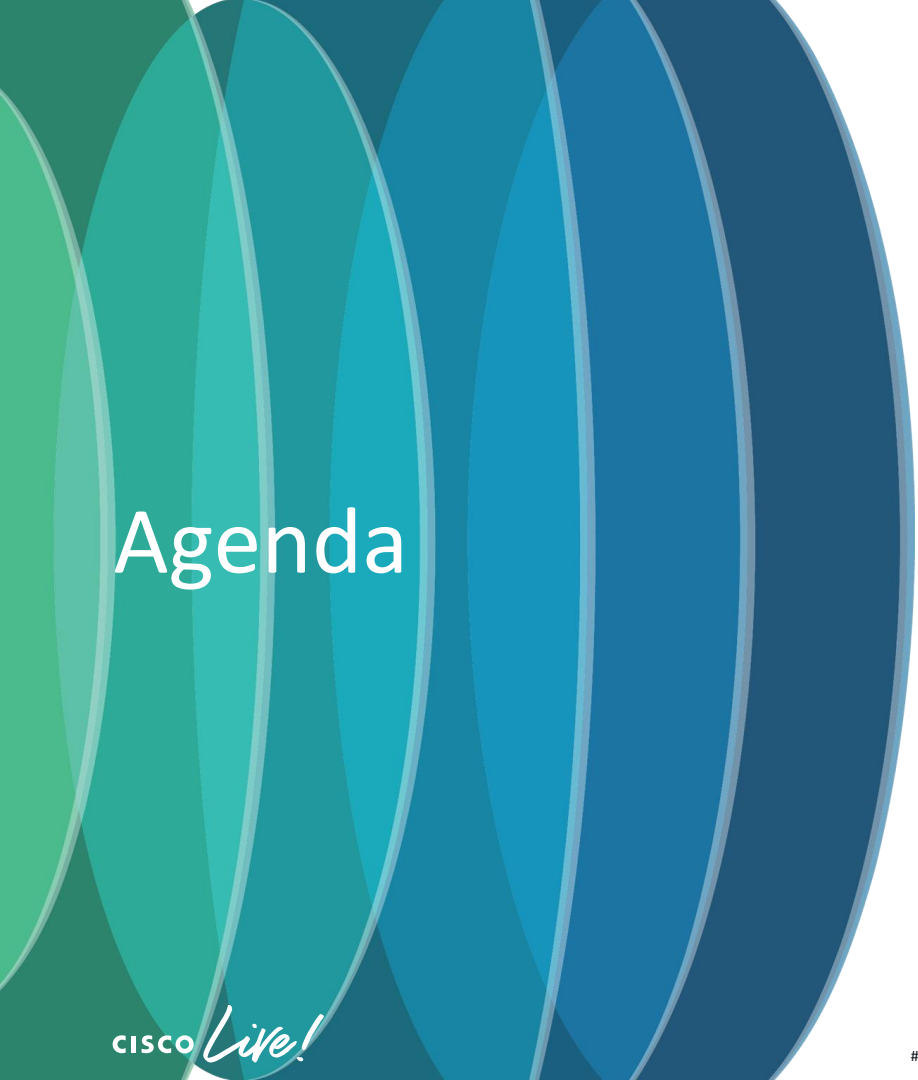
<https://ciscolive.ciscoevents.com/ciscolivebot/#BRKDCN-2617>



Session Objectives



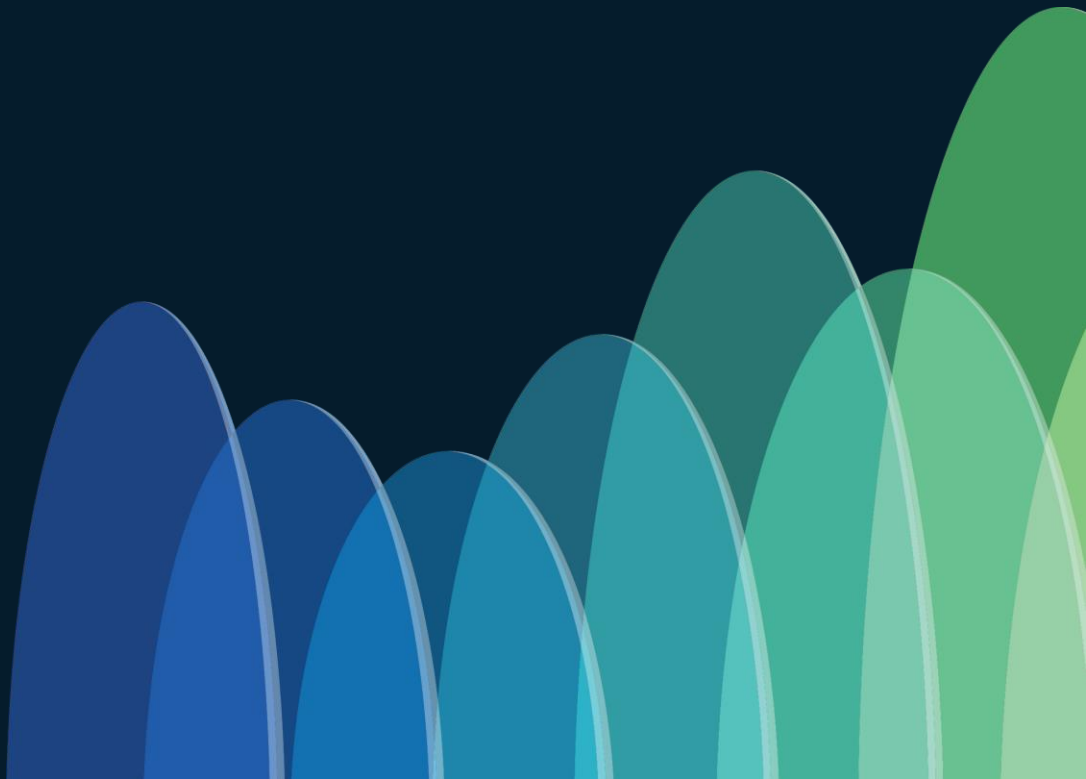
- At the end of the session, the participants should be able to:
 - Articulate the different deployment options to interconnect Cisco ACI networks (Multi-Pod and Multi-Site) and when to choose one vs. the other
 - Understand the functionalities and specific design considerations associated to the ACI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



Agenda

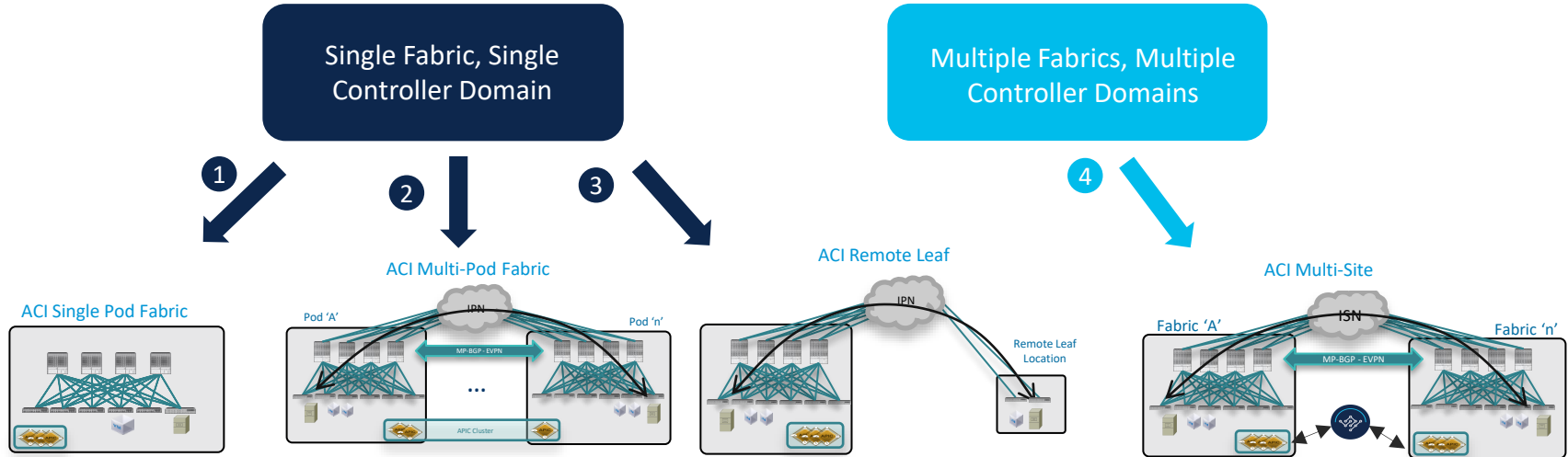
- Introduction
- IPN/ISN Requirements
- Multi-Pod APIC Cluster Deployment Considerations
- Control and Data Plane
- Nexus Dashboard Orchestrator (NDO)
- Provisioning Policies on NDO
- Connecting to the External L3 Domain
- Network Services Integration

Introduction



ACI Architectural Options

Fabric and Policy Domain Evolution



Multi-Pod or Multi-Site?

Where to Get More Information

- ACI Multi-Pod White Paper

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

- ACI Multi-Site White Paper

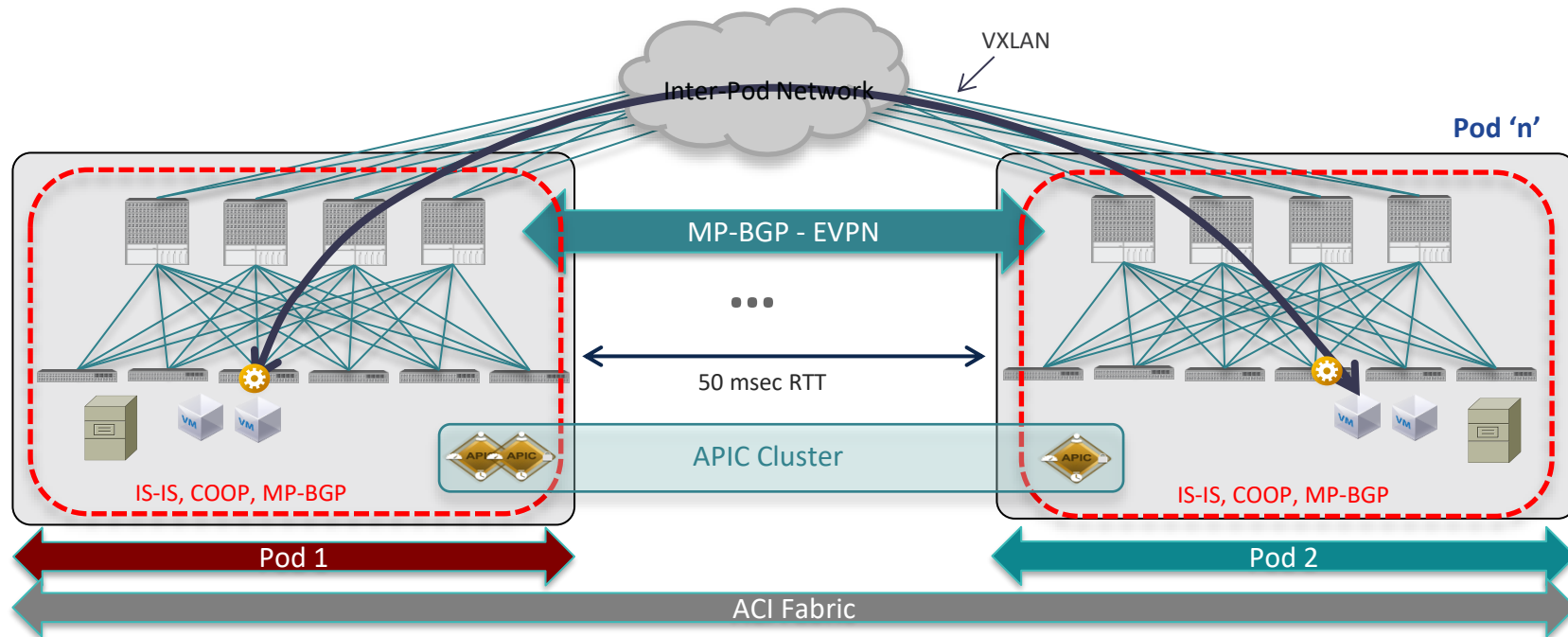
<https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-739609.html>

- ACI Multi-Site Cisco Live 2020 Digital Breakout Session

<https://www.ciscolive.com/on-demand/on-demand-library.html?search=ardica&search=ardica#/video/1636411349156002rlx8>

ACI Multi-Pod

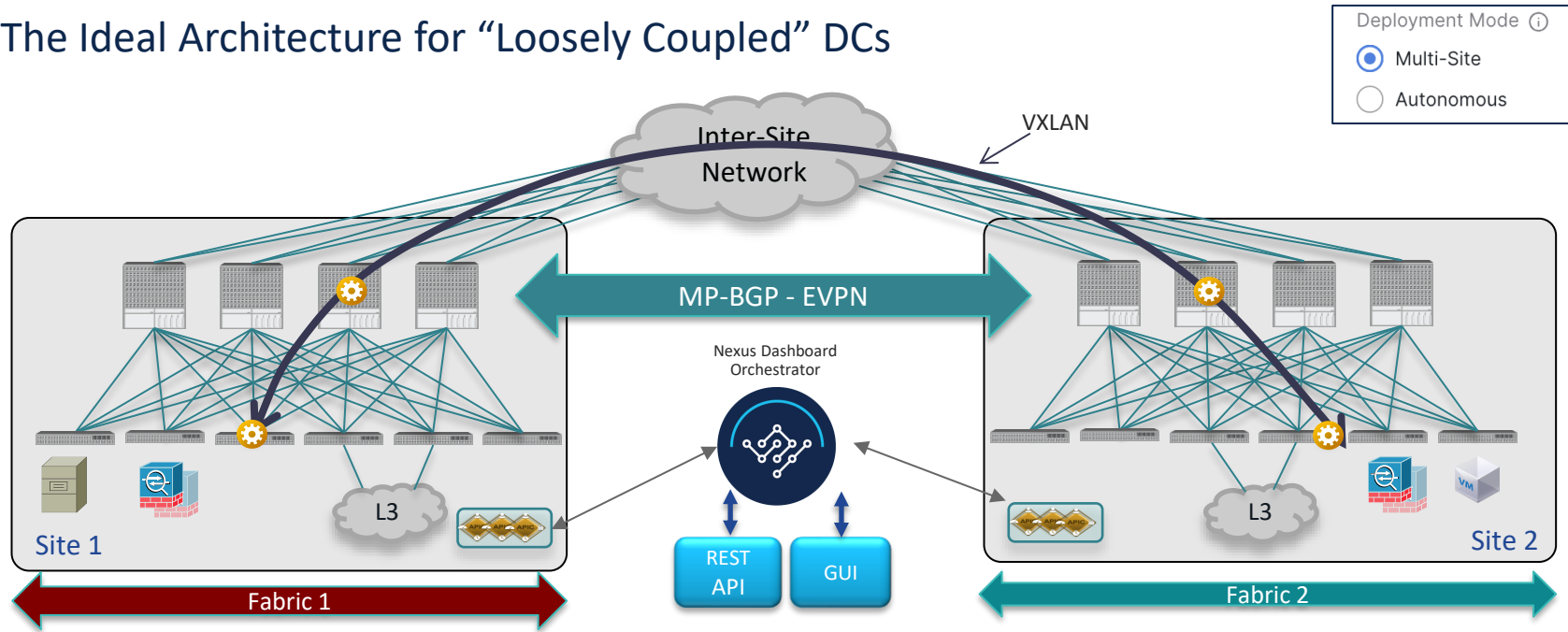
The Ideal Architecture for Active/Active DC Deployments



- 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-Site

The Ideal Architecture for “Loosely Coupled” DCs

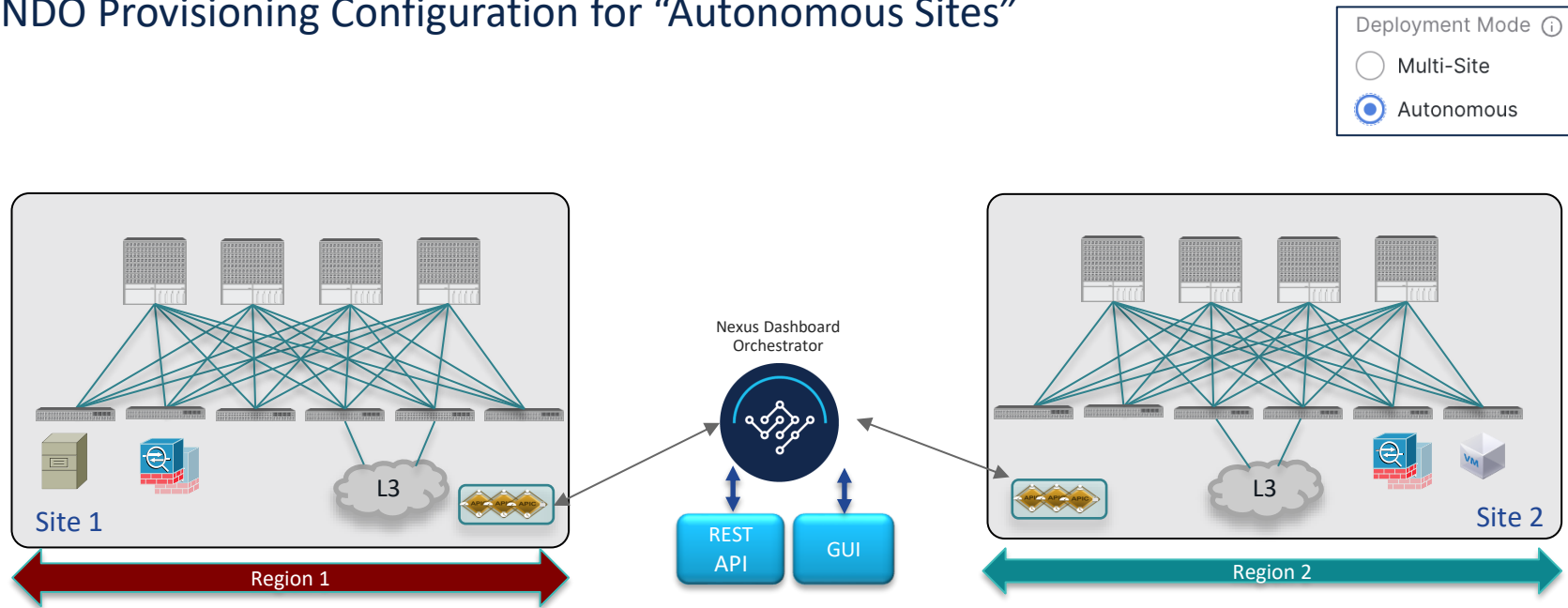


- Separate ACI Fabrics with independent APIC clusters
- No latency limitation between Fabrics
- Nexus Dashboard Orchestrator pushes cross-fabric configuration to multiple APIC clusters providing scoping of all configuration changes

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

ACI Multi-Site

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

ACI Multi-Site Architecture

Most Common Use Cases

- Compartmentation/Scale

Building Multiple Fabrics inside a single Data Center

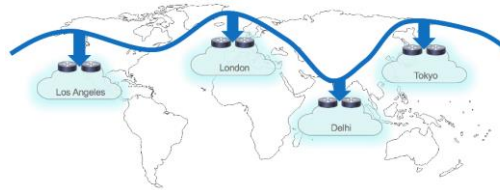


Optimised and controlled L2/L3 connectivity (including optimised/controlled BUM forwarding), scale out total number of leaf nodes (SP use case)

- Data Center Interconnect (DCI)

Extend connectivity/policy between 'loosely coupled' DC sites

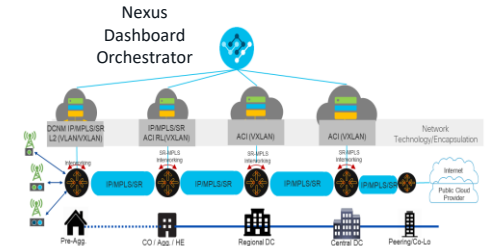
Disaster Recovery and IP mobility use cases



- SP 5G Telco DC/Cloud*

Centralised DC Orchestration for "Autonomous Fabrics"

Optional SR-MPLS/MPLS Handoff on Border Leaf nodes

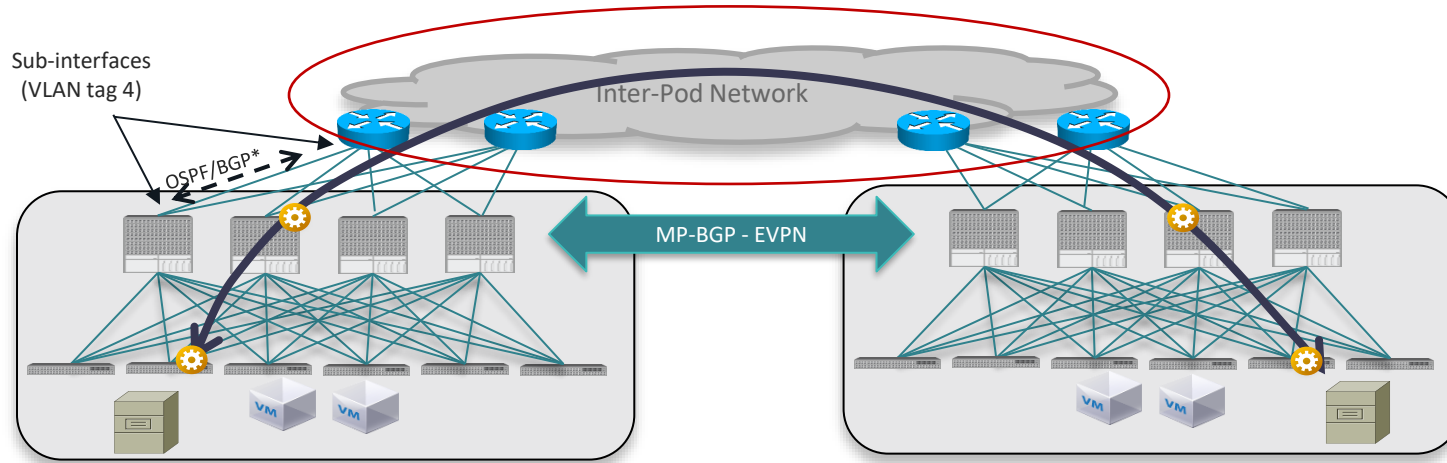


*May also apply to Enterprise deployments

IPN/ISN Requirements



Inter-Pod Network (IPN) Functional Requirements



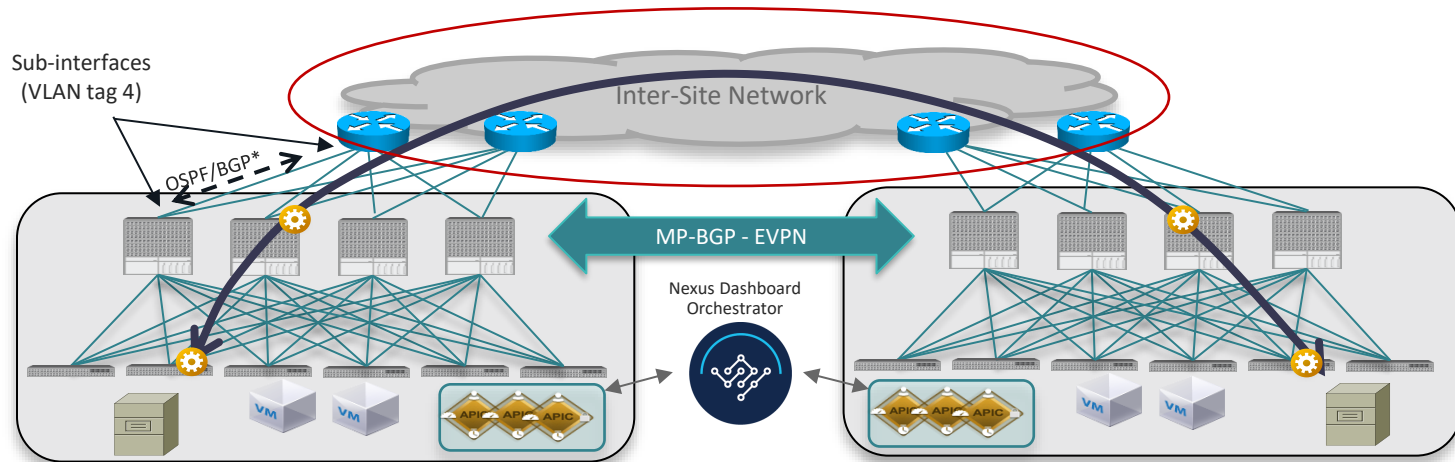
- Not managed by APIC, must be independently configured (day-0 configuration)
- IP topology can be arbitrary, not mandatory to connect all the spine nodes to the IPN
- IPN main functional requirements:
 - ✓ Multicast BiDir PIM → needed to handle BUM*/TRM** traffic
 - ✓ Increased end-to-end MTU support (at least 50/54 extra Bytes)
 - ✓ OSPF/BGP to peer with the spine nodes and exchange TEP address reachability
 - ✓ DHCP Relay

Must use sub-interfaces (with VLAN tag 4) toward the spines

*BUM: Broadcast, Unknown Unicast, Multicast

**TRM: Tenant Routed Multicast

Inter-Site Network (ISN) Functional Requirements



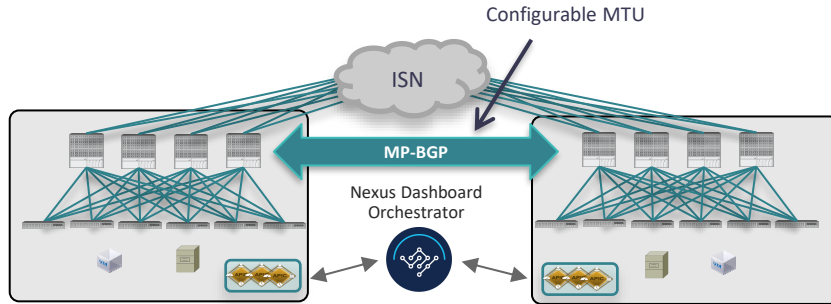
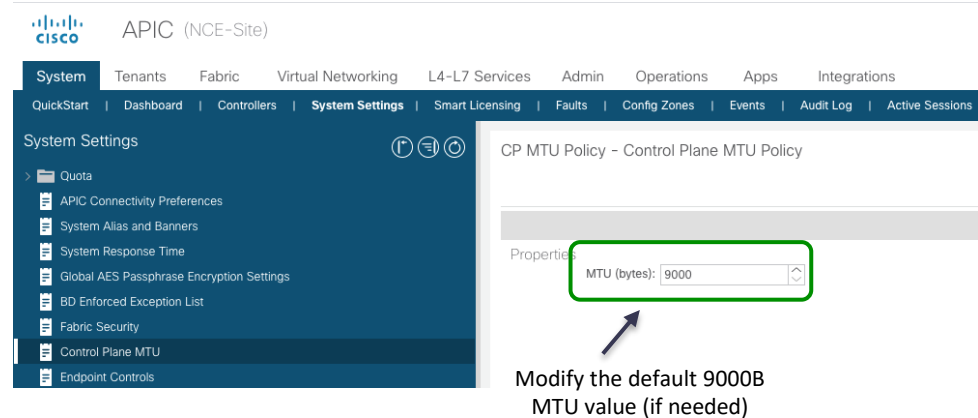
- Not managed by APIC or NDO, must be independently configured (day-0 configuration)
- IP topology can be arbitrary, not mandatory to connect all the spine nodes to the ISN
- ISN main functional requirements:
 - ✓ OSPF/BGP to peer with the spine nodes and exchange TEP address reachability
 - Must use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

What if the ISN Supports Only
1500B MTU Size?



ACI Multi-Site and MTU Size

Tuning MTU Size for EVPN Control-Plane Traffic

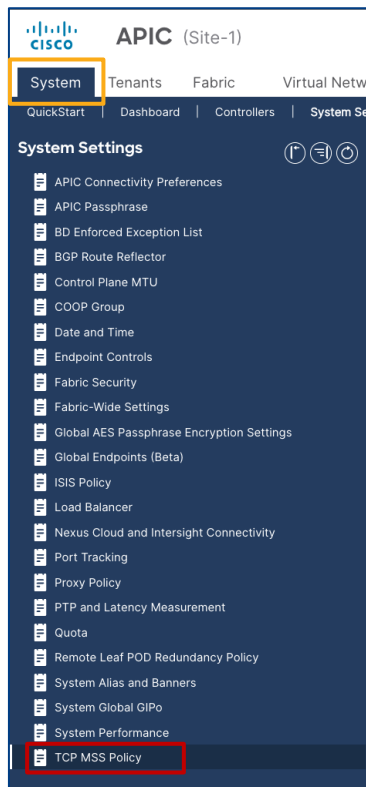


- Control-Plane MTU can be set leveraging the “Control Plane MTU Policy” on APIC
 - The setting applies to all the control-plane traffic generated by ACI leaf/spine nodes
- The required MTU in the ISN would hence depend on this setting and on the MTU of the traffic generated by endpoints/devices connected to the fabric
 - Always need to consider the VXLAN encapsulation overhead for data plane traffic

ACI Multi-Site and MTU Size

Introducing the TCP-MSS Adjust Functionality

ACI Release 6.0(3)F



TCP MSS Policy

Properties

Type: **Global** | RL and Msite | RL Only | Disable

IPv4: 8888

IPv6: 8868

Supported values are
688-9104 bytes

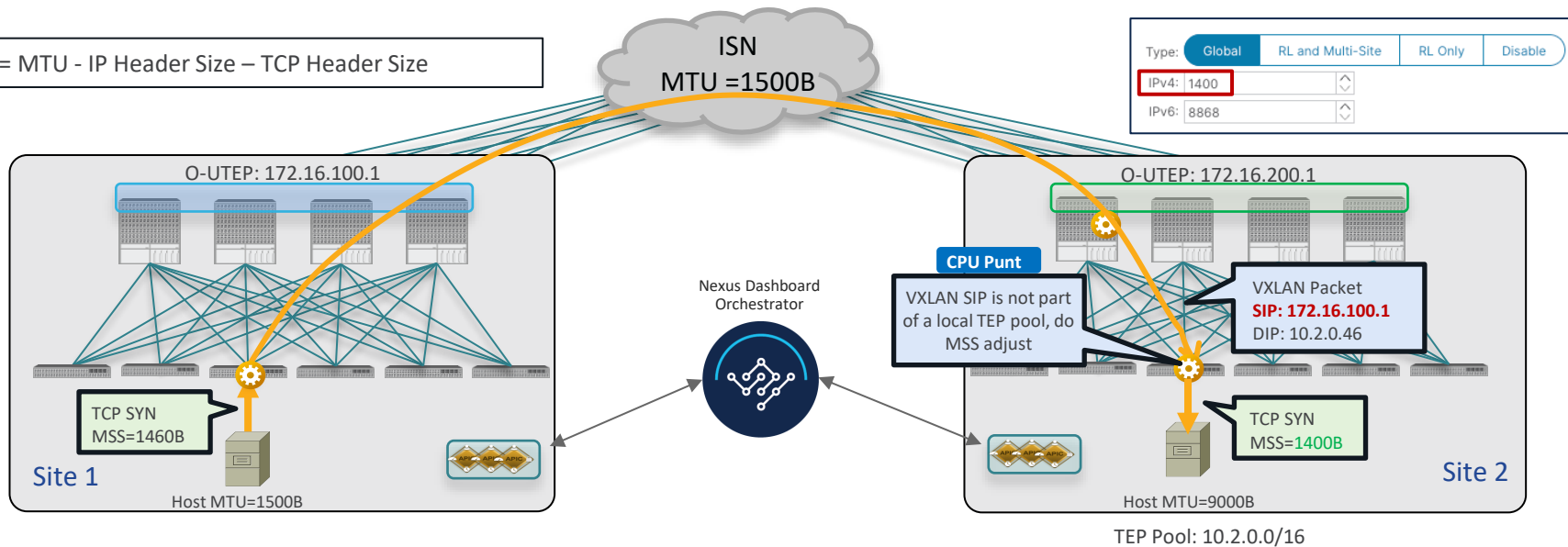
- TCP MSS adjust policy is enabled at System Settings level
- Supports different TCP MSS adjust setting for IPv4 and IPv6
- Supports three different options:
 1. **Global**: applies to all flows (Multi-Pod, Multi-Site, Remote Leafs)
 2. **RL and Msite**: applies to Multi-Site and Remote Leafs flows
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

TCP-MSS Adjust Functionality

SYN Packet

ACI Release 6.0(3)F

$MSS = MTU - IP \text{ Header Size} - TCP \text{ Header Size}$

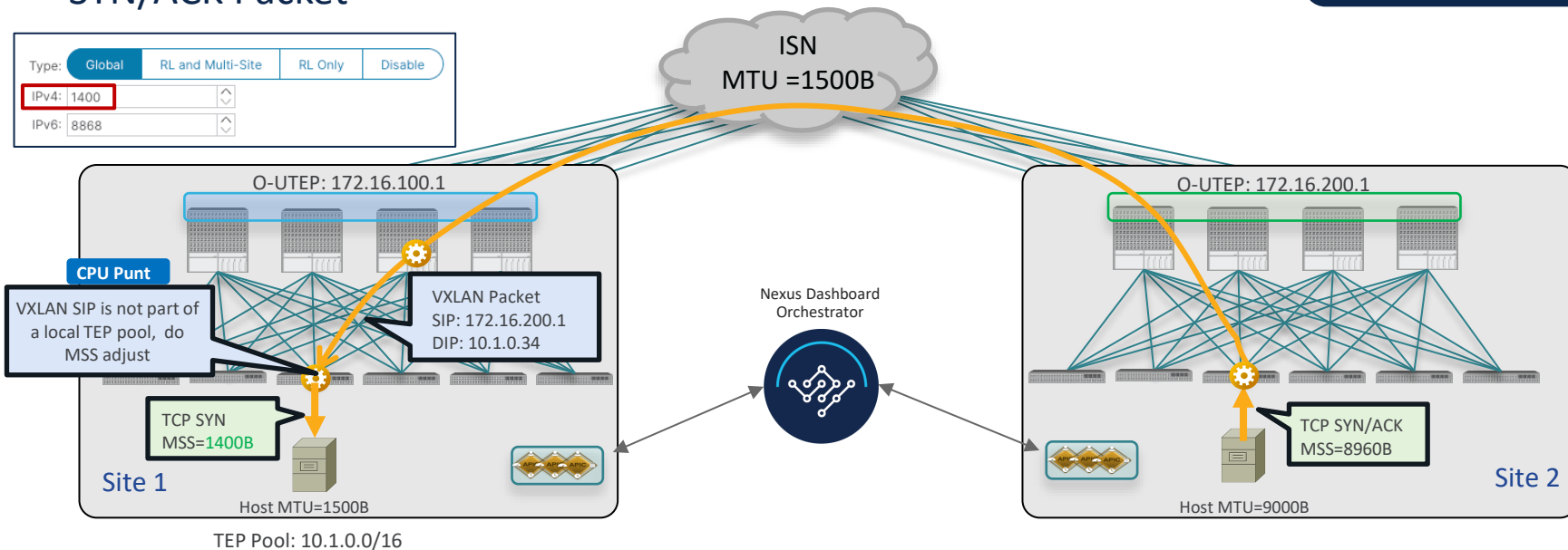


- TCP MSS adjust is always performed on the egress leaf node
- Adjusts TCP MSS value on SYN and SYN/ACK packets
- Checks for Source IP in the VXLAN header → TCP-MSS adjusts performed if the source IP is not part of the fabric's internal TEP pool

TCP-MSS Adjust Functionality

SYN/ACK Packet

ACI Release 6.0(3)F

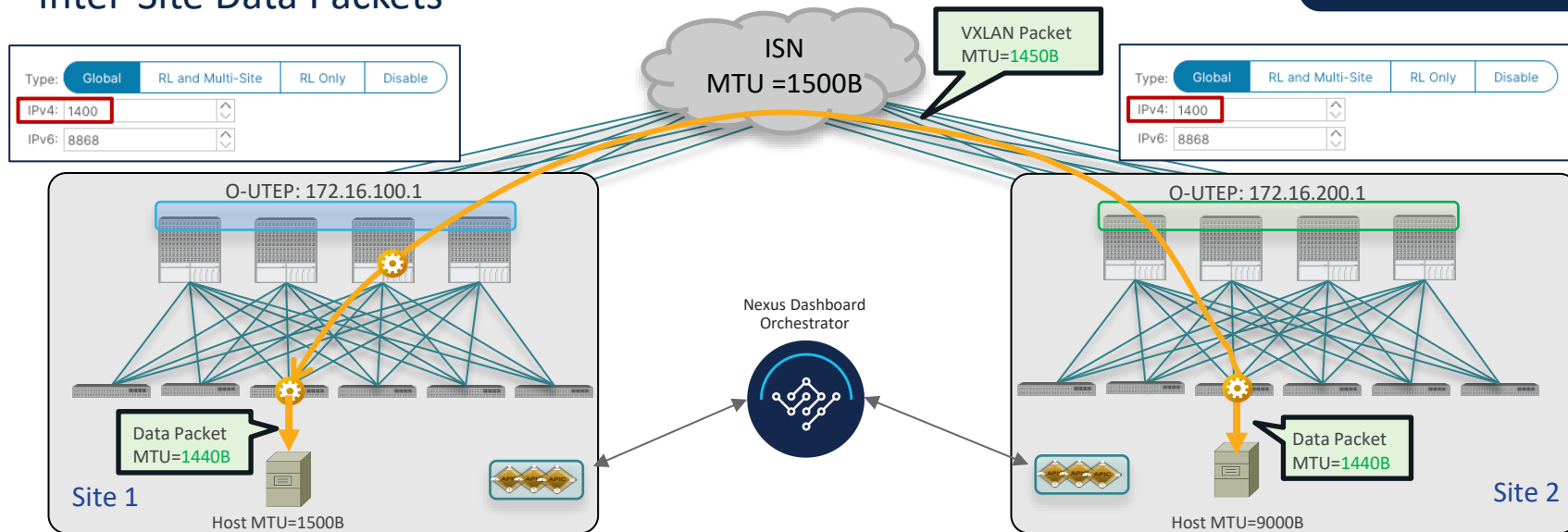


- TCP MSS adjust is always performed on the egress leaf node
- Adjusts TCP MSS value on SYN and SYN/ACK packets
- Checks for Source IP in the VXLAN header → TCP-MSS adjusts performed if the source IP is not part of the fabric's internal TEP pool

TCP-MSS Adjust Functionality

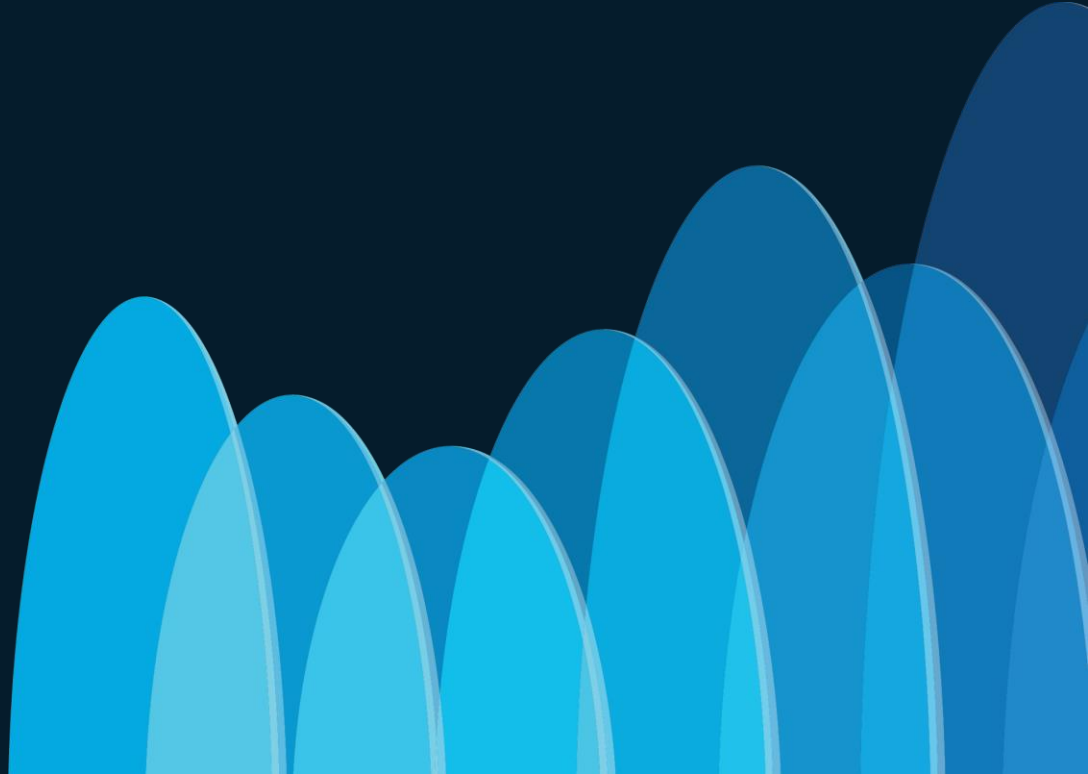
ACI Release 6.0(3)F

Inter-Site Data Packets

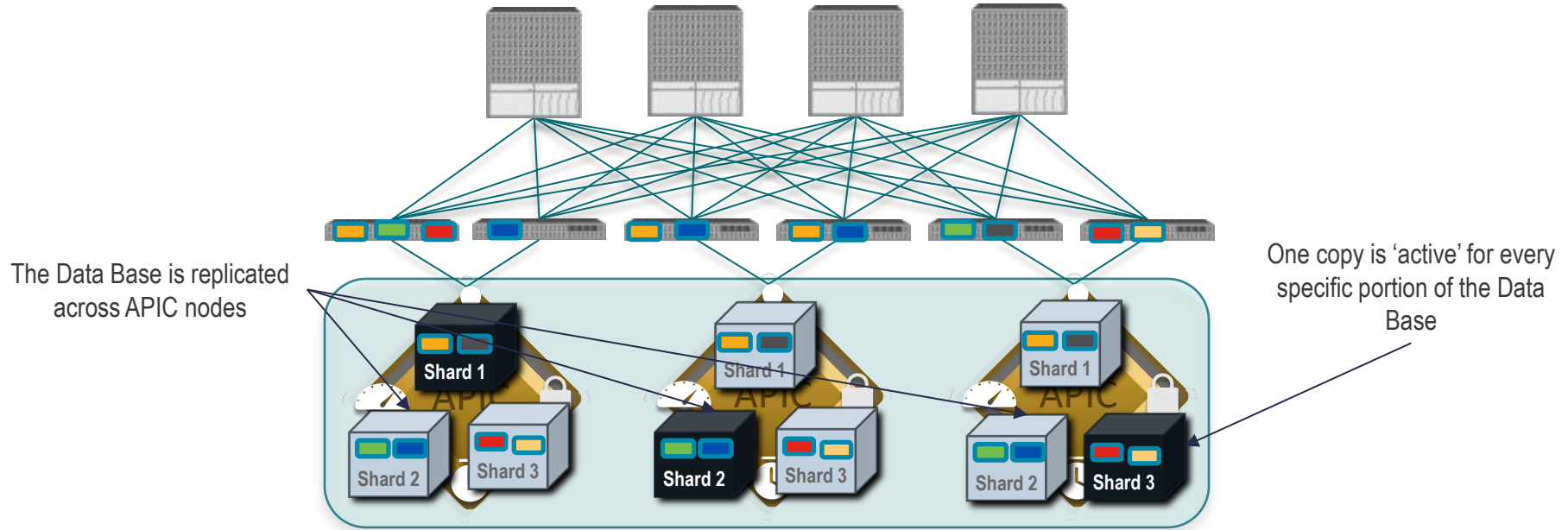


- As a result of the MSS negotiation, the endpoints generate packets for that TCP communication with total MTU 1440B (irrespectively of the local Host MTU)
- The VXLAN encapsulated traffic can be successfully forwarded across the ISN

Multi-Pod APIC Cluster Deployment Considerations



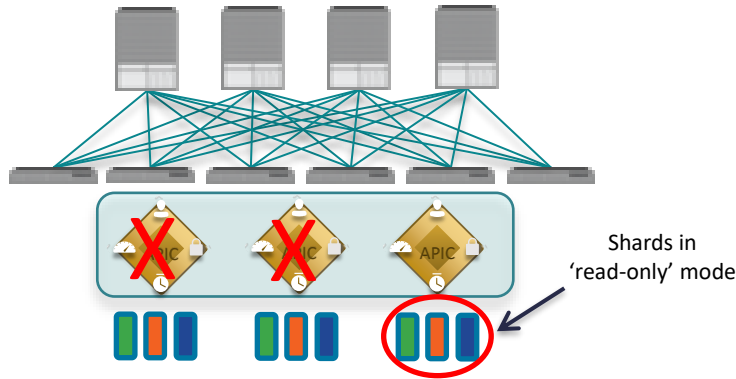
APIC – Distributed Multi-Active Data Base



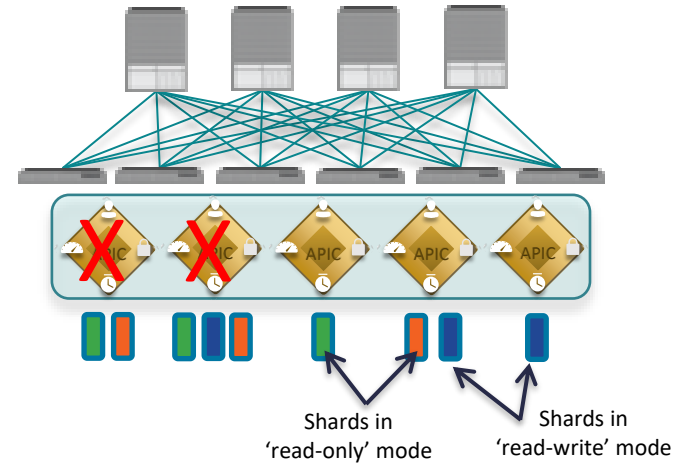
- Processes are active on all nodes (not active/standby)
- The Data Base is distributed as active + 2 backup instances (shards) for every attribute

APIC Cluster Deployment Considerations

Single Pod Scenario



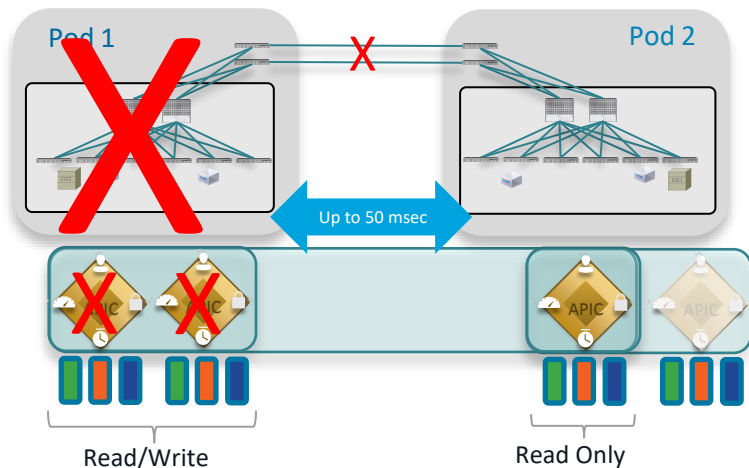
- APIC will allow read-only access to the DB when only one node remains active (standard DB quorum)
- **Hard failure** of two nodes cause all shards to be in 'read-only' mode (of course reboot etc. heals the cluster after APIC nodes are up)



- 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)

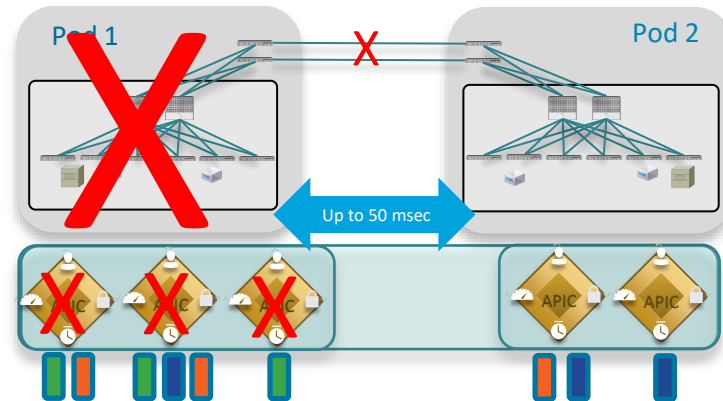
APIC Cluster Deployment Considerations

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

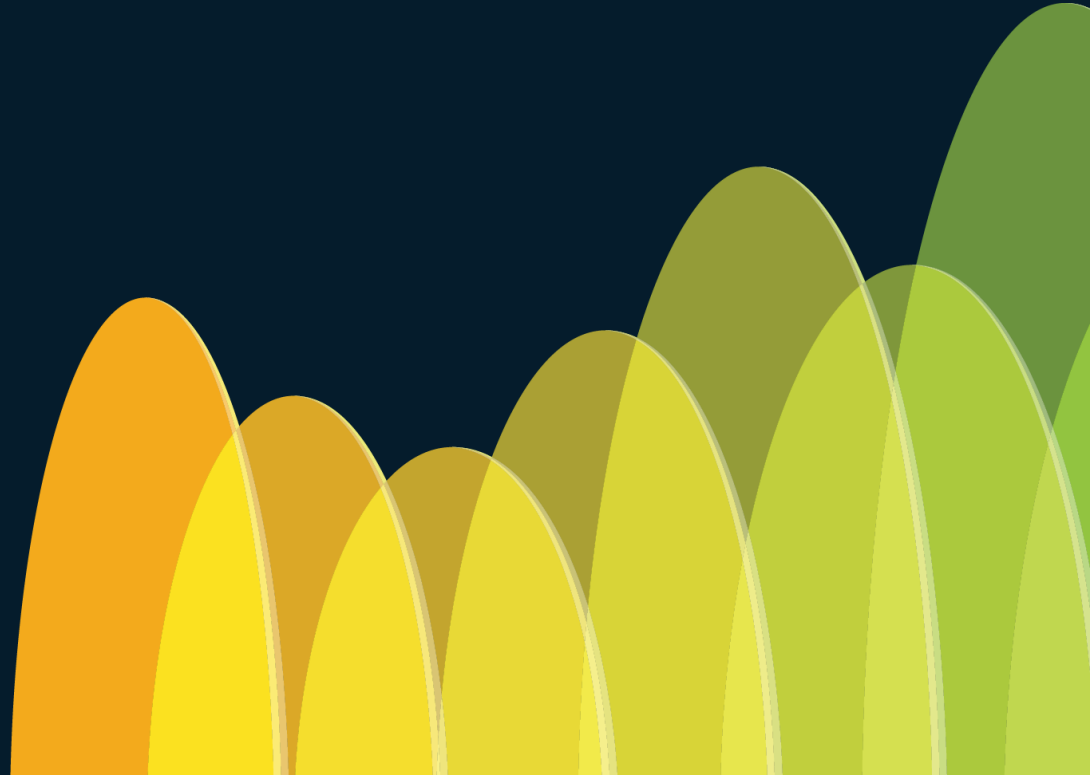
CISCO Live!



- **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**)

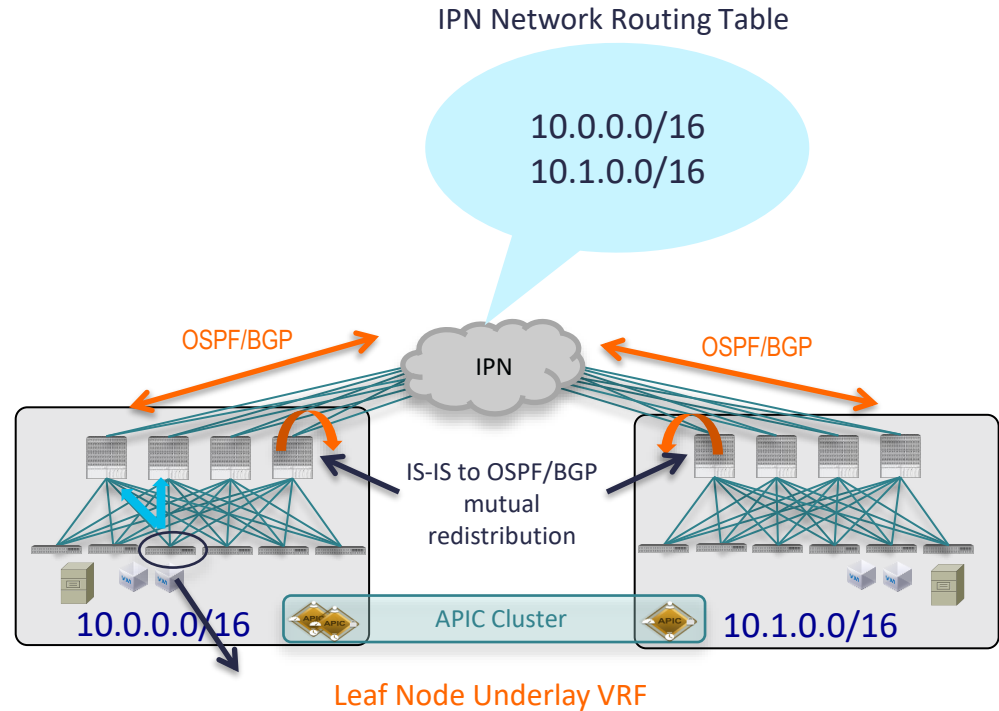
Multi-Pod Control and Data-Plane



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



ACI Multi-Pod

Inter-Pod MP-BGP EVPN Control Plane

- MP-BGP EVPN to sync Endpoint (EP) and Multicast Group information

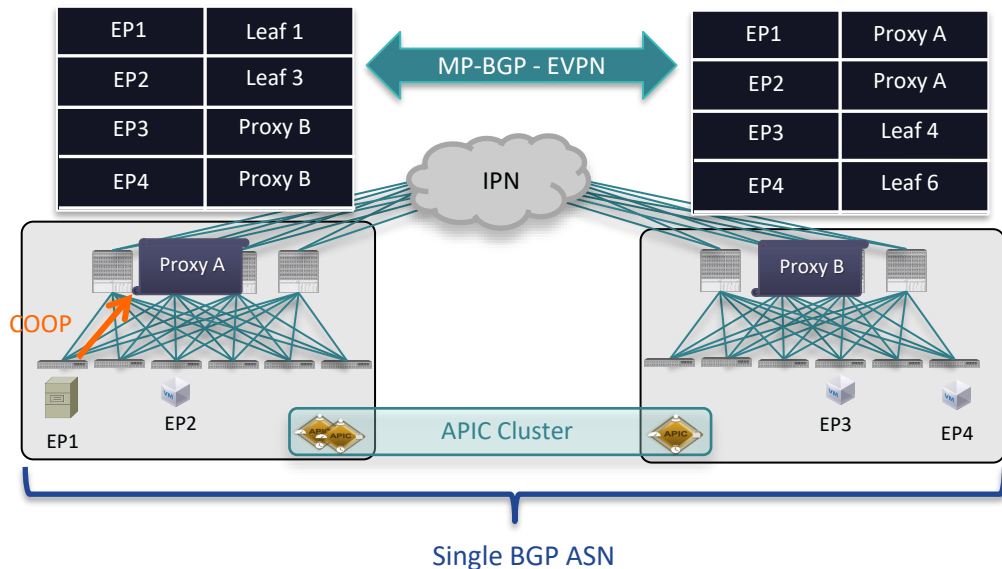
All remote Pod entries associated to a Proxy VTEP next-hop address (not part of local TEP Pool)

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)



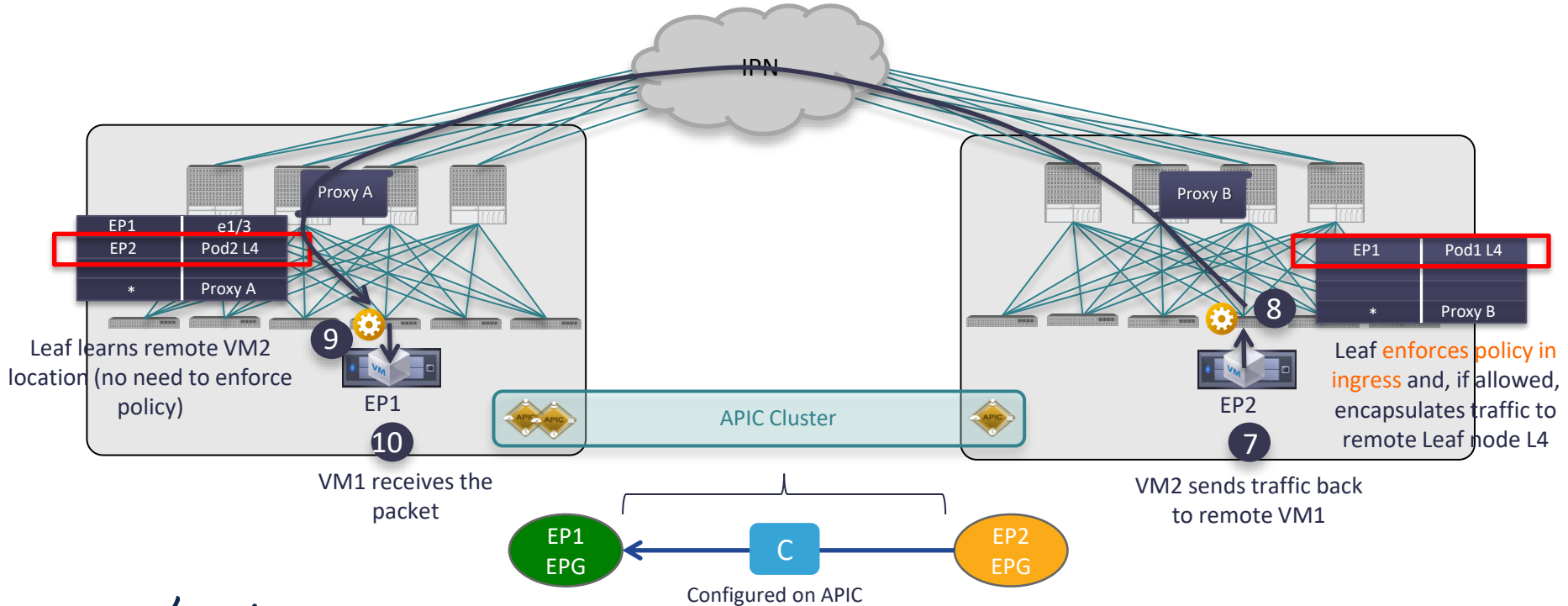
Inter-Pod Data Plane



ACI Multi-Pod

Inter-Pod Data Plane (2)

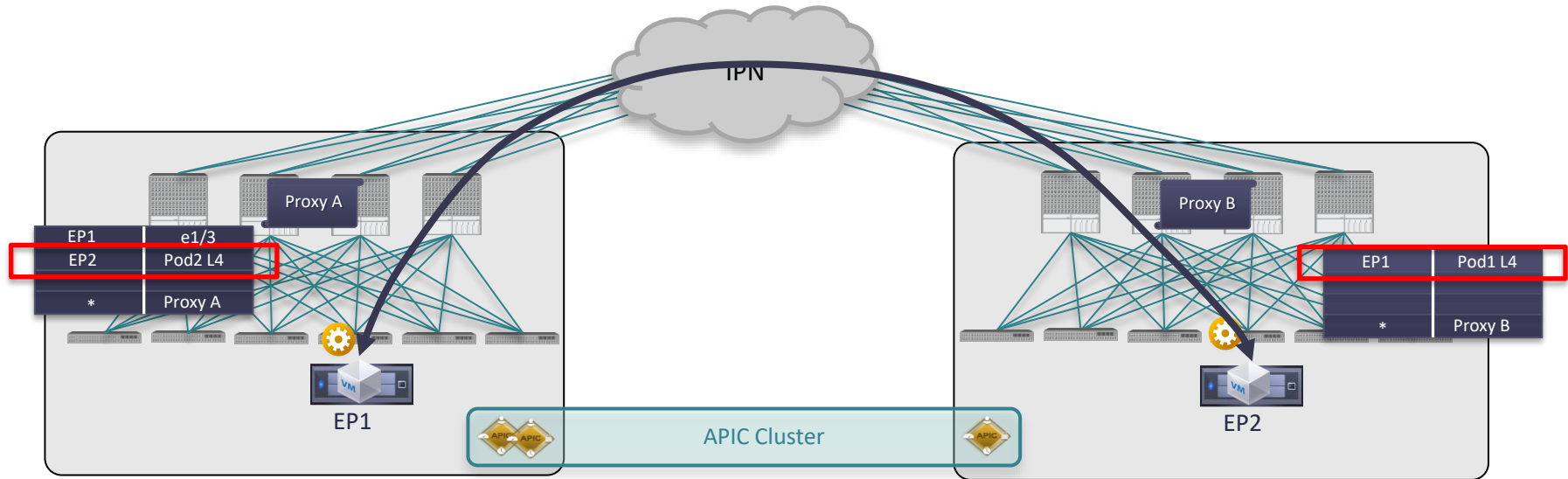
⚙️ = VXLAN Encap/Decap



ACI Multi-Pod

Inter-Pod Data Plane (3)

 = VXLAN Encap/Decap

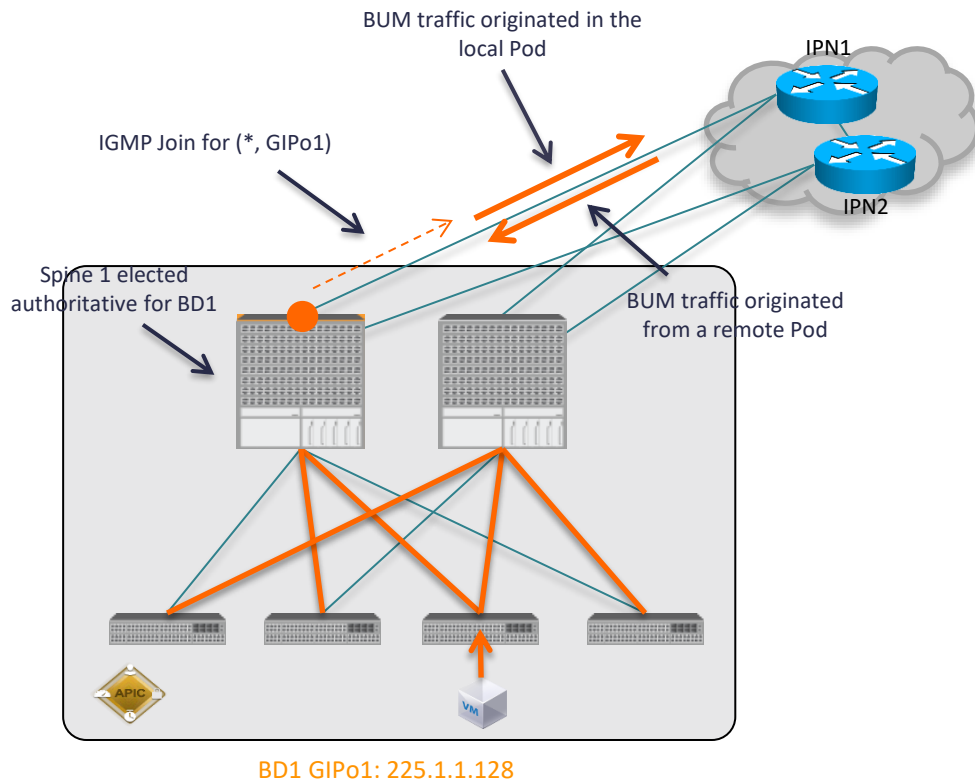


11

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

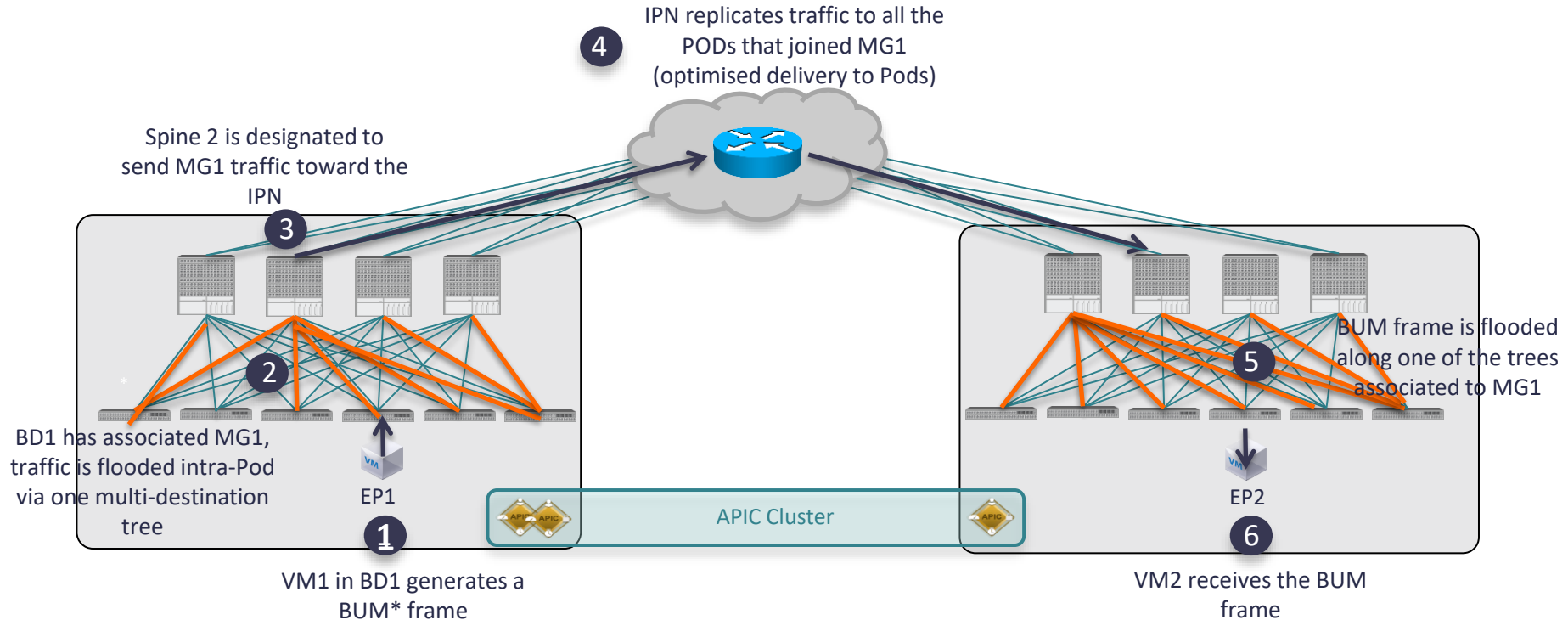
Use of Multicast for Inter-Pod Layer 2 BUM Traffic



- 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

ACI Multi-Pod

Use of Multicast for Inter-Pod BUM Traffic

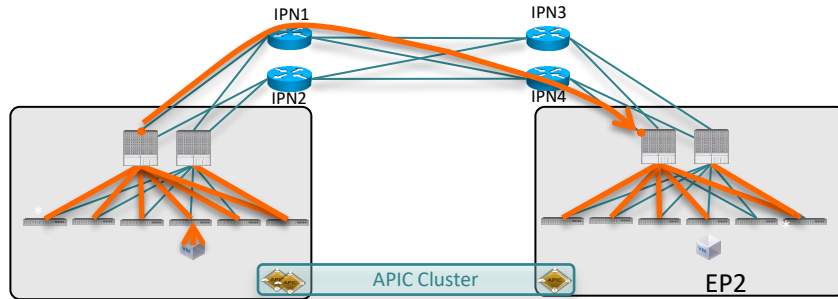


BUM: Layer 2 Broadcast, Unknown Unicast, Multicast

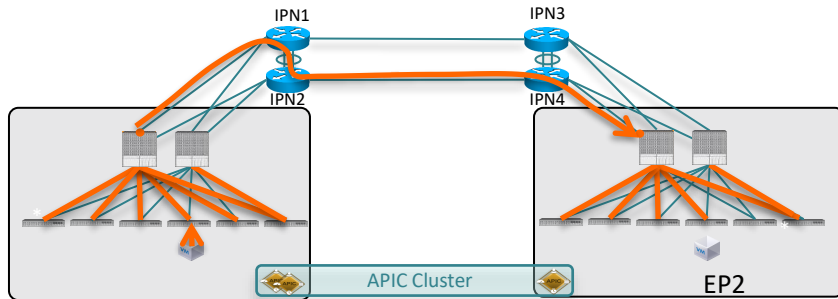
ACI Multi-Pod

PIM Bidir for BUM – Supported Topologies

Full Mesh between remote IPN devices

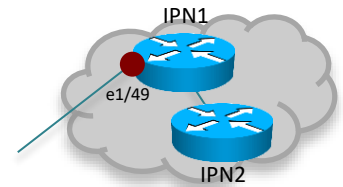


Directly connect local IPN devices



- 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
```



ACI Multi-Pod

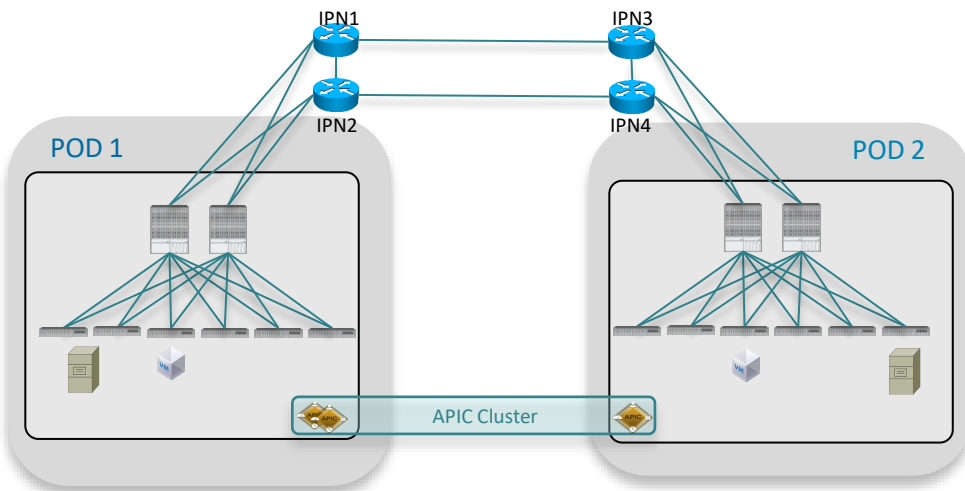
RP Redundancy with PIM Bidir

Active RP (IPN1)

```
ip pim rp-address 192.168.100.2 group-list
225.0.0.0/15 bidir
ip pim rp-address 192.168.100.2 group-list
239.255.255.240/28 bidir
interface loopback1
 description Phantom RP
 ip address 192.168.100.1/30
 ip ospf network point-to-point
 ip router ospf IPN area 0.0.0.0
 ip pim sparse-mode
```

Standby RP (IPN3)

```
ip pim rp-address 192.168.100.2 group-list
225.0.0.0/15 bidir
ip pim rp-address 192.168.100.2 group-list
239.255.255.240/28 bidir
interface loopback1
 description Phantom RP
 ip address 192.168.100.1/29
 ip ospf network point-to-point
 ip router ospf IPN area 0.0.0.0
 ip pim sparse-mode
```



- In PIM Bidir, only one device functions as active RP for a given group at the same time
- Can leverage the 'Phantom RP' configuration for providing resiliency
http://www.cisco.com/c/dam/en/us/products/collateral/ios-nx-os-software/multicast-enterprise/prod_white_paper0900aecd80310db2.pdf
- The RP role is 'active' on the device announcing the most specific route for the RP's address
- In case of failure of the 'active' RP device, the shared tree is immediately rebuilt toward the 'standby' RP based on routing convergence event
- Can deploy multiple RPs, each active for a sub-range of multicast groups

Multi-Site Control and Data-Plane



ACI Multi-Site

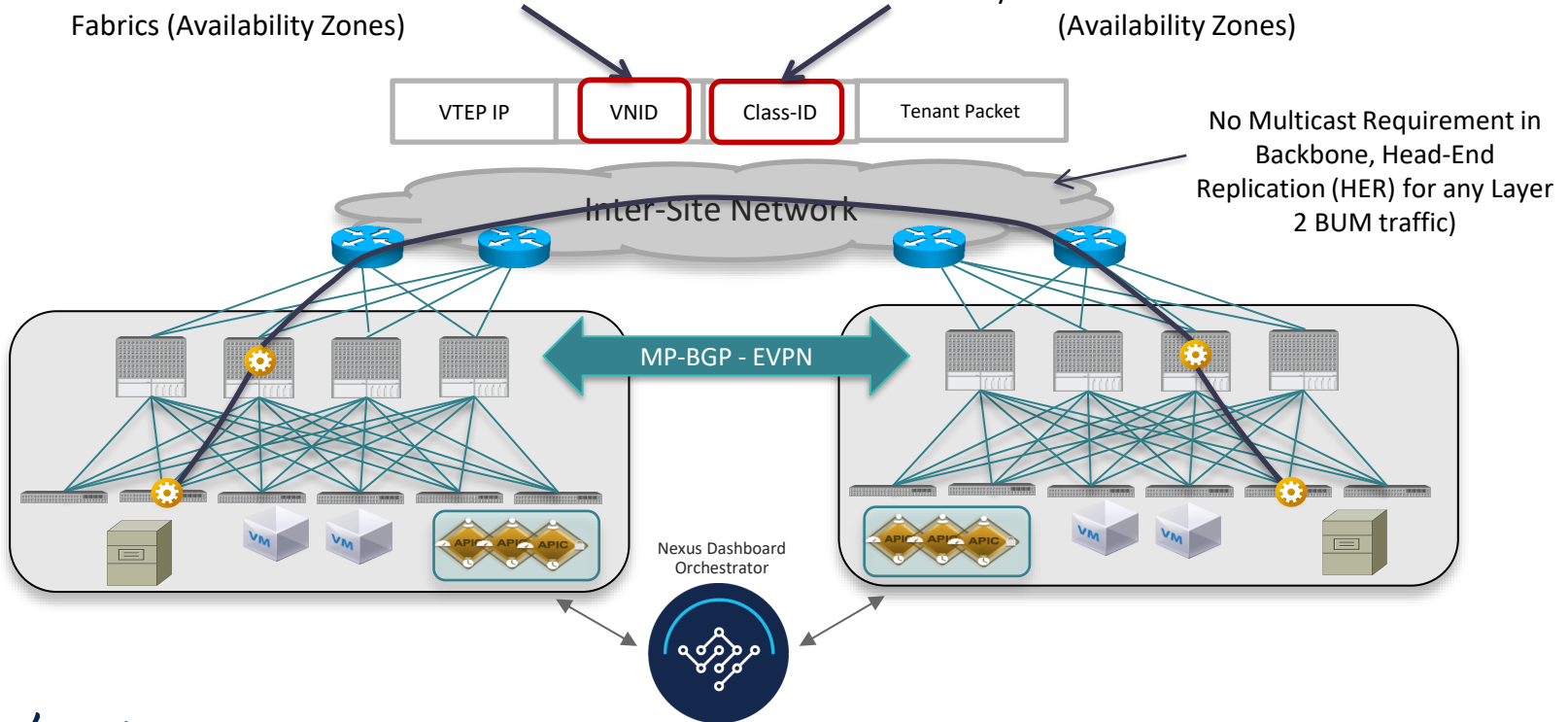
Network and Identity Extended between Fabrics

Deployment Mode ①

- Multi-Site
- Autonomous

Network information carried across Fabrics (Availability Zones)

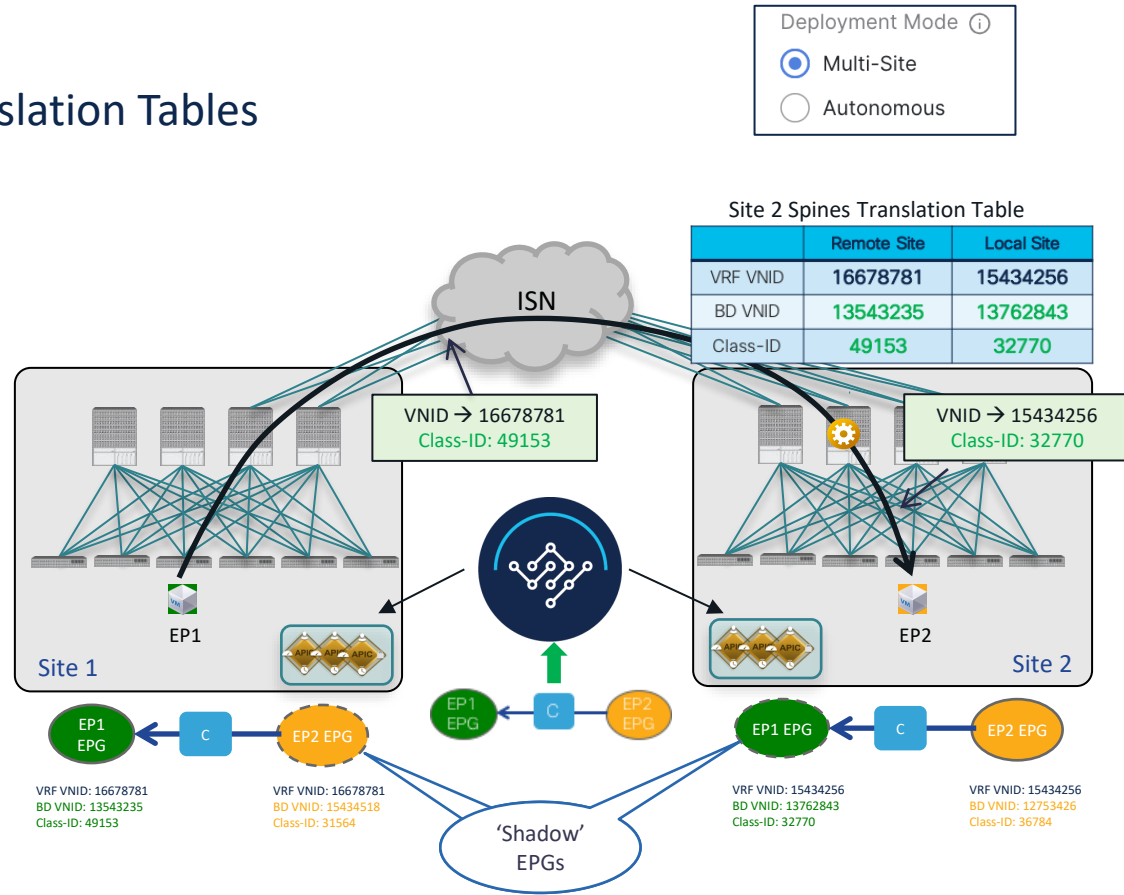
Identity information carried across Fabrics (Availability Zones)



ACI Multi-Site

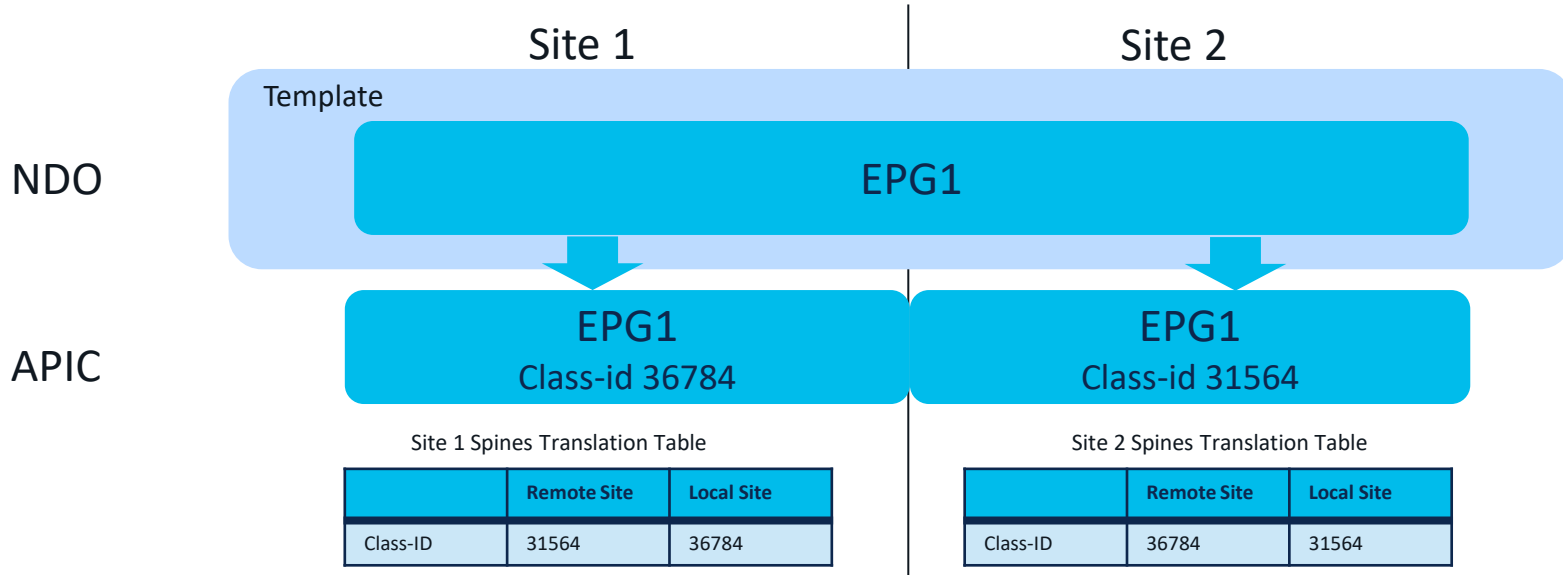
Inter-Site Policies and Spines' Translation Tables

- 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 normalisation)
- **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



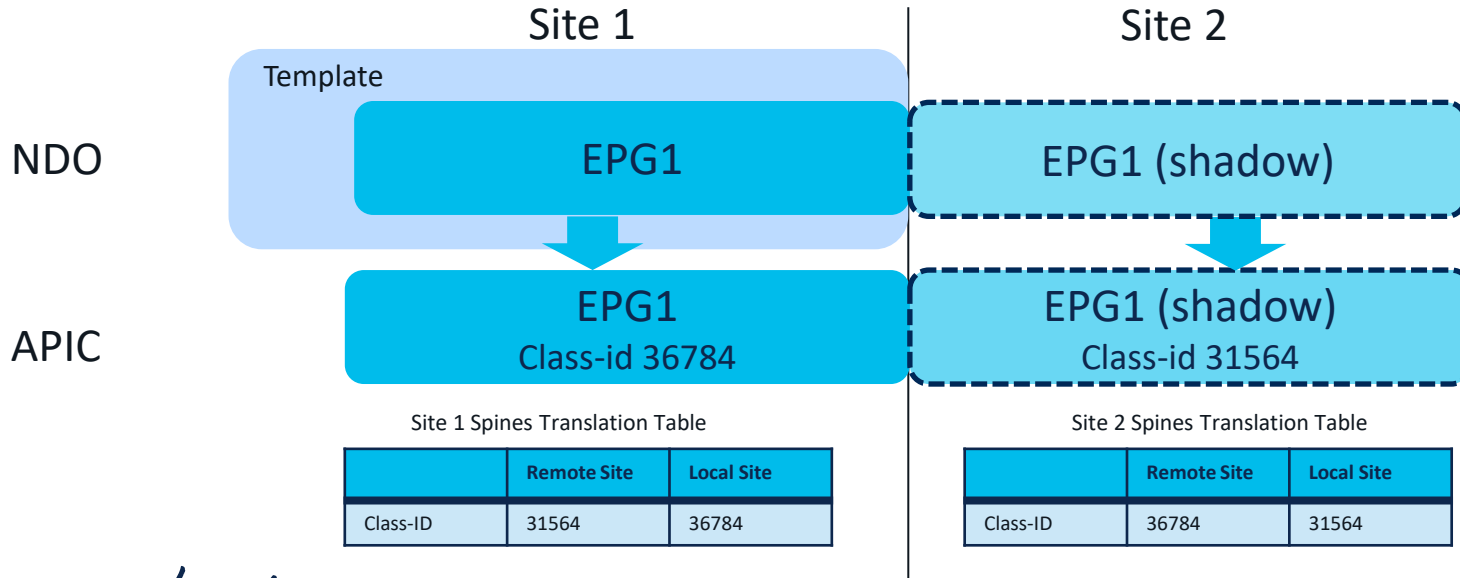
What is a stretched object?

- When an object is defined in a template associated to multiple sites the object considered a stretched object
- If that object is a VRF, Bridge Domain, EPG, or External EPG, NDO will cross program VNIDs and class-ids (translations)



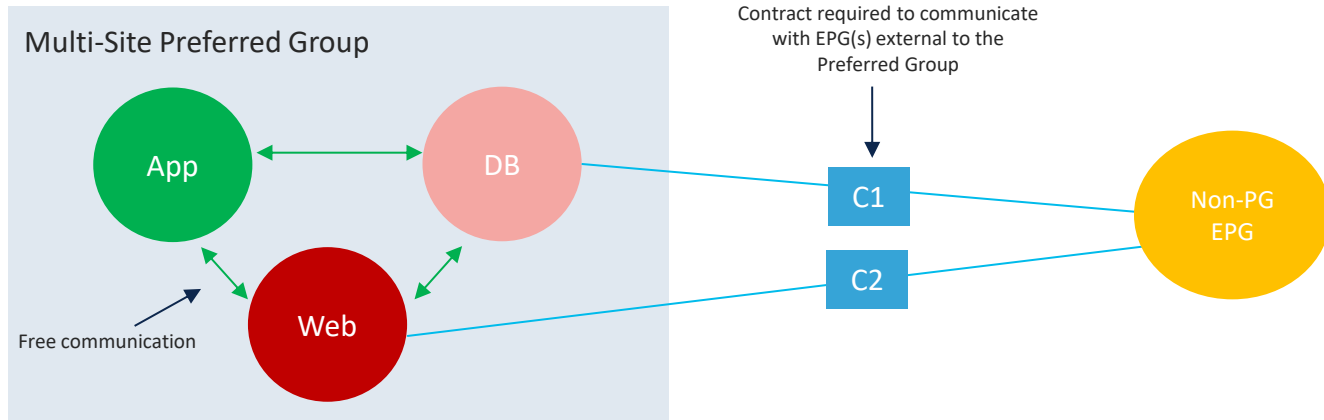
What is a shadow object?

- If an object is defined in a template associated to one site but translations are required in the other site to allow cross site communication (contract associations, preferred group, etc), NDO will program a shadow object in that site
- The shadow object created on ACI is like a stretched object with translations rules but is not user configurable. It is created automatically by NDO



ACI Multi-Site

Simplify Policy Enforcement: Preferred Groups

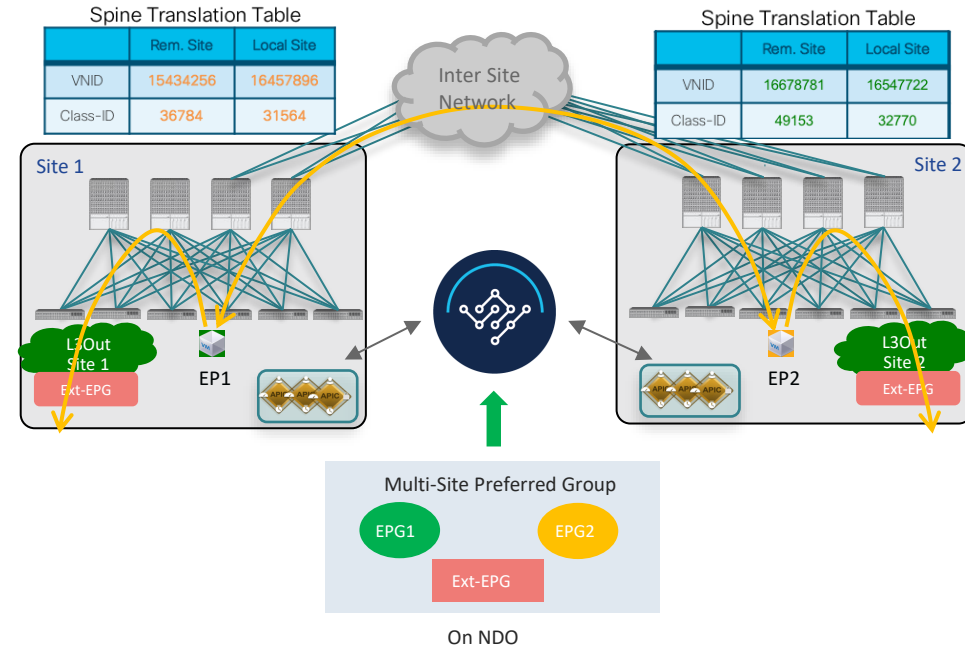


- “VRF unenforced” not supported with Multi-Site
- Multi-Site Preferred Group configuration can be provisioned directly from NDO
 - Creates ‘shadow’ EPGs and translation table entries ‘under the hood’ to allow ‘free’ inter-site communication
 - 5000 total EPGs part of preferred group supported in NDO 4.x release
- Typically desired in legacy to ACI migration scenarios

Simplify Policy Enforcement

Preferred Groups for E-W and N-S Flows

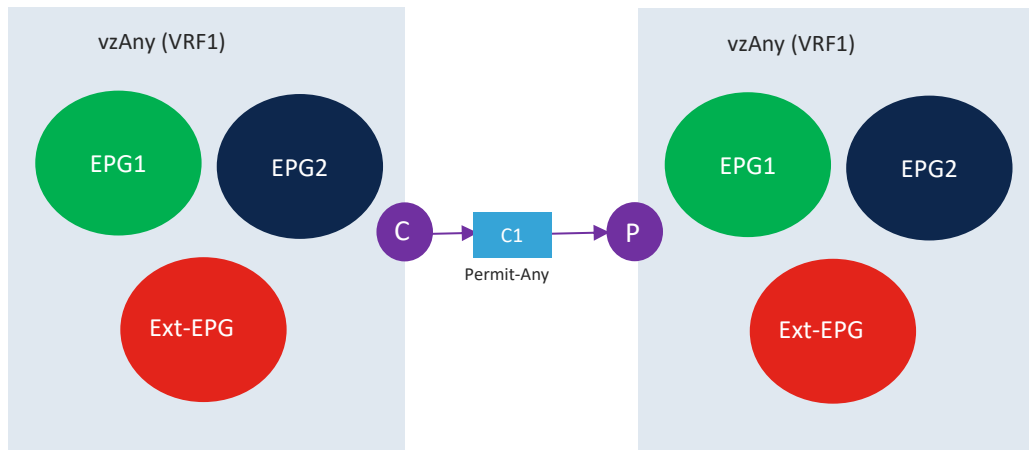
- Adding internal EPGs and External EPGs (associated to L3Outs) to the Preferred Group allows to enable free east-west and north-south connectivity
- When adding the Ext-EPG to the Preferred Group:
 - Can't use 0.0.0.0/0 for classification, needs more specific prefixes
 - As workaround it is possible to use 0.0.0.0/1 and 128.0.0.0/1 to achieve the same result
 - Must ensure Ext-EPG is a stretched object
- **Intersite L3Out not supported if the Ext-EPG is part of a Preferred Group**



Simplify Policy Enforcement

Enabling Free Communication inside a VRF

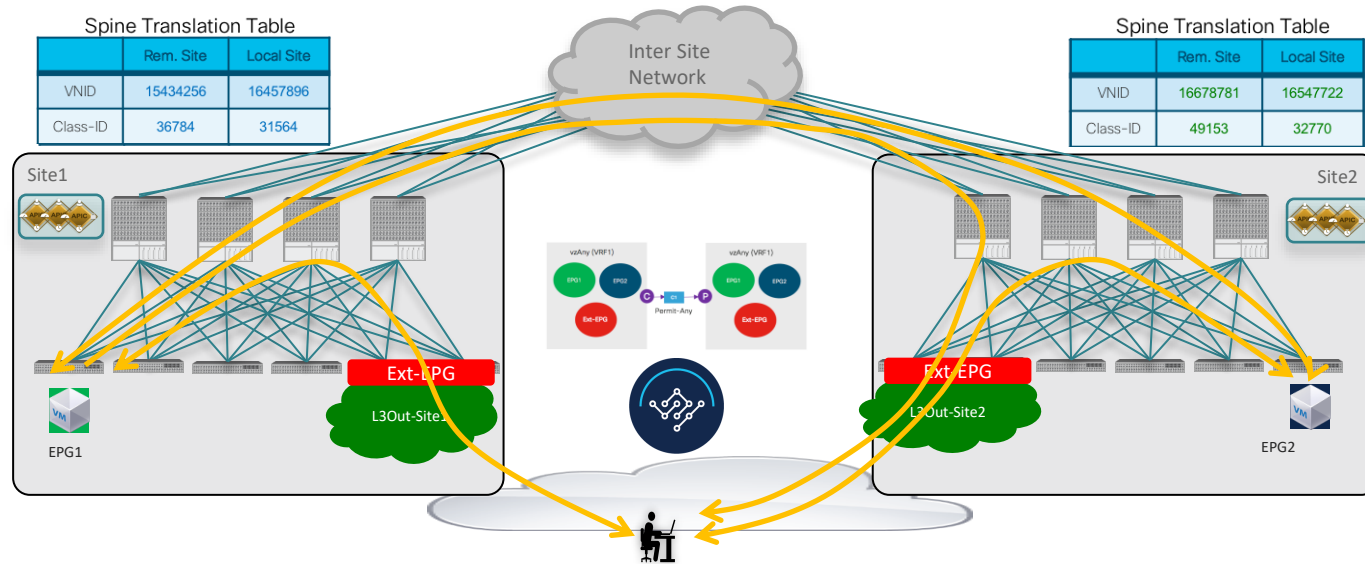
What is vzAny? Logical object representing all the EPGs/Ext-EPGs in a VRF



- vzAny provides and consumes a contract with an associated “Permit-any” filter
- Use ACI fabric only for network connectivity without policy enforcement
- Equivalent to “VRF unenforced”

Simplify Policy Enforcement

Enabling Free Communication inside a VRF

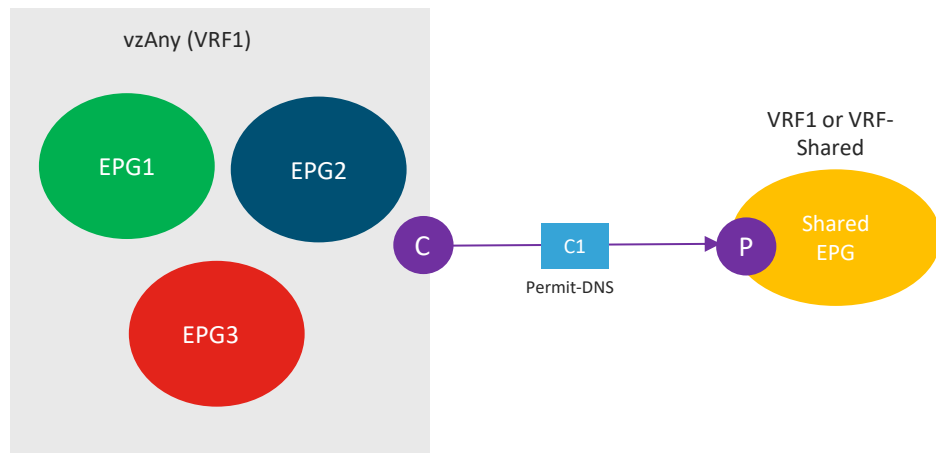


- Proper translation entries are created on the spines of both fabrics to enable east-west and north-south communication
- Supported also for connecting to the external Layer 3 domain
- vAny + PBR support available from NDO release 4.2(3) and ACI release 6.0(4)

Simplify Policy Enforcement

Enabling Many-to-One communication (Shared Services)

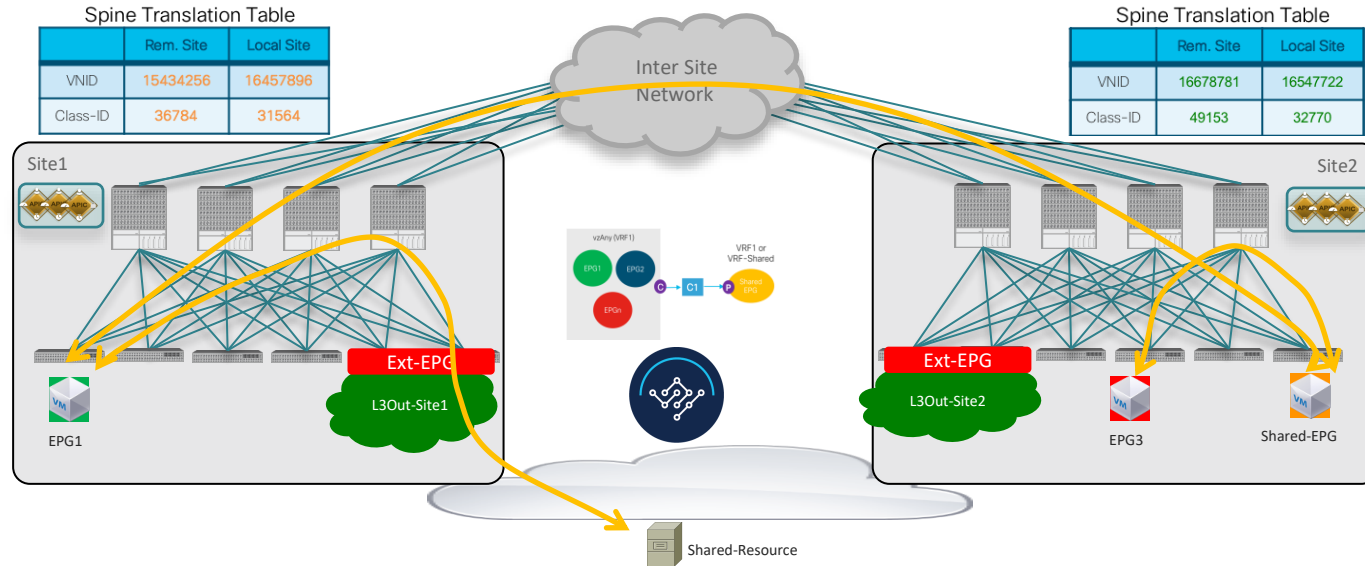
What is vzAny? Logical object representing all the EPGs/Ext-EPGs in a VRF



- Multiple EPGs part of a specific VRF1 consume the services provided by a shared EPG (part of VRF1 or of a VRF-shared)
- VRF-shared can be part of the same tenant or of a different tenant

Simplify Policy Enforcement

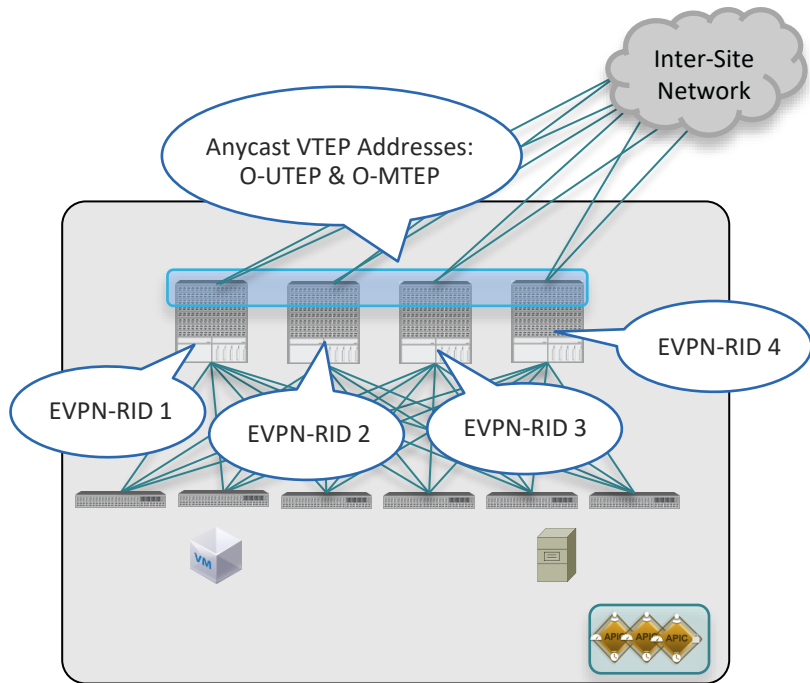
Enabling Many-to-One communication (Shared Services)



- Proper translation entries are created on the spines of both fabrics to enable east-west communication
- Supported also for Shared Services behind an L3Out

ACI Multi-Site

BGP Inter-Site Peers

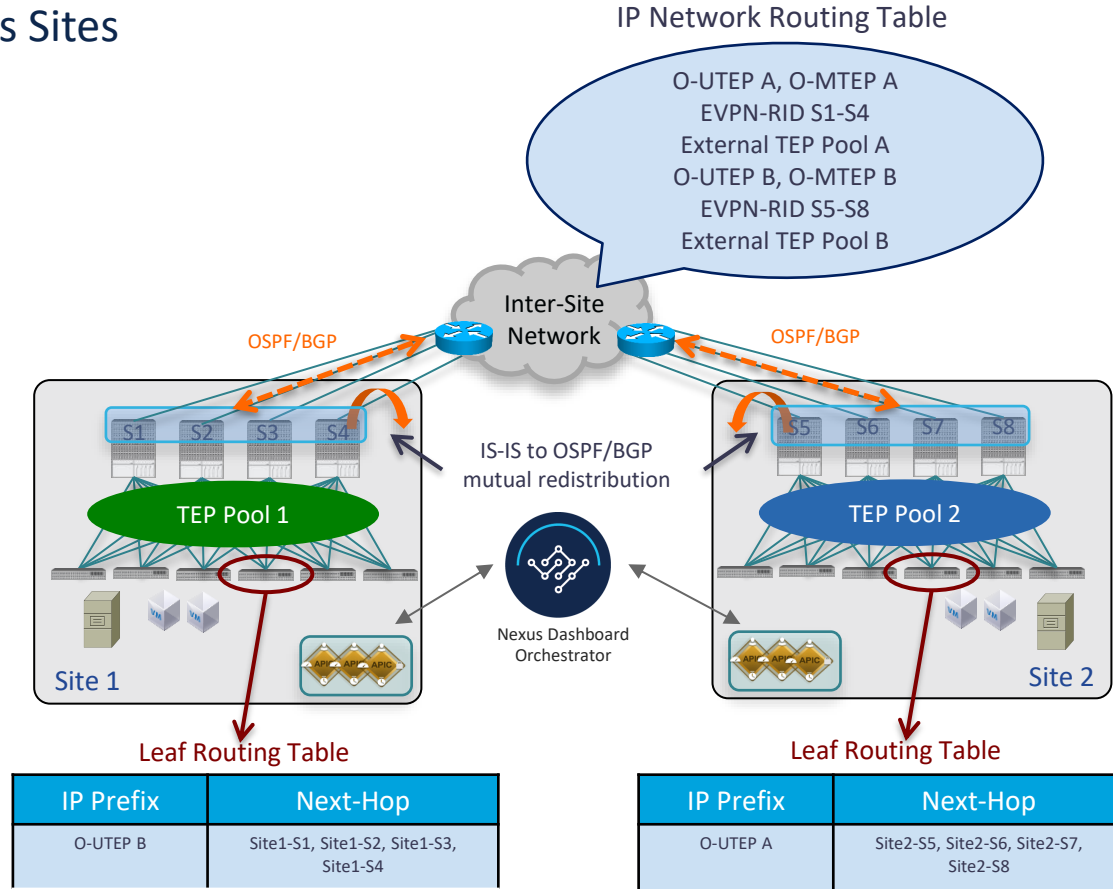


- Spines connected to the Inter-Site Network perform two main functions:
 1. Establishment of MP-BGP EVPN peerings with spines in remote sites
 - One dedicated Control-Plane address (EVPN-RID) is assigned to each spine running MP-BGP EVPN
 2. Forwarding of inter-sites data-plane traffic
 - Anycast Overlay 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
- EVPN-RID, O-UTEP and O-MTEP addresses are assigned from the Nexus Dashboard Orchestrator and must be routable across the ISN

ACI Multi-Site

Exchanging TEP Information across Sites

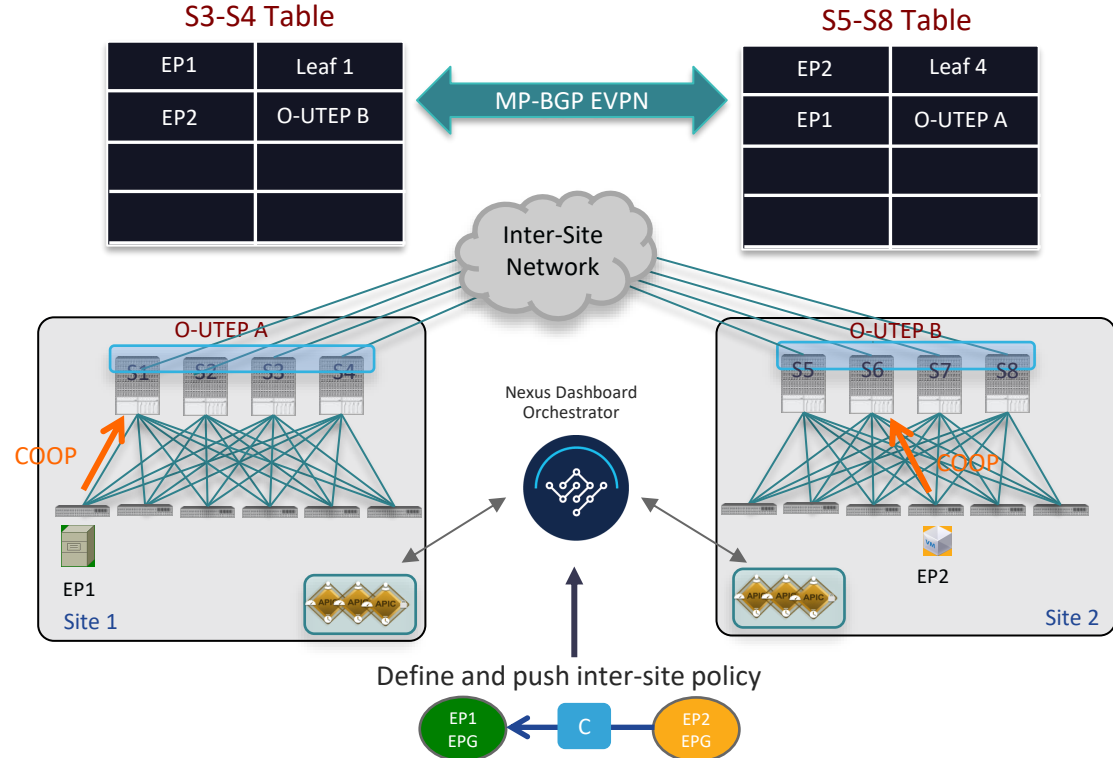
- OSPF or BGP peering between spines and Inter-Site network
- Mandates the use of L3 sub-interfaces (with VLAN 4 tag) between the spines and the ISN
- Exchange of External Spine TEP addresses (EVPN-RID, O-UTEP, O-MTEP, and External TEP Pool) across sites
 - Use of overlapping internal TEP Pools across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



ACI Multi-Site

Inter-Site MP-BGP EVPN Control Plane

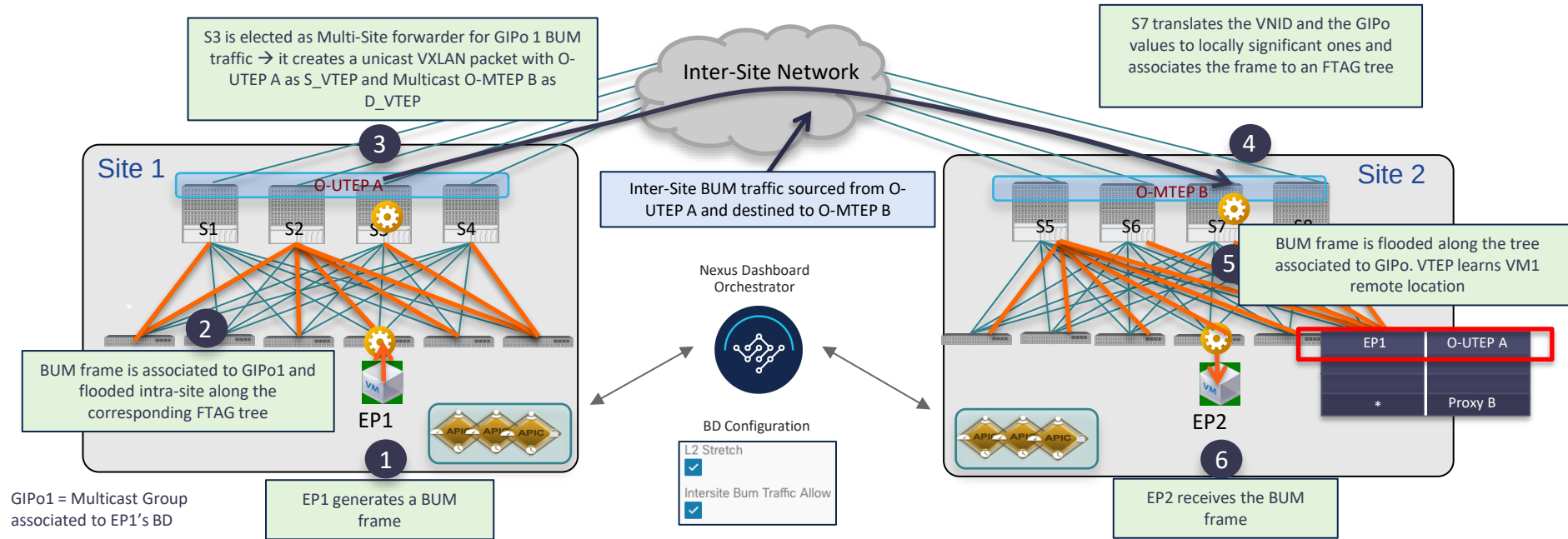
- MP-BGP EVPN used to communicate Endpoint (EP) information across Sites
 - MP-iBGP or MP-EBGP peering options supported
 - Required MP-BGP configuration fully automated via NDO
 - Remote host route entries (EVPN Type-2) are associated to the remote site Anycast O-UTEP address
- Automatic filtering of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



ACI Multi-Site

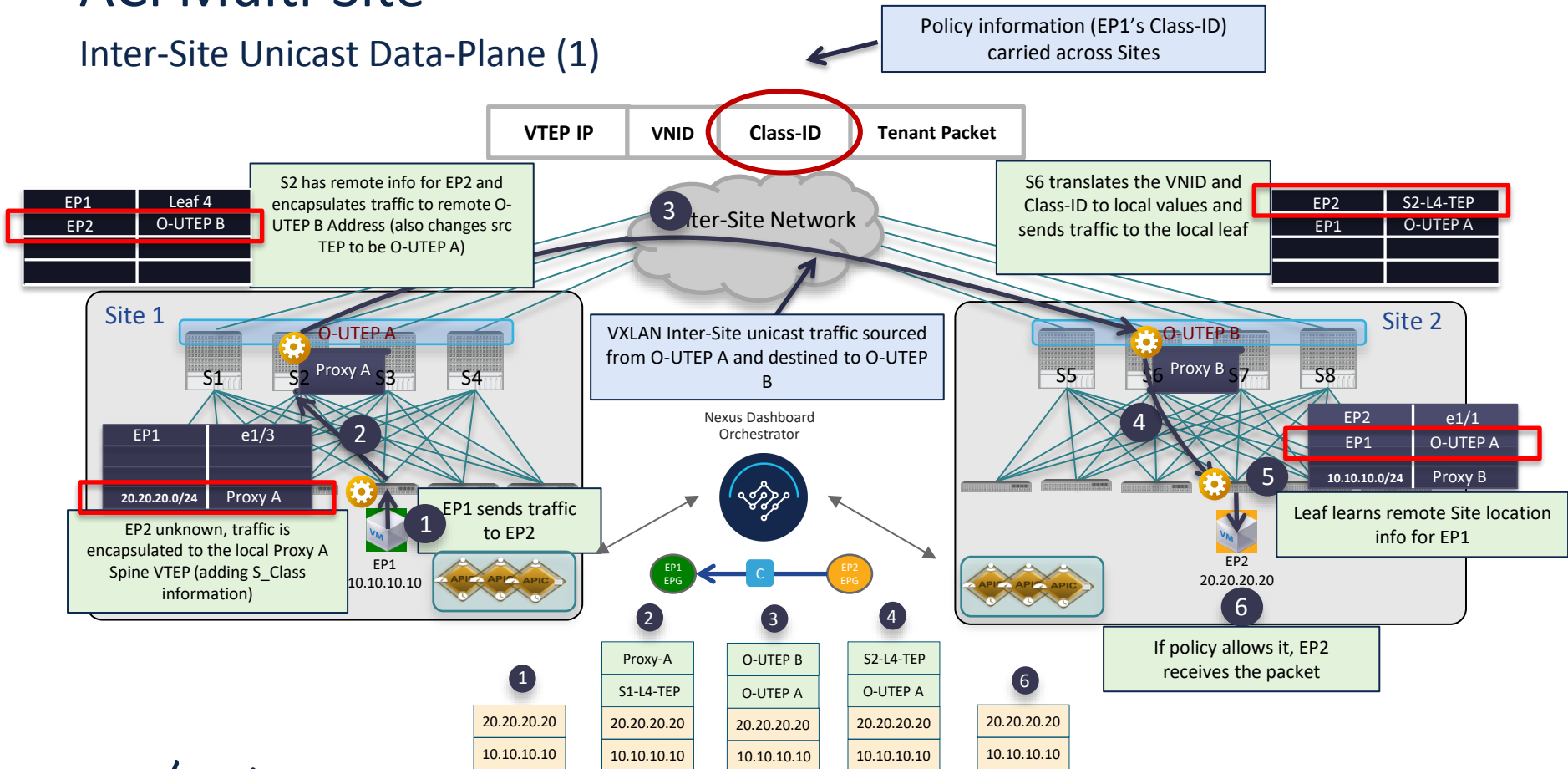
Inter-Site Layer 2 BUM* Forwarding

*BUM – Broadcast, Unknown Unicast, Multicast



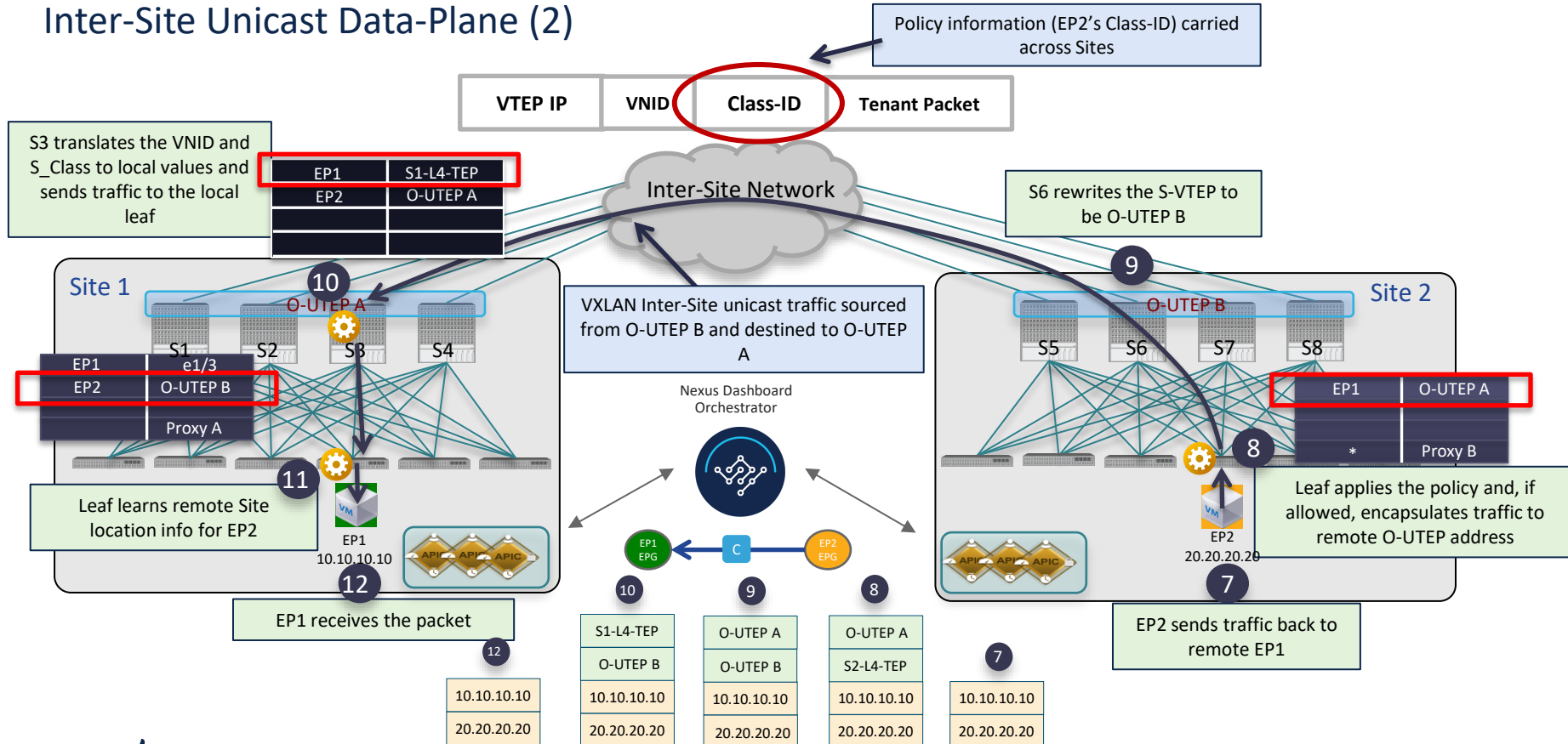
ACI Multi-Site

Inter-Site Unicast Data-Plane (1)



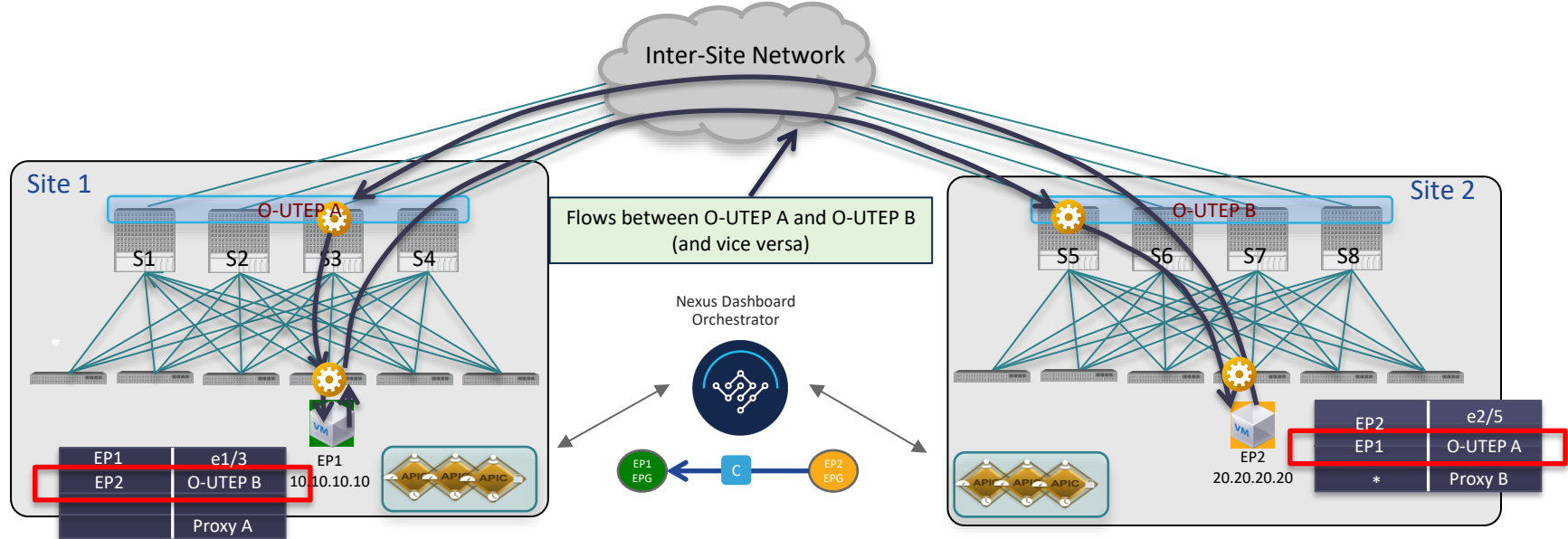
ACI Multi-Site

Inter-Site Unicast Data-Plane (2)

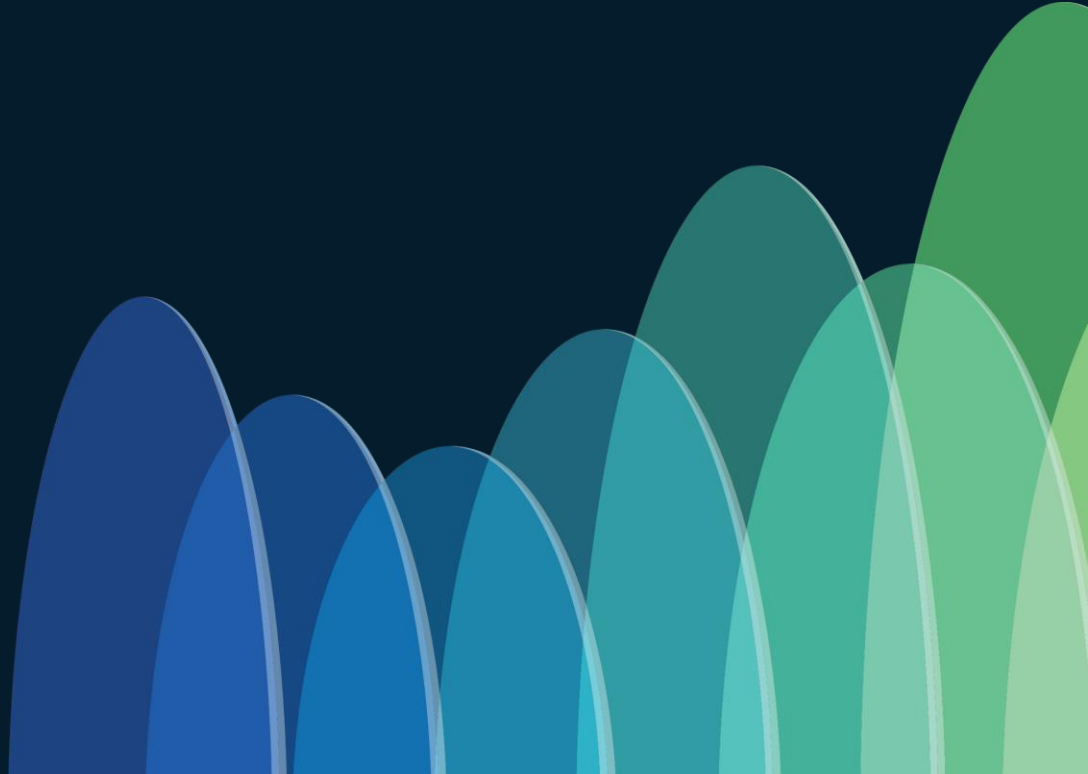


Inter-Site Unicast Data-Plane (3)

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

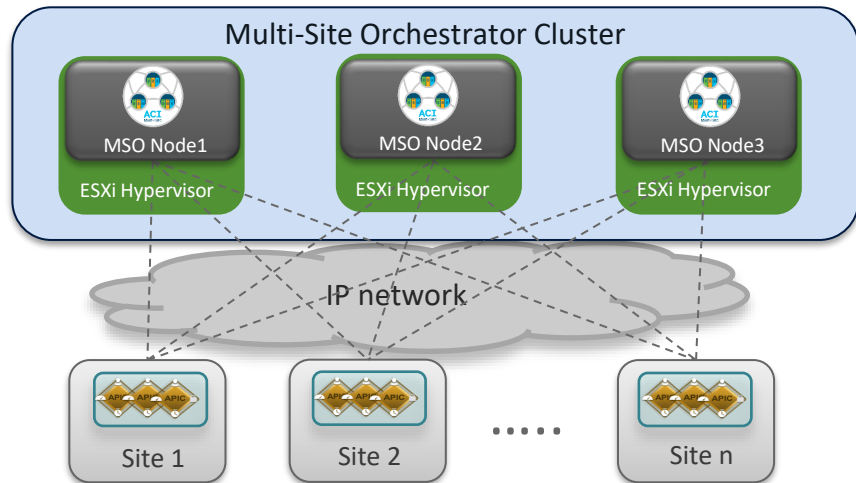


Nexus Dashboard Orchestrator (NDO)



Original Multi-Site Orchestrator Option

VM Based MSO Cluster (OVA), Now EoL/EoS



- Supported from the beginning (MSO release 1.0(1))
- Each Cisco Multi-Site Orchestrator node is packaged in a VMware vSphere virtual appliance (OVA)
- For high availability, you should deploy each Cisco Multi-Site Orchestrator virtual machine on its own VMware ESXi host
- Requirements for MSO Release 1.2(x) and above:
 - VMware ESXi 6.0 or later
 - Minimum of eight virtual CPUs (vCPUs), 48 Gbps of memory, and 100 GB of disk space
- **MSO 3.1(1) last supported release with this form factor, now EoL/EoS**

Cisco Multi-Site Orchestrator has become Cisco Nexus Dashboard Orchestrator



Up to release 3.1(1)



From release 3.2(1)

Cisco Nexus Dashboard

Simple to Automate, Simple to Consume

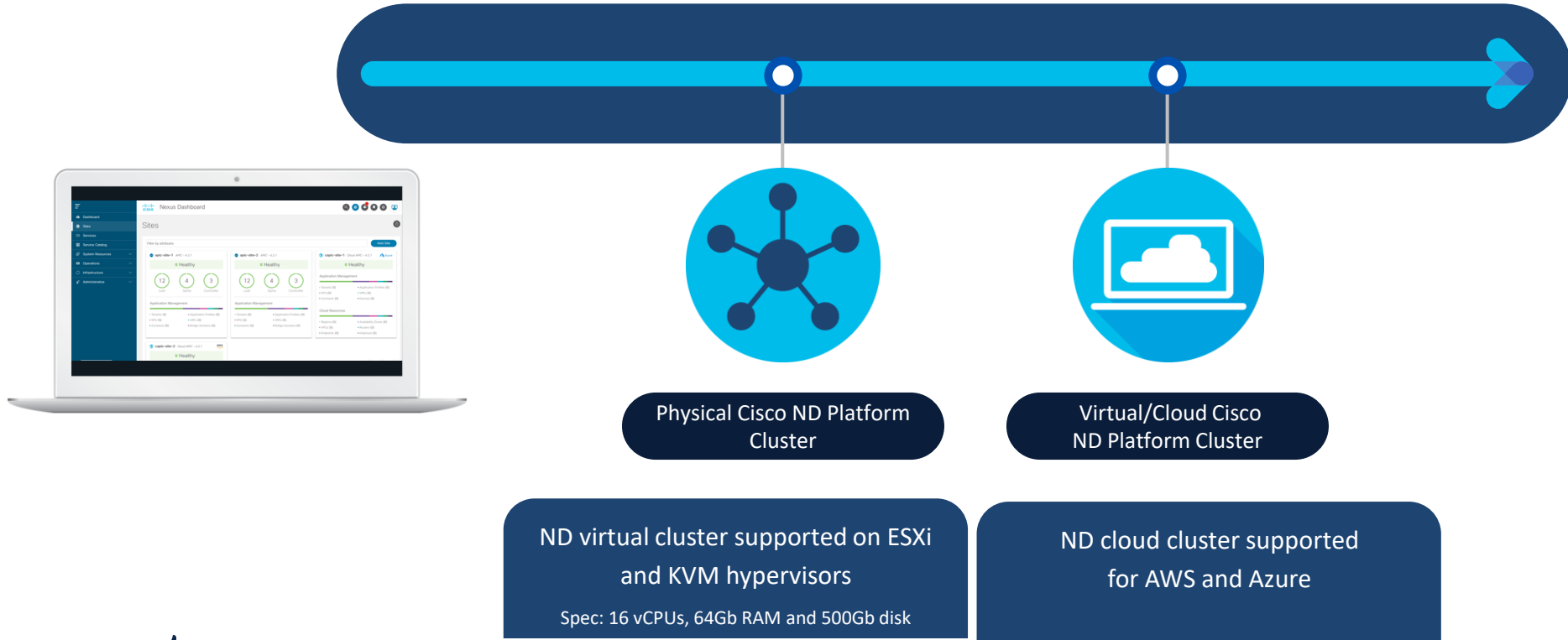
Powering **automation**
Unified **agile** platform



CISCO *Live!*

Cisco Nexus Dashboard

Deployment Evolution



NDO Upgrade/Migration Considerations

Recommended Release per Scenario

Recommended Releases per Scenario	
Current Release	Target Release
MSO/NDO 1.1(x) to 3.7(2)	NDO 4.2(3)
NDO 4.0(1) to 4.2(2)	NDO 4.2(3)
None - Greenfield	NDO 4.2(3)

- Migration procedure required between any old MSO release to NDO
- Direct upgrade supported from any old NDO release to NDO 4.2(3) (and newer)
- Note that Nexus Dashboard may need to be upgraded first

Upgrade Options

Upgrading to the latest NDO version may also require Nexus Dashboard Upgrade. The two upgrade options are a direct upgrade or deploy a new cluster and restore from backup.

Physical cluster?

Direct upgrade may be the best option but may require multiple upgrade steps if on an older release

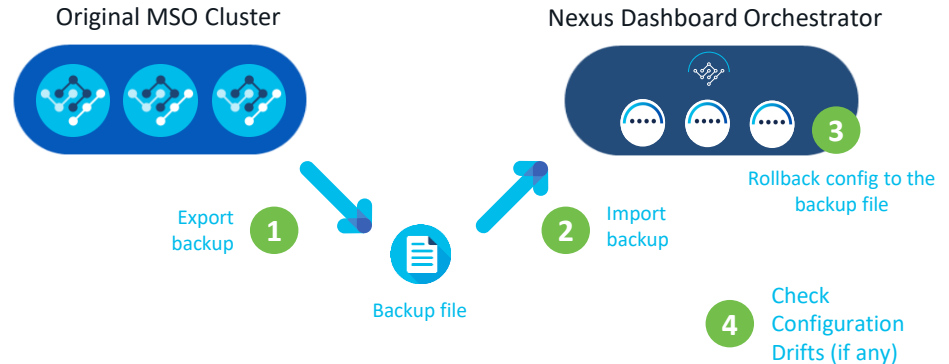
Virtual Cluster? MSO?

For MSO you can only do the restore from backup. May be preferred over direct upgrade for all virtual form factors. Avoids multi-step upgrade and provides an easy rollback option

Migrating the MSO Cluster to Cisco NDO

Migration Procedure

- Moving from MSO to NDO requires executing a migration procedure
 - Exporting a backup config file from MSO and importing it on NDO
- The rollback of the config file on NDO includes a specific “database cleaning” procedure
 - Important for removing any stale object that may exist in the original MSO database



Upgrading/Migrating to an NDO 4.x release may involve a template transformation →
Be aware of this and ideally try this upfront in the lab

<https://www.cisco.com/c/en/us/td/docs/dcn/ndo/4x/deployment/cisco-nexus-dashboard-orchestrator-deployment-guide-421/ndo-deploy-migrate-42x.html>

Cisco Nexus Dashboard 3.1

- Nexus Dashboard 3.1 and later is deployed as a unified image
- Services, Orchestrator, Fabric Controller, and/or Insights are enabled when deploying the cluster
- It is no longer required to install services as apps on the platform (no more .nap files)
- The Nexus Dashboard will only allow you to select supported combinations based on cluster being deployed (virtual, physical, node type, disk, CPU, etc)
- Supports upgrade* and backup/restore migrations

*Refer to migration guide for supported release combinations for upgrades

Nexus Dashboard Capacity Planning

- [See the Nexus Dashboard Capacity Planning Tool](#) to plan your deployment
- Select the form factor and services you wish to run
- Cluster size, CPU, memory, and disk requirements are displayed for the selected options



Nexus Dashboard Capacity Planning

This page provides cluster sizing guidelines based on the type and number of services you plan to run in your Nexus Dashboard as well as the target fabrics' sizes. The provided information has been tested and validated by Cisco, Cisco partners, or both.

For feedback on this tool, send email to ciscodcnapps-docfeedback@cisco.com.

Nexus Dashboard Release:

Cluster Form Factor:

- ☐ Physical
- ☒ Virtual (vCenter/ESXi)
- ☐ Virtual (KVM)
- ☐ Cloud (AWS/Azure)

Nexus Dashboard Services:

- ☐ Insights
- ☒ Orchestrator
- ☐ Fabric Controller

Cluster size:

- 3 'primary' OVA-App nodes
- 1 'primary' OVA-App node for non-production lab/testing environments only

Each OVA-App node requires 16 vCPU, 64GB memory, 550GB HDD or SSD/NVMe.

Maximum scale:

- NDO deployments do not differentiate between small, medium, or large fabrics. Minimum size cluster supports full scale as listed in [Verified Scalability Guide](#) for your release.

Nexus Dashboard 3.1 Deployment

Cluster Bringup
Answer some questions, select the services you want to enable and have Nexus Dashboard ready to use in a few minutes.

Deployment Mode
Select which of these available services you would like to enable. [Learn More](#)

- ☐ **Fabric Controller**
Automate and manage network connectivity for LAN and SAN fabrics
- ☒ **Orchestrator**
Automate Data Center and Cloud Interconnect while centralizing network and policy configurations
- ☐ **Insights**
Accelerate time to remediation through telemetry and smart analytics

Select the services you would like to enable during deployment. Services that cannot be co-hosted will be greyed out

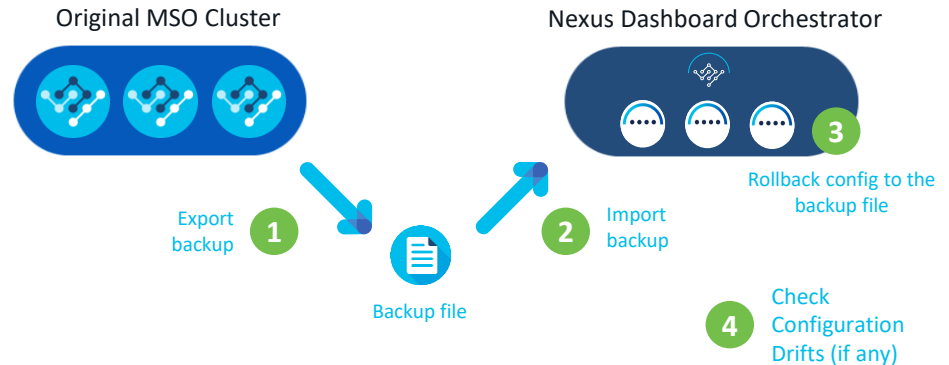


The deployment mode cannot be changed after the cluster is deployed

Migrating the MSO Cluster to Cisco NDO

Migration Procedure

- Moving from MSO to NDO requires executing a migration procedure
 - Exporting a backup config file from MSO and importing it on NDO
- The rollback of the config file on NDO includes a specific “database cleaning” procedure
 - Important for removing any stale object that may exist in the original MSO database

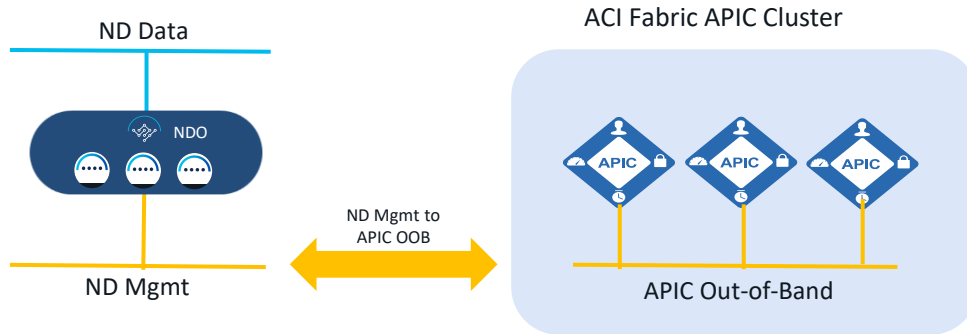


Upgrading/Migrating to an NDO 4.x release may involve a template transformation →
Be aware of this and ideally try this upfront in the lab

<https://www.cisco.com/c/en/us/td/docs/dcn/ndo/4x/deployment/cisco-nexus-dashboard-orchestrator-deployment-guide-421/ndo-deploy-migrate-42x.html>

NDO and APIC Connectivity Considerations

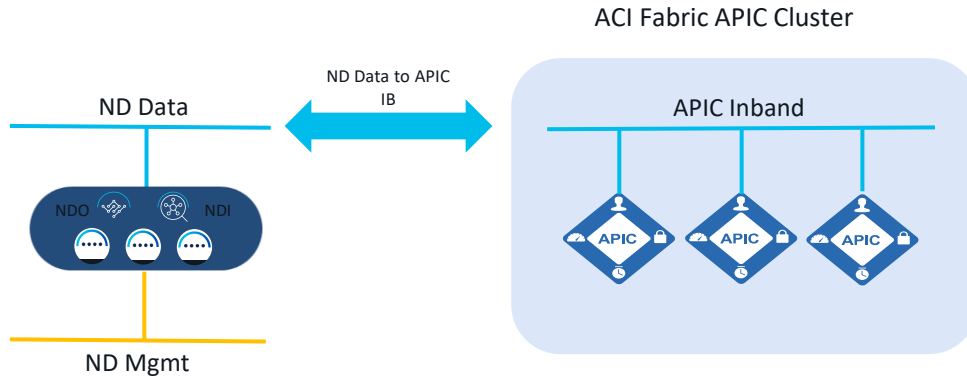
Option 1 – ND Mgmt to APIC OOB



- This option is the natural evolution of the original MSO deployment model
 - MSO used a single interface per cluster node and only communication to APIC OOB was supported
- Recommended if NDO is the only service running on the ND cluster
 - This is because other services (as NDI for example) may mandate connectivity to APIC IB instead
- Can be achieved in two ways:
 1. Connecting ND Mgmt interface to the same subnet of the APIC OOB interfaces
 2. Adding a static route on ND associated to the ND Mgmt interface to reach APIC OOB subnet

NDO and APIC Connectivity Considerations

Option 2 – ND Data to APIC OOB



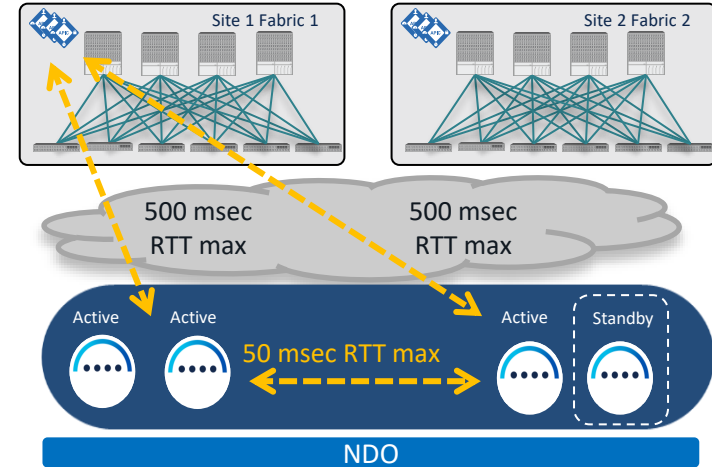
- When running NDO and NDI on the same ND cluster, each fabric must be onboarded using the APIC nodes In-Band addresses
- Even when starting only with NDO, the recommendation is to use connectivity to APIC In-Band if there are plans to add NDI later
- By default the ND Data interface is used to connect to APIC IB (for onboarding and for API calls), as long as the APIC IB interfaces are not in the same subnet of the ND Mgmt interfaces

*Connectivity to the APIC's IB addresses is the only supported option when running also NDI on the same ND cluster

Nexus Dashboard Orchestrator High Availability

Distributed ND Cluster Deployment for NDO – Physical ND Form Factor

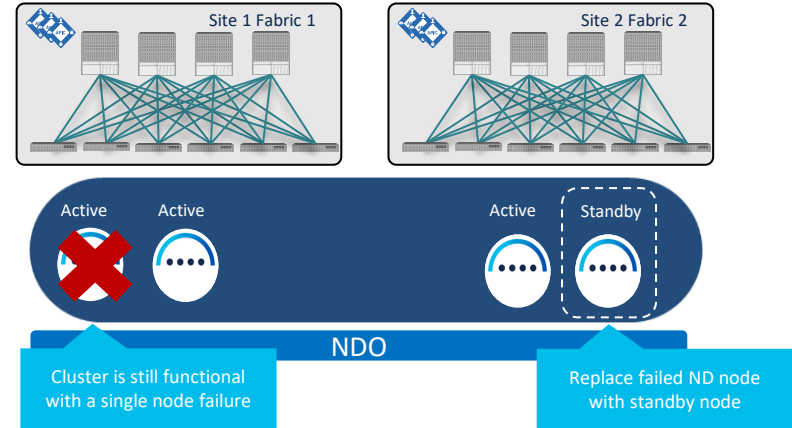
- At least 2 ND active nodes are needed to keep the cluster operational
- Deploy standby node to recover cluster in the event of a failure of primary nodes
- Maximum supported latency values
 - 50 msec RTT: between ND nodes
 - 500 msec RTT: between an ND node and an APIC node
 - 150 msec RTT: between an ND node and switches for Insights service (consideration if cohosting services on distributed cluster)



Nexus Dashboard Orchestrator High Availability

Distributed ND Cluster Deployment for NDO – Physical ND Form Factor

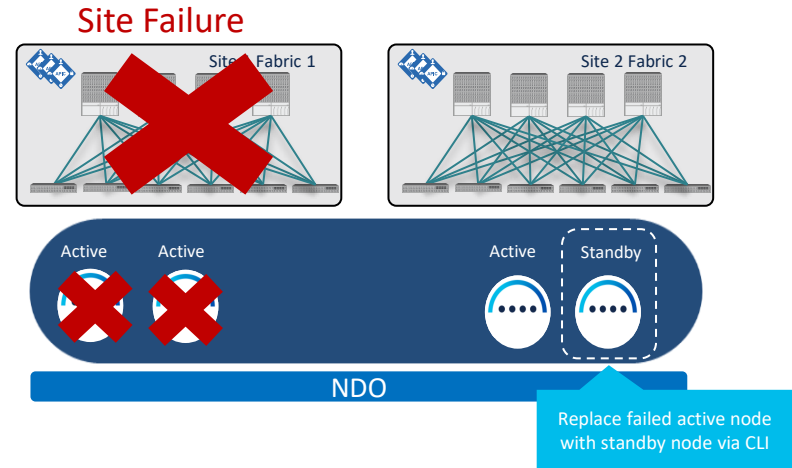
- At least 2 ND active nodes are needed to keep the cluster operational
- In case of failure of a primary node, a standby node can be used to replace the failed node. Can be done via UI.



Nexus Dashboard Orchestrator High Availability

Distributed ND Cluster Deployment for NDO – Physical ND Form Factor

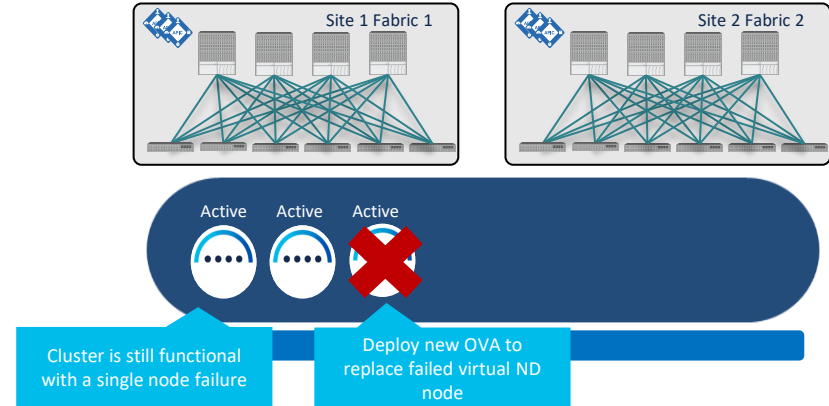
- At least 2 ND active nodes are needed to keep the cluster operational
- In case of concurrent failure of 2 ND active nodes, a standby node can be used to replace a failed primary node. Procedure done via CLI.



Nexus Dashboard Orchestrator High Availability

Centralised ND Cluster Deployment for NDO – Virtual ND Form Factor

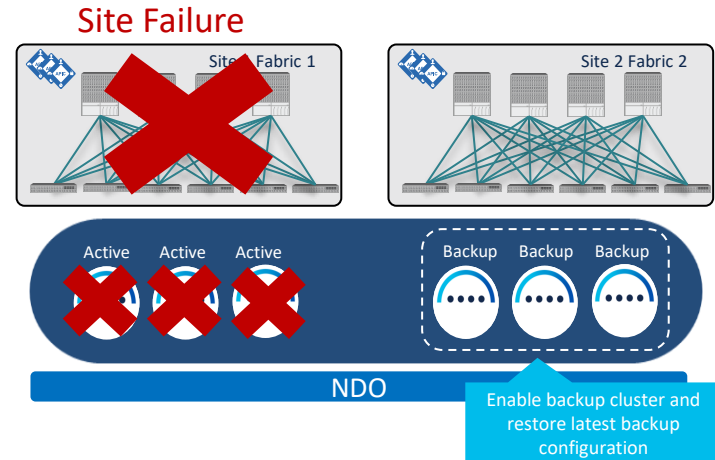
- Only physical form factor supports standby nodes
- Deploy cluster in one site/availability zone (centralized deployment)
- In case of single virtual ND node failure, you can quickly restore a failed node by deploying a new virtual ND node
- Recommended deployment mode for virtual ND cluster. Distributed deployment is possible but adds more complexity without benefit



Nexus Dashboard Orchestrator High Availability

Centralized ND Cluster Deployment for NDO – Virtual ND Form Factor

- Only physical form factor supports standby nodes
- In case of site failure or loss of entire ND cluster, quickly deploy a new virtual ND replacement cluster or cluster can be pre-deployed and ready to replace failed cluster
- Restore configuration from most recent backup*
- To limit the impact of a loss of ND cluster, backup configuration on a regular basis. Store backup files in multiple geographic locations

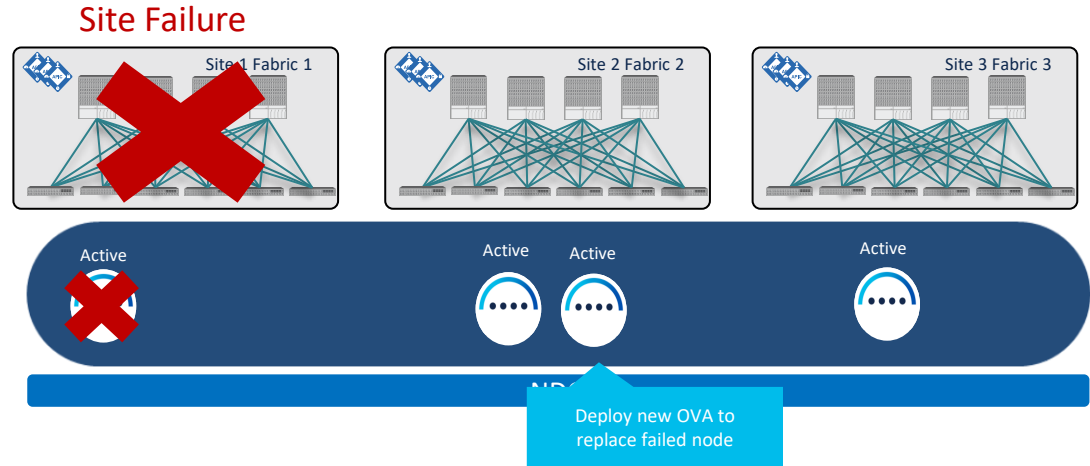


*Separate ND and NDO backups on ND 3.1.1. A single combined backup will be supported from ND 3.2

Nexus Dashboard Orchestrator High Availability

Distributed ND Cluster Deployment for NDO – Virtual ND Form Factor

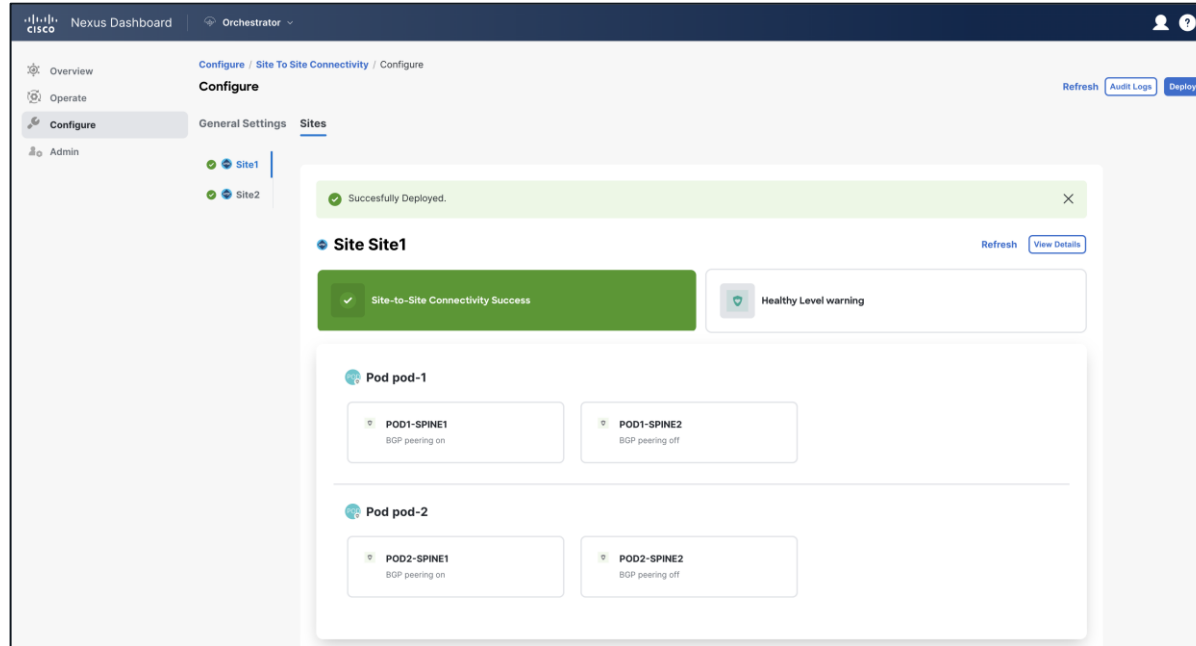
- Distributed deployment can be used for virtual cluster with more than two sites or availability zones
- Recovery from single node or single site failure is possible by replacing failed virtual node
- Backup cluster and recent backup file can be used for more than two node failures



Provisioning Policies on NDO



Provisioning Infra Configuration for the Fabrics



- Provisioning of OSPF/BGP peering between the spine nodes in each fabric and the ISN devices
- Provisioning of full mesh MP-BGP EVPN adjacencies between spines in different fabrics
- Enablement of VXLAN Data Plane (provisioning of O-UTEP and O-MTEP addresses for each fabric)

Supporting Different Types of Templates

The screenshot displays the Cisco Nexus Dashboard Orchestrator interface. The top navigation bar includes the Cisco logo, 'Nexus Dashboard', and 'Orchestrator'. A left-hand sidebar contains navigation links: Overview, Operate, Configure, and Admin. The main content area is titled 'Welcome, admin' and features a 'Global View' tab. Under this tab, a 'Configure' dropdown menu is open, listing 'Configure', 'Site to Site Connectivity', 'Tenant Templates', and 'Fabric Templates'. Two callout boxes provide details: a green-bordered box for 'Tenant Level Configuration' (Applications Policies, Tenant Policies, L3Outs, Service-Stitching) points to 'Tenant Templates', and a blue-bordered box for 'Fabric Management Configuration' (Used to define fabric access policies, interface and monitoring policies) points to 'Fabric Templates'. A small diagram on the right shows two sites, 'Site1 ACI' and 'Site2 ACI', connected by a line.

- Provisioning Tenant level configuration from NDO is mandatory for the VXLAN Multi-Site use case (drives creation of translation entries, etc.)
- Provisioning Fabric level configuration from NDO is advantageous (single pane of glass) but optional

Provisioning Policies on NDO

Multiple Template Types

Tenant Level Configuration

Optimised in
4.0(1)

New in 4.0(1)

New in 4.1(1)

New in 4.2(1)

EPGs, BDs, VRFs,
Contracts, etc.

Route-Maps, DHCP-
Relay, QoS...

Full L3Out
Configuration

Full Service-Stitching
Configuration (node, PBR
policy, ...)

Applications

Tenant Policies

L3Out

Service Device

Fabric Management Configuration

New in 4.0(1)

New in 4.0(1)

New in 4.0(1)

VLAN pools, Physical
Domains, NTP, etc.

Interfaces, SyncE,
MACsec, etc.

SPAN Tenant and
Access Sessions

Fabric Policies

Fabric Resources
Policies

Monitoring
Policies

Benefits

Simplify

Single Pane of Glass

Application Templates

Multi-Site Templates

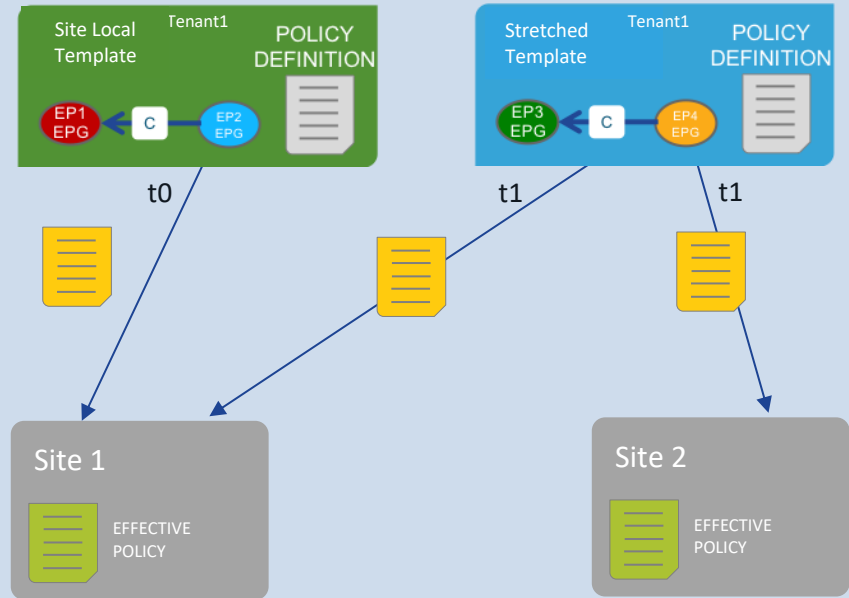
- Application Template = ACI policy definition (ANP, EPGs, BDs, VRFs, etc.)
- Schema = container of Application Templates sharing a common use-case
 - As a typical use case, a schema can (and should) be dedicated to a Tenant
- The template is the atomic unit of change for policies
 - A Multi-Site template associated to a single site can be pushed only to that site
 - A Multi-Site template associated to multiple sites is concurrently pushed to all those sites

Deployment Mode ⓘ

☒ Multi-Site

☐ Autonomous

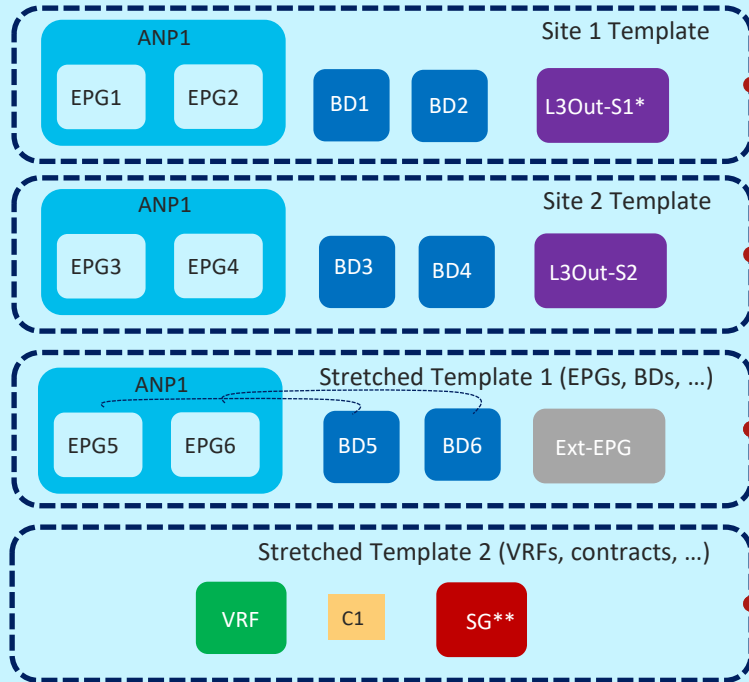
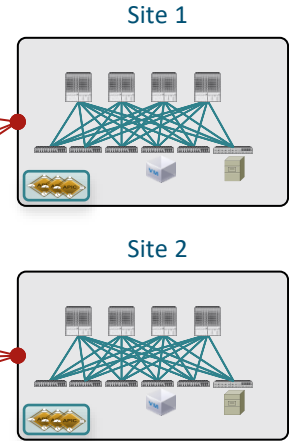
Schema



Best Practices for Multi-Site Templates

One Template per Site, plus Two Templates for “Stretched Objects”

Schema (dedicated to Tenant1)



*L3Out defined in a separate “L3Out Template” from NDO 4.1(1)

**Service-Graph implicitly created using Service Device Templates from NDO 4.2(3)

Application Templates

Autonomous Templates

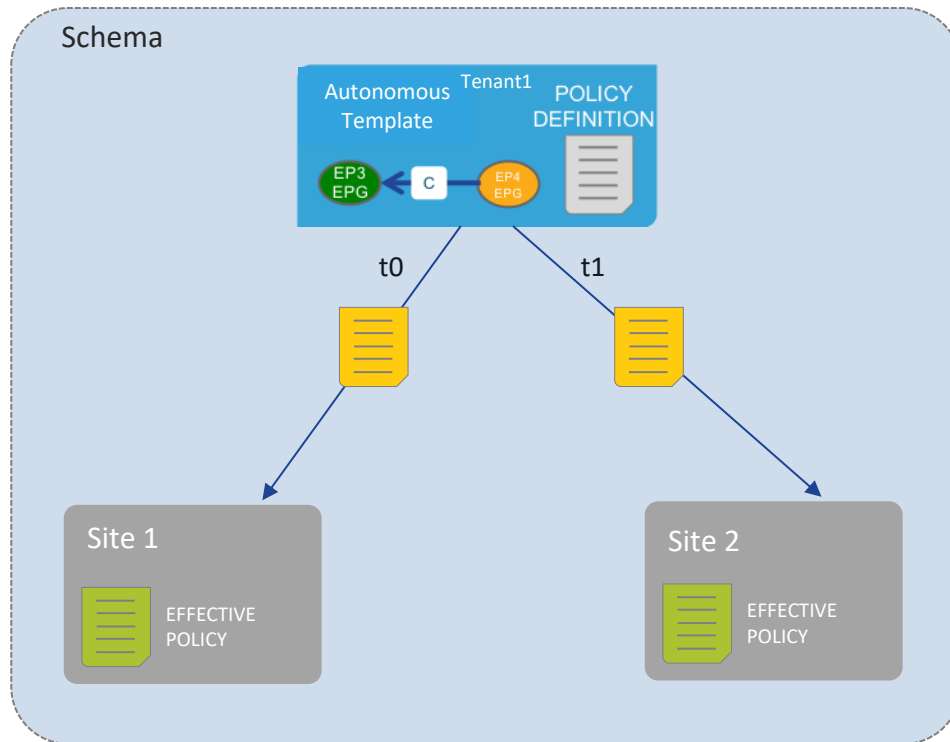
Deployment Mode ⓘ

○ Multi-Site

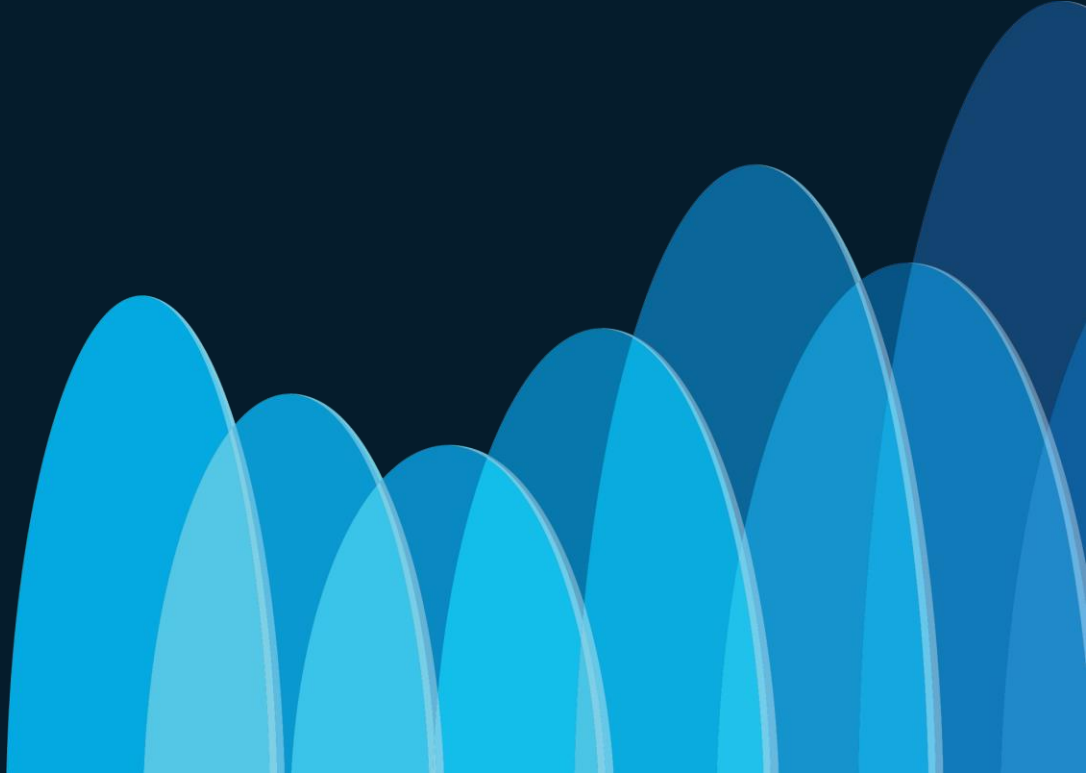
● Autonomous

- Autonomous templates can also be associated to one or more fabrics
- Differently than for Multi-Site templates, the deployment of an Autonomous template to different sites won't cause the “stretching” of configuration objects (VRFs, BDs, EPGs,...)
- NDO performs a “configuration replication” function to multiple sites
 - Site level configuration customisation is possible (BD subnet, VMM domain association, etc.)
- Autonomous Templates can be deployed to different fabrics at different points in time*

Schema

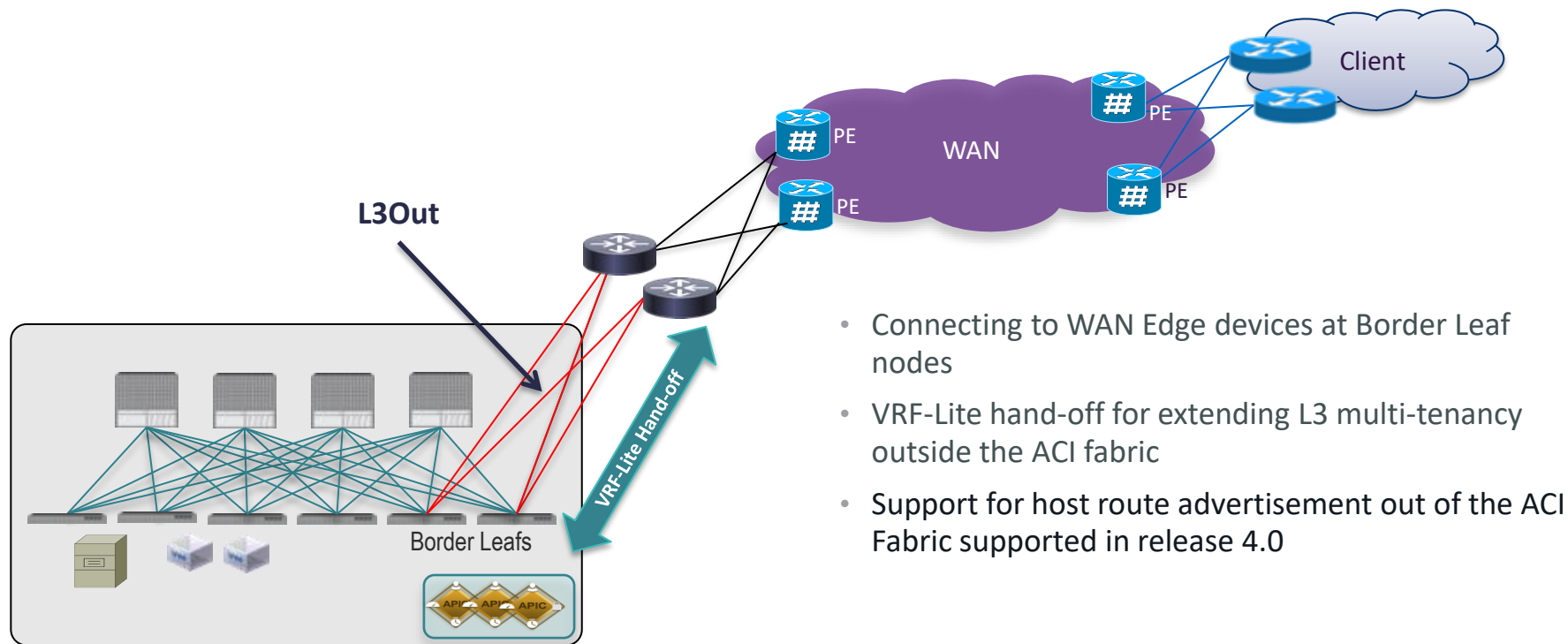


Connecting to the External L3 Domain



Connecting ACI to Layer 3 Domain

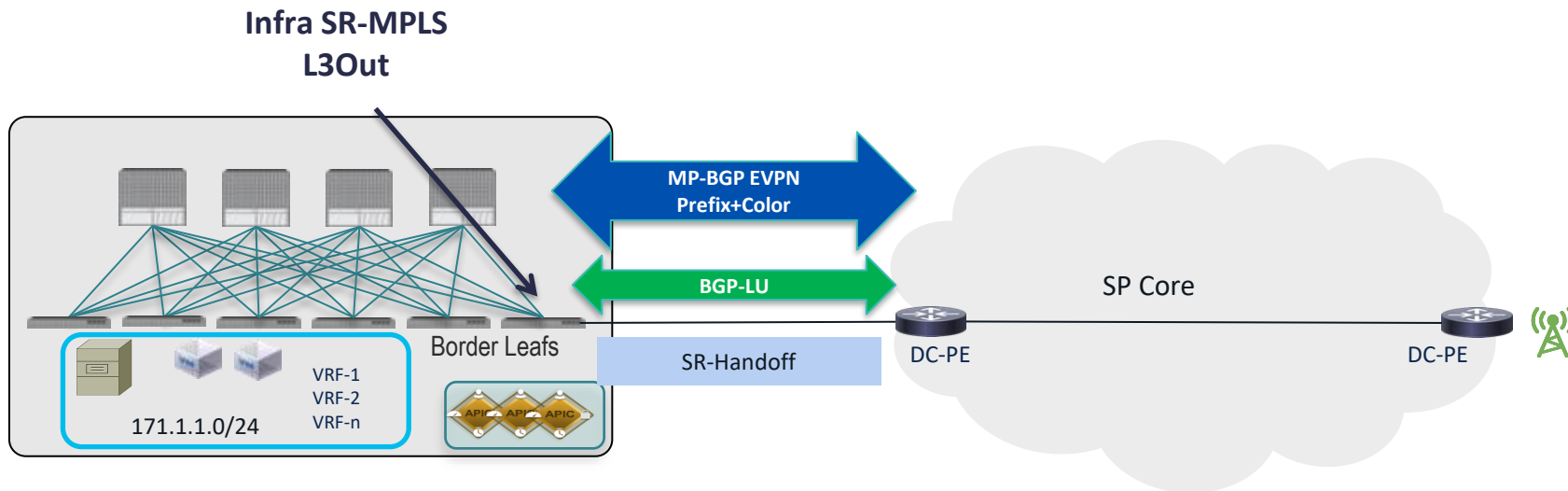
'Traditional' L3Out on the BL Nodes



Connecting ACI to Layer 3 Domain

'SR-MPLS Handoff'

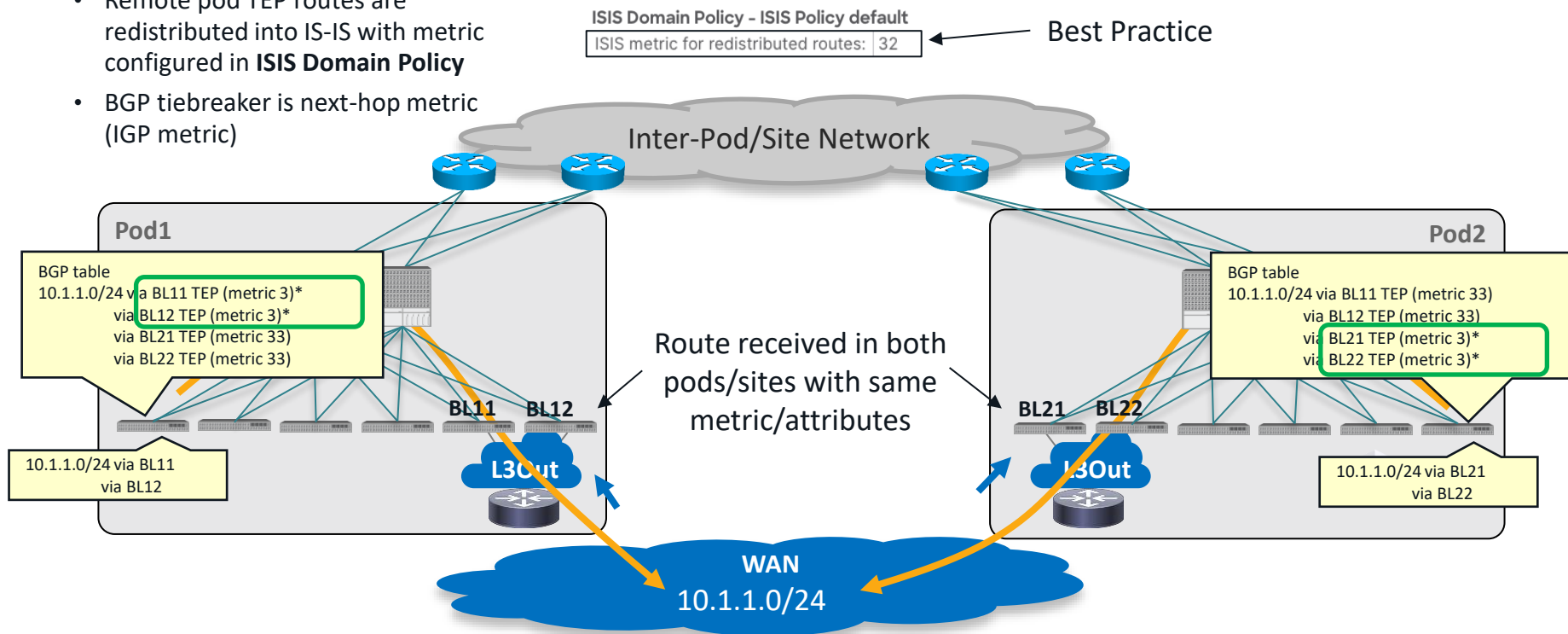
- Border Leafs connect to PE router in SP core
- Single BGP EVPN session for all VRFs
- ACI BL is advertising EVPN type-5 routes with BGP color community



Connecting to the External L3 Domain

Local L3Outs preferred over L3Outs in remote pods/sites

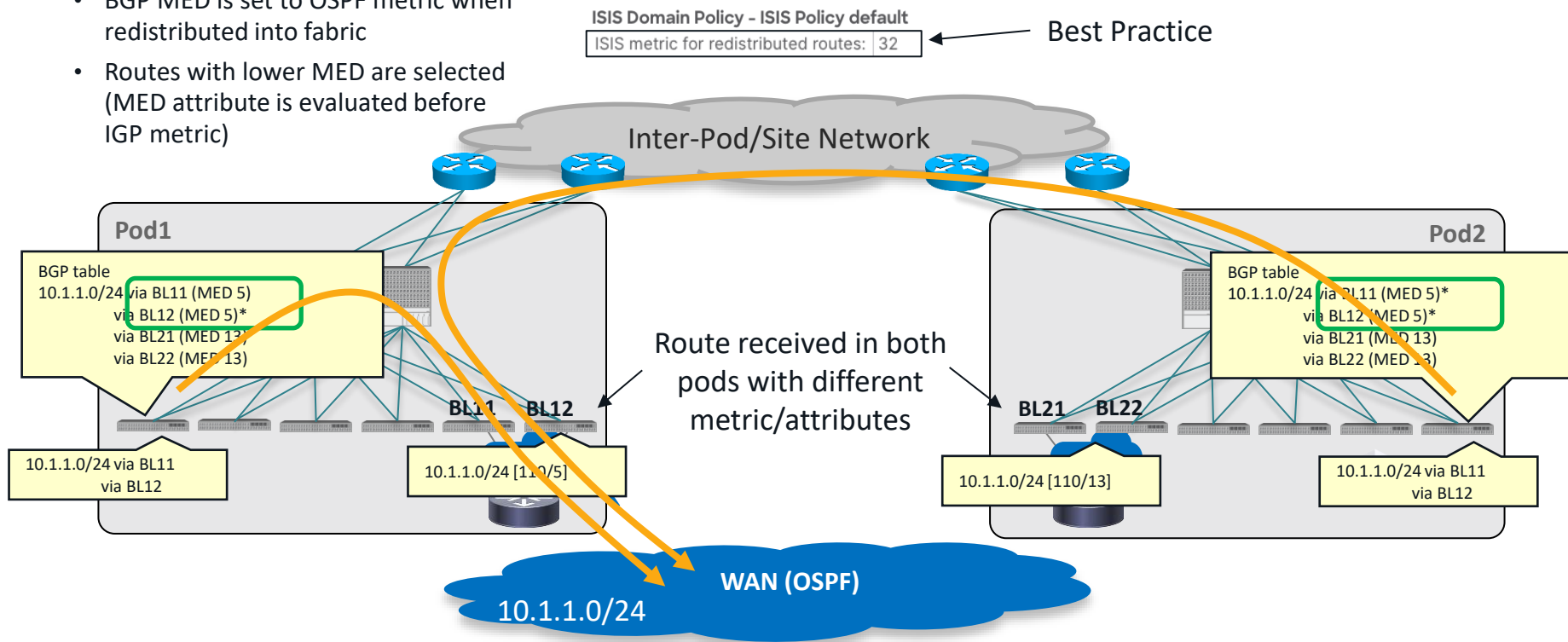
- Remote pod TEP routes are redistributed into IS-IS with metric configured in **ISIS Domain Policy**
- BGP tiebreaker is next-hop metric (IGP metric)



Connecting to the External L3 Domain

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

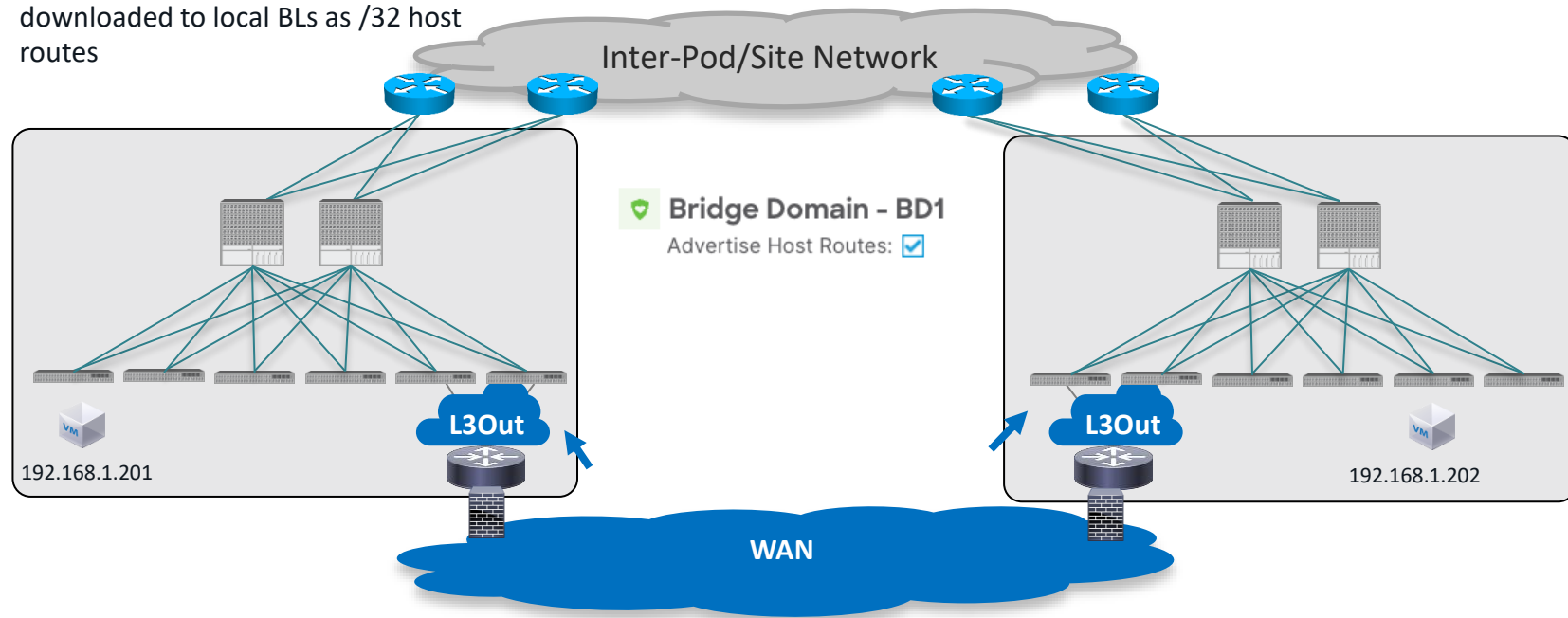
- BGP MED is set to OSPF metric when redistributed into fabric
- Routes with lower MED are selected (MED attribute is evaluated before IGP metric)



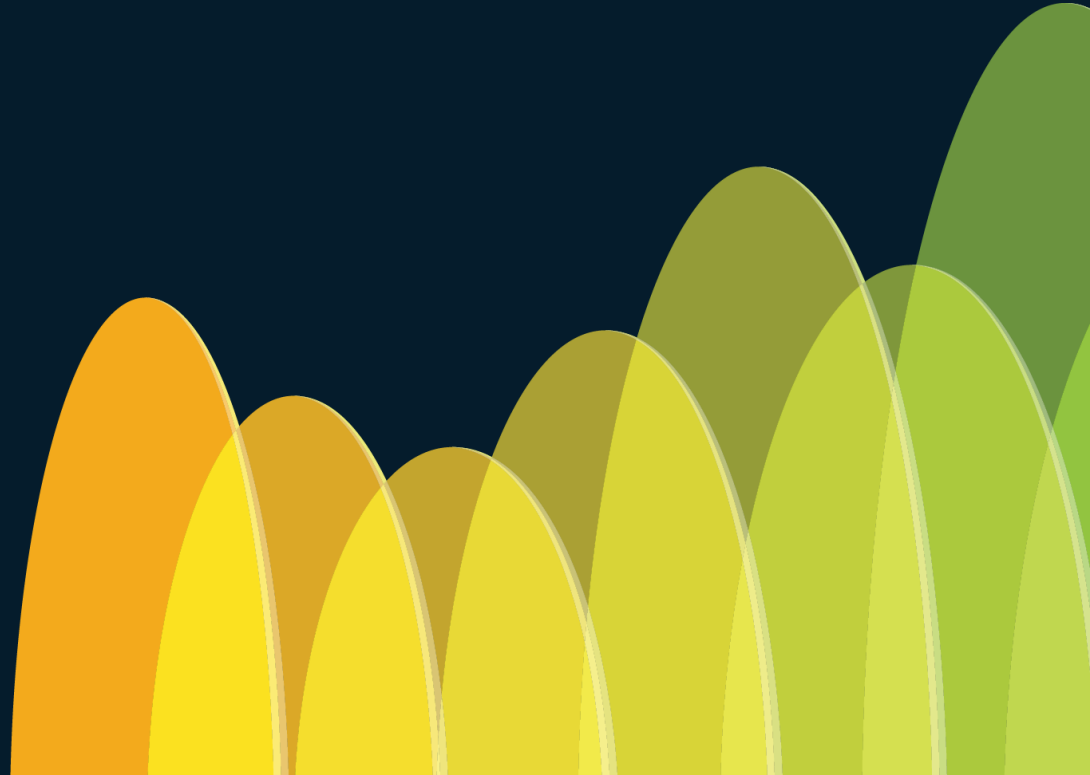
Connecting to the External L3 Domain

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

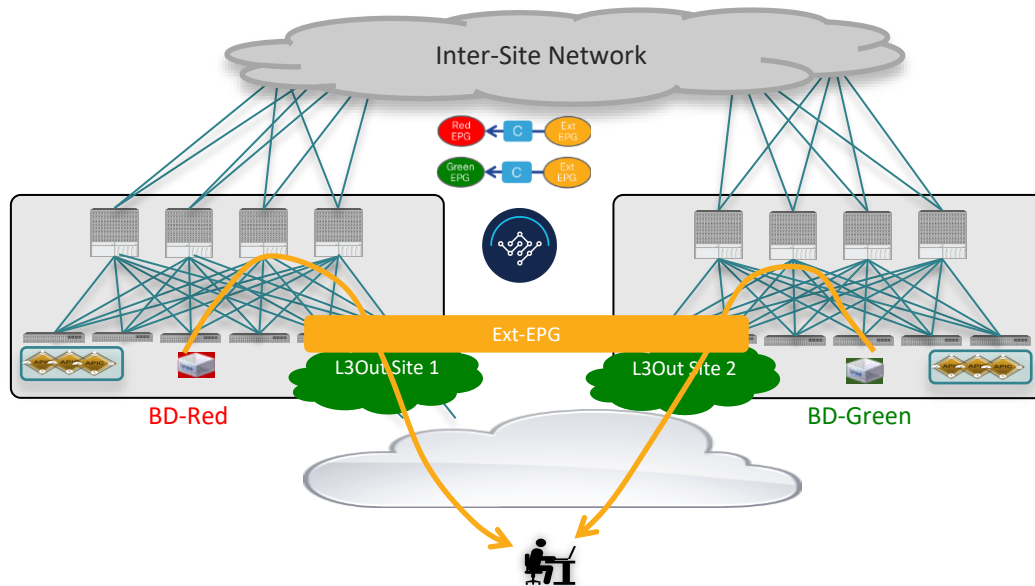


Multi-Site Considerations for External L3 Domains



ACI Multi-Site and L3Out

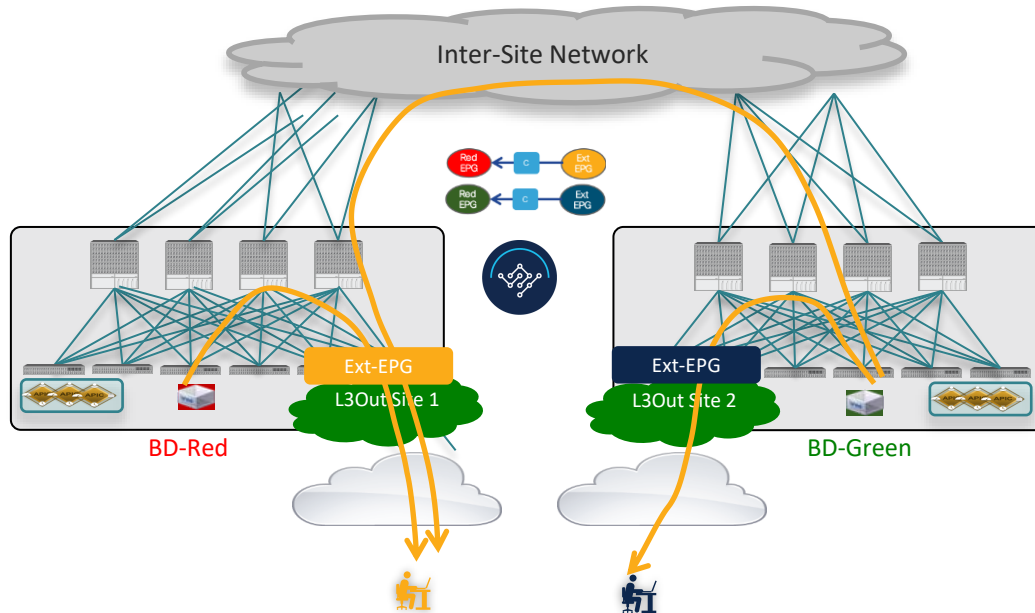
Stretching or Not Stretching the Ext-EPG?



- 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)

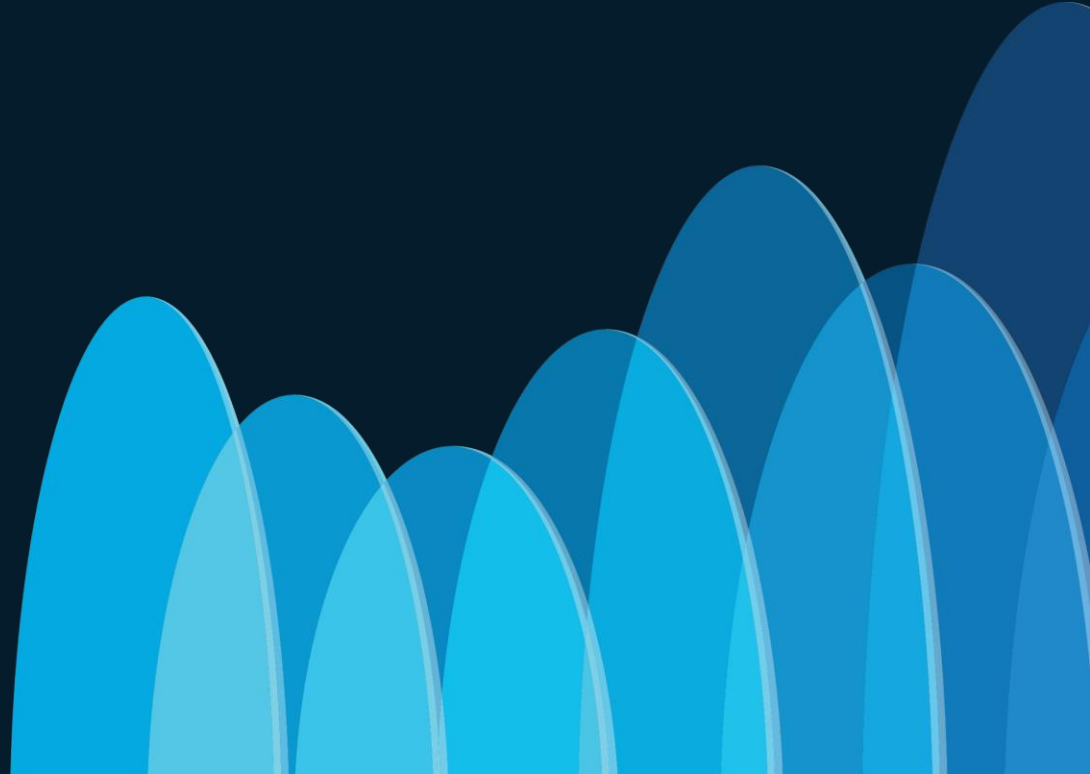
ACI Multi-Site and L3Out

Stretching or Not Stretching the Ext-EPG?



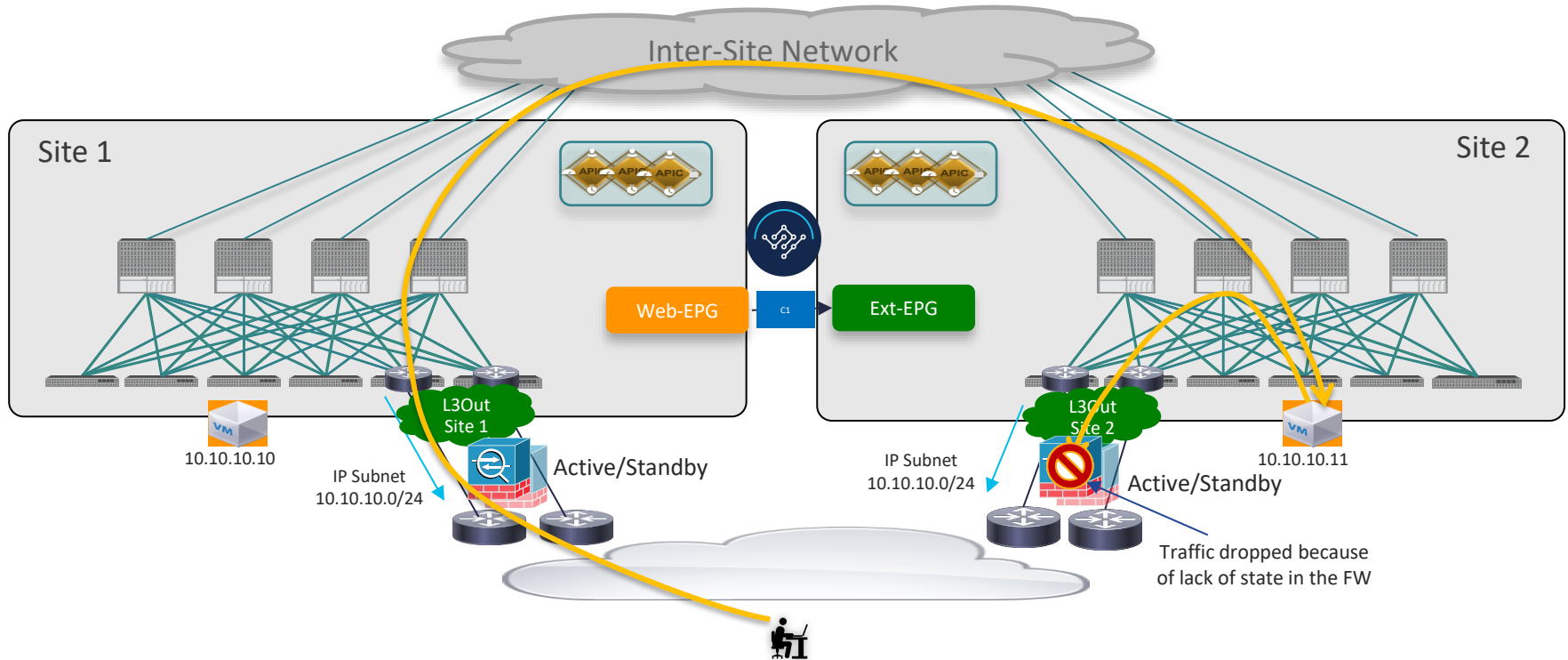
- Separate Ext-EPGs can be defined in templates mapped to separate sites (non stretched objects)
 - Each Ext-EPG can be mapped to the local L3Out in the “global” or “site level” section of the template configuration
- Allows to apply different policies to each Ext-EPGs at different time
- Can still use the same 0.0.0.0/0 network configuration for classification on both sites
- May require enablement of Intersite L3Out

Solving Asymmetric Routing Issues with the External Network



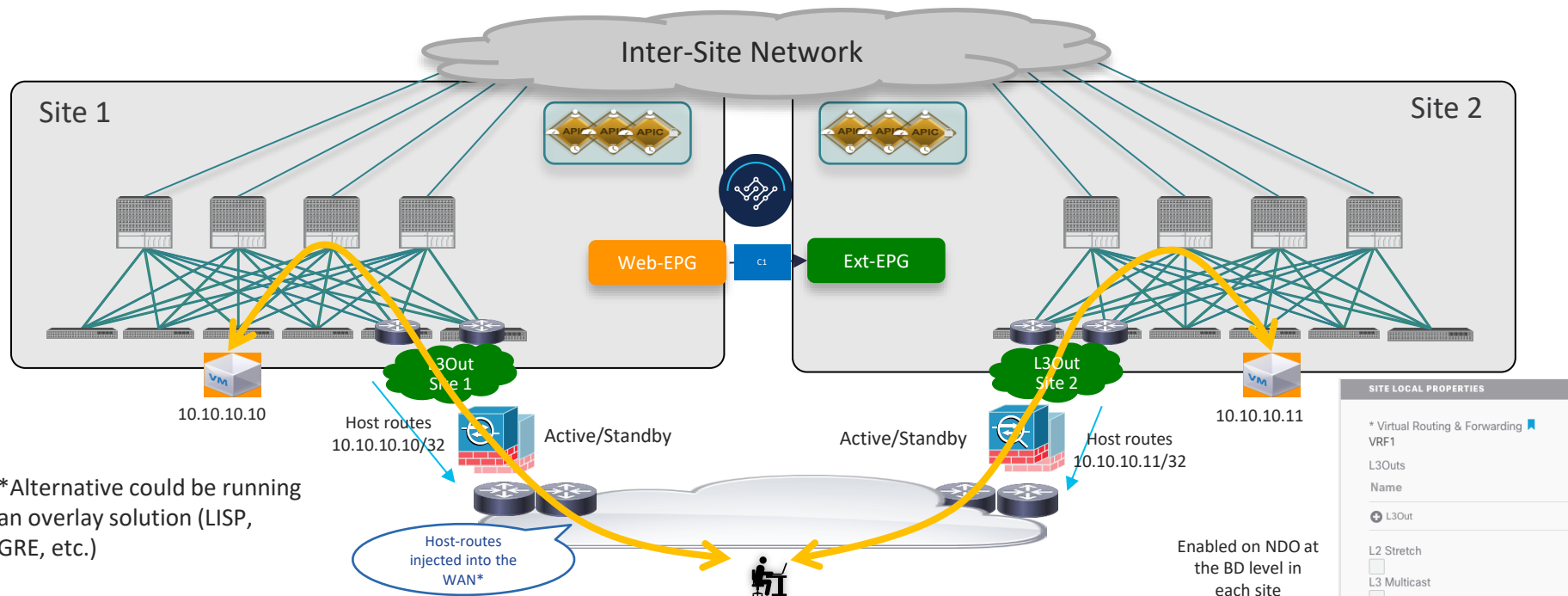
ACI Multi-Site and L3Out

Typical Deployment of Perimeter FWs



Solving Asymmetric Routing Issues

Use of Host-Routes Advertisement

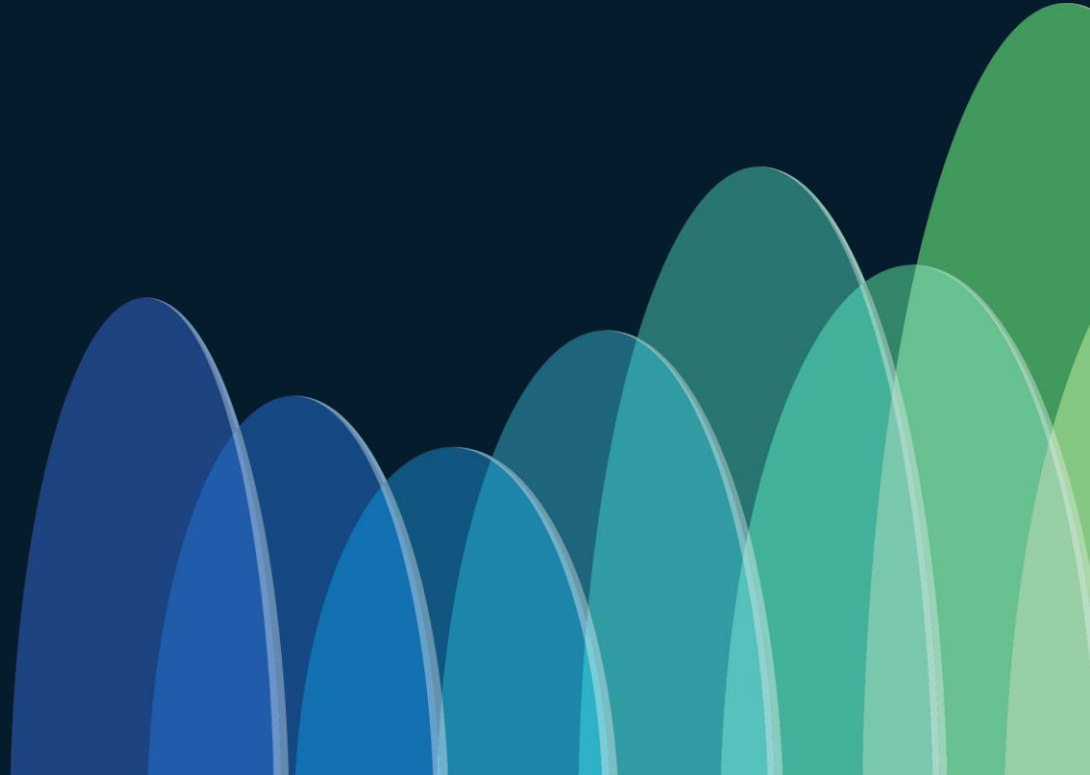


*Alternative could be running an overlay solution (LISP, GRE, etc.)

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



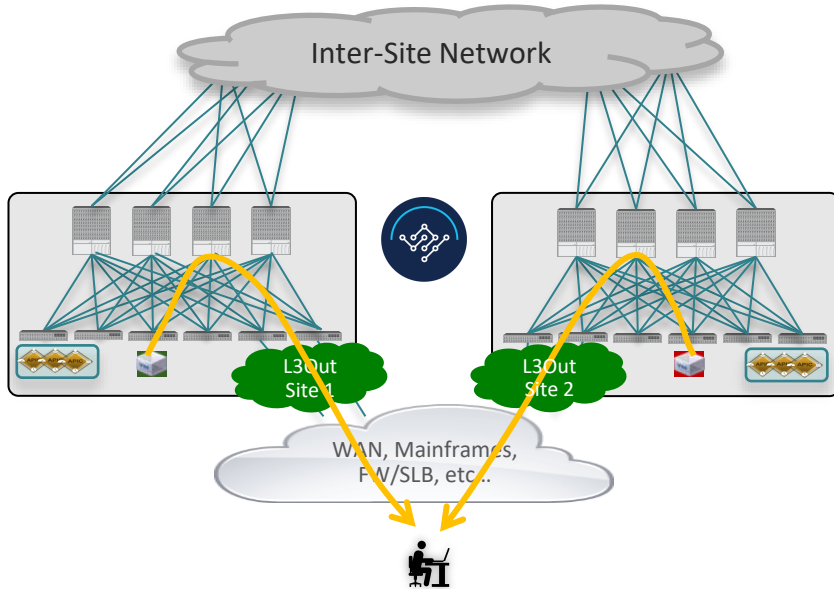
Intersite L3Out Support



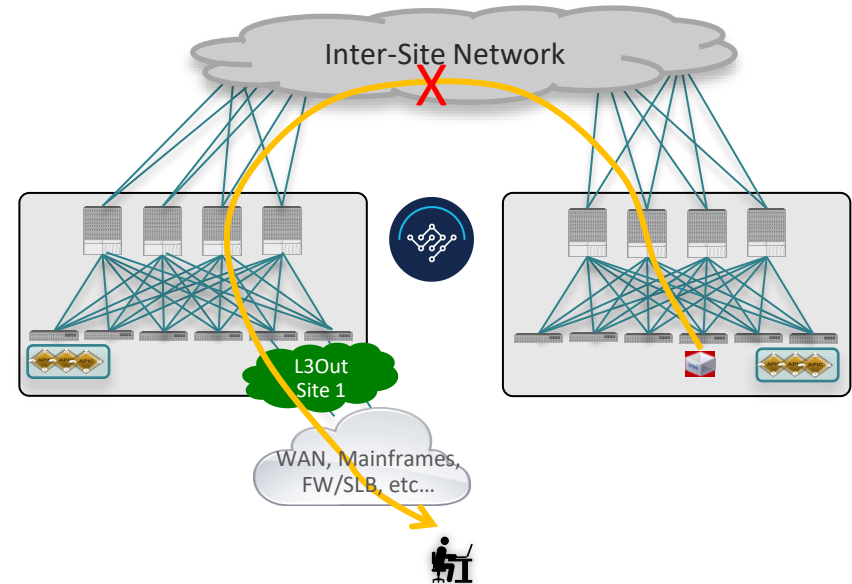
Problem Statement

Behavior before ACI Release 4.2(1)

Supported Design



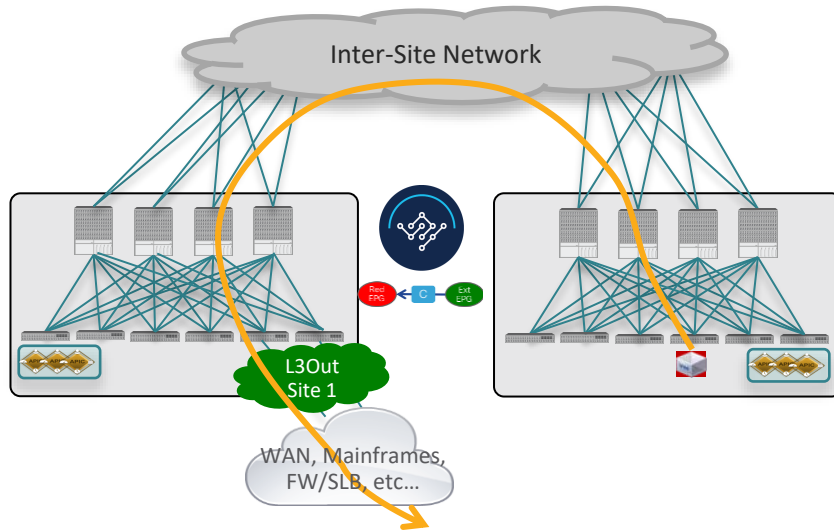
Not Supported Design



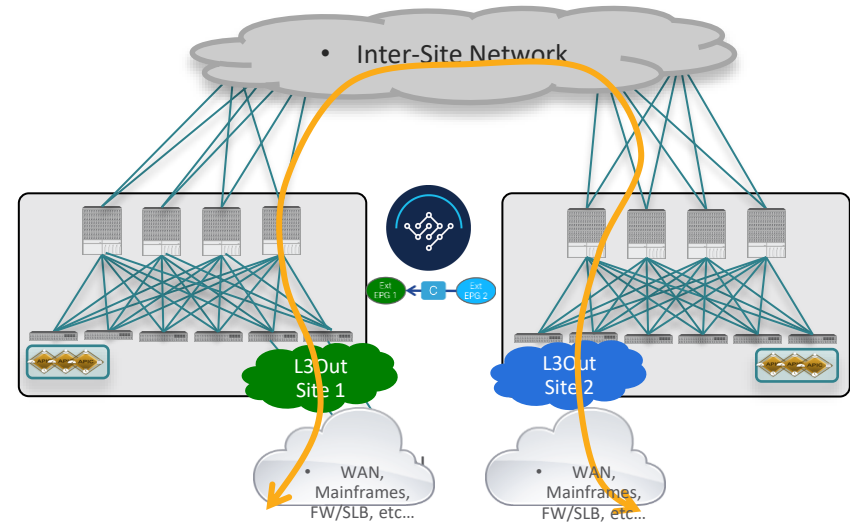
Note: the same consideration applies to both IP-Based L3Outs and SR-MPLS L3Outs

ACI Multi-Site and Intersite L3Out

Supported Scenarios



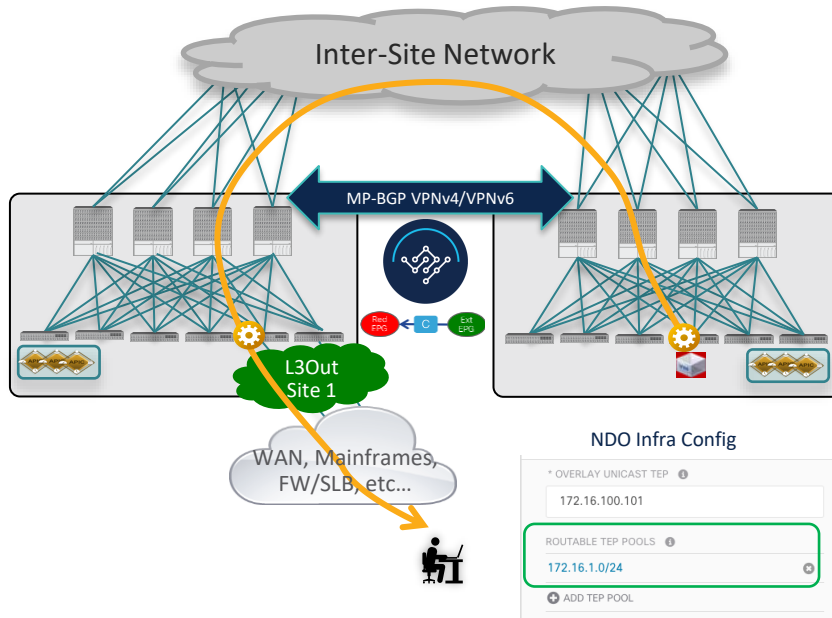
- Endpoint to remote L3Out communication (intra-VRF)
- Endpoint to remote L3Out communication (inter-VRF)



- Inter-site transit routing (intra-VRF)
- Inter-site transit routing (inter-VRF)

ACI Multi-Site and L3Out

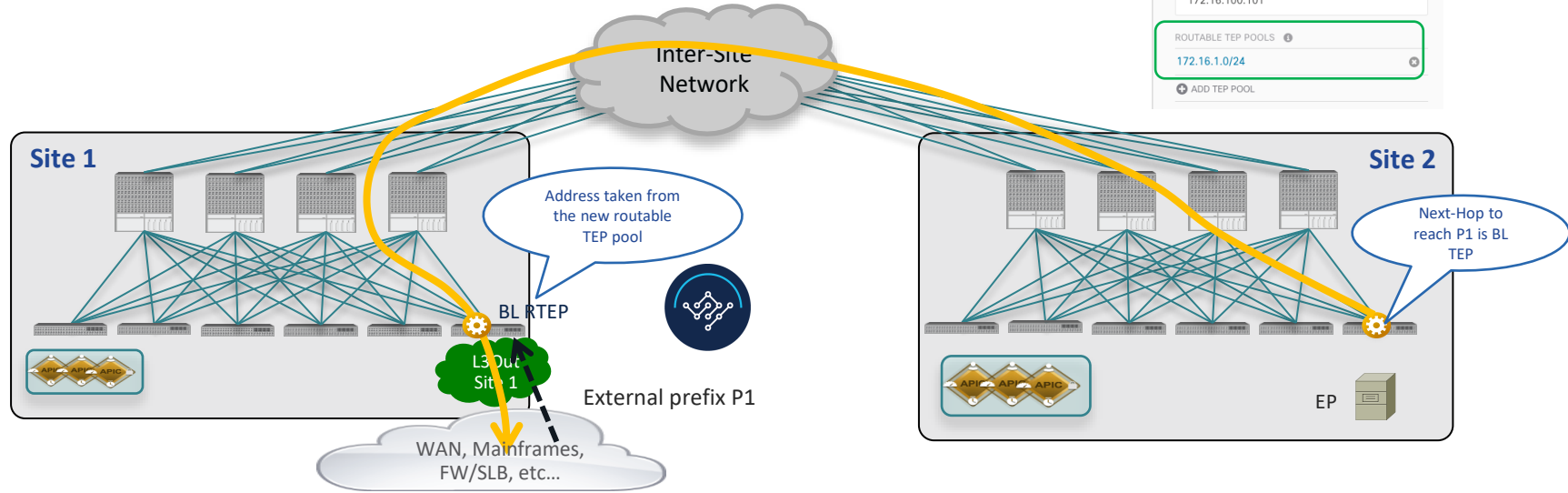
Support of Intersite L3Out



- Starting with ACI Release 4.2(1) it is possible for endpoints in a site to send traffic to resources (WAN, Mainframes, FWs/SLBs, etc.) accessible via a remote L3Out connection
- External prefixes are exchanged across sites via MP-BGP VPNv4/VPNv6 sessions between spines
- Traffic will be 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 will also support transit routing across sites (L3Out to L3Out)

Enabling Intersite L3Out

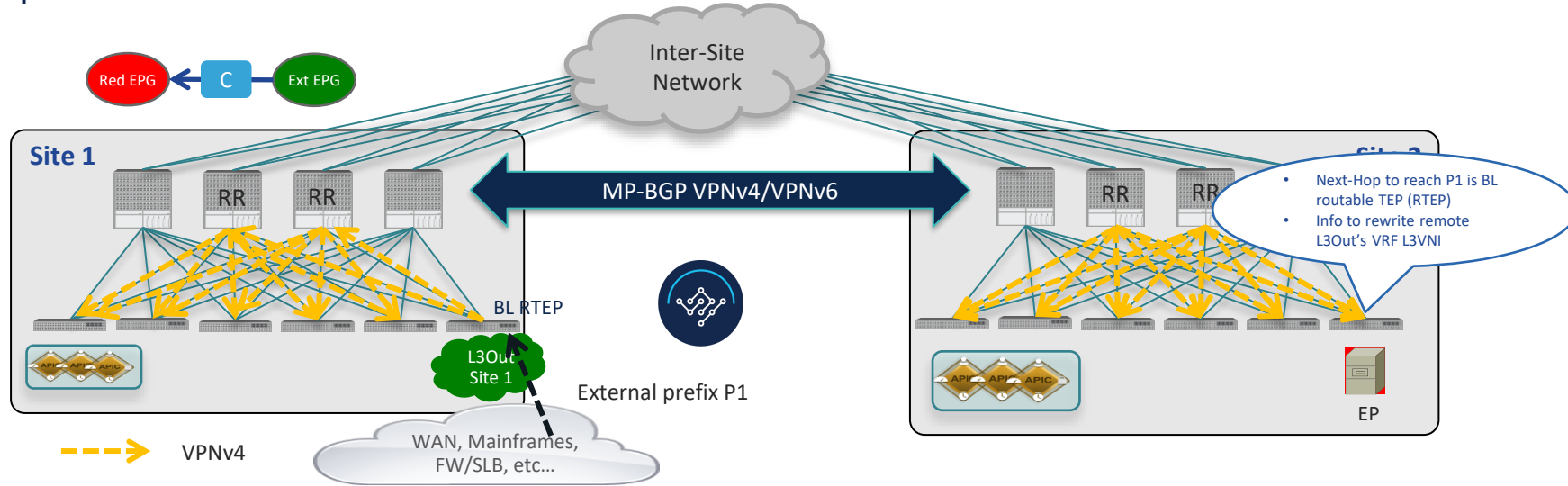
Step 1 - Provisioning of a Routable TEP Pool



- The BL TEP is normally taken from the original TEP pool assigned during the fabric bring-up procedure
- Since we don't want to assume that the original TEP pool can be reached across the ISN, a separate routable TEP pool is introduced to support intersite L3Out
 - The routable TEP pool can be directly configured on NDO
 - One or more routable TEP pools can be configured (pool size is /22 to /29), even at different times

Enabling Intersite L3Out

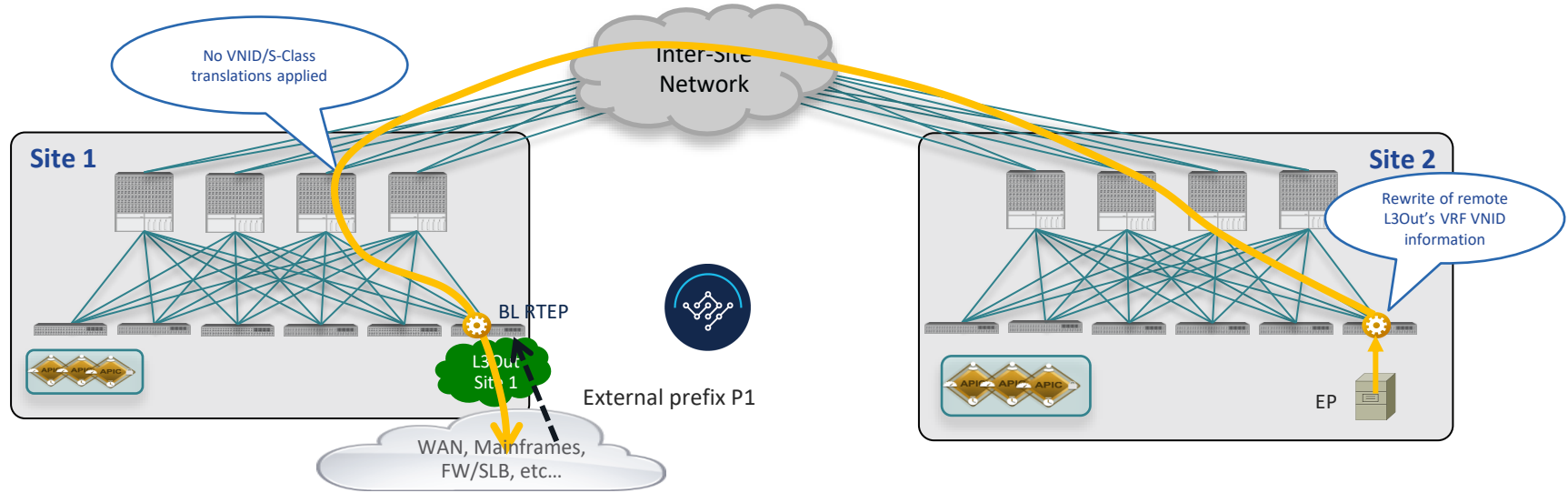
Step 2 – Provision a Contract for North-South Communication



- External prefix advertisements received via the L3Out are redistributed to the leaf nodes in the local site via MP-BGP VPNv4/VPNv6 through the RRs in the spines (normal ACI intra-fabric behavior)
- MP-BGP VPNv4 advertisements are also used to distribute this information to the remote sites
- The prefixes are then redistributed inside the remote sites via VPNv4/VPNv6 by the RR spines
 - The next-hop VTEP for the prefixes is the BL routable TEP (RTEP) that received the routes from the external network
 - Associated to the prefix information are the info to rewrite the VRF L3VNI value to match the one in the remote site

ACI Multi-Site and Intersite L3Out

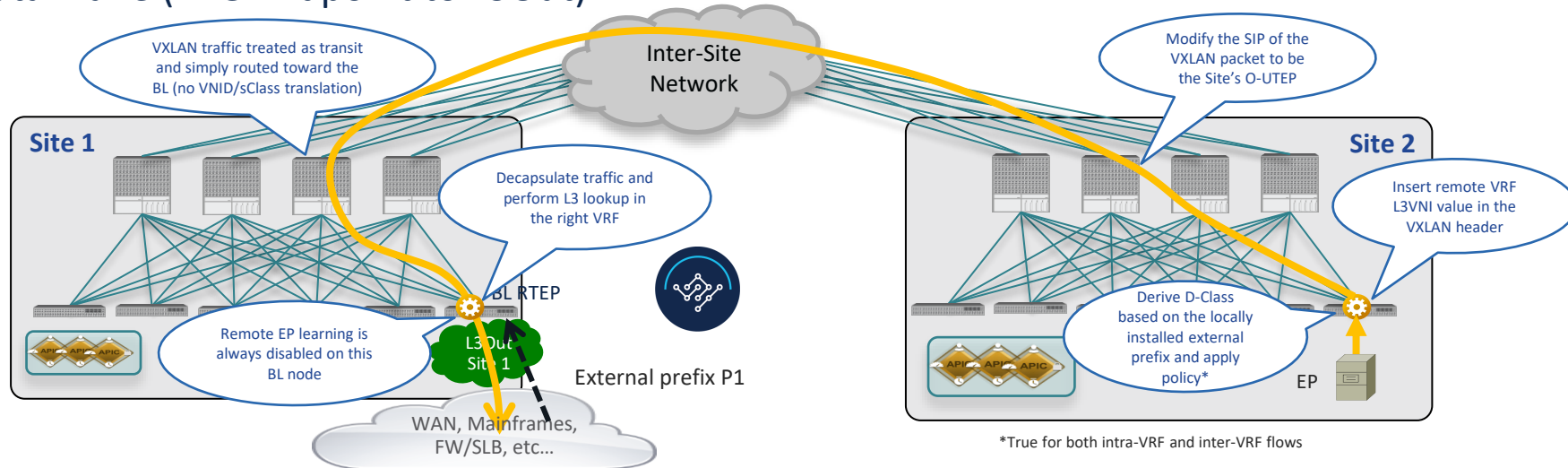
No VNID/S-Class Translations on Receiving Spines



- Differently from regular inter-site (east-west) communication between endpoints, the spines on the receiving site are not involved in VNID/Class-ID translations for communication to external prefixes
- BGP program the rewrite of VNIDs of remote site routes directly on ToRs

ACI Multi-Site and Intersite L3Out

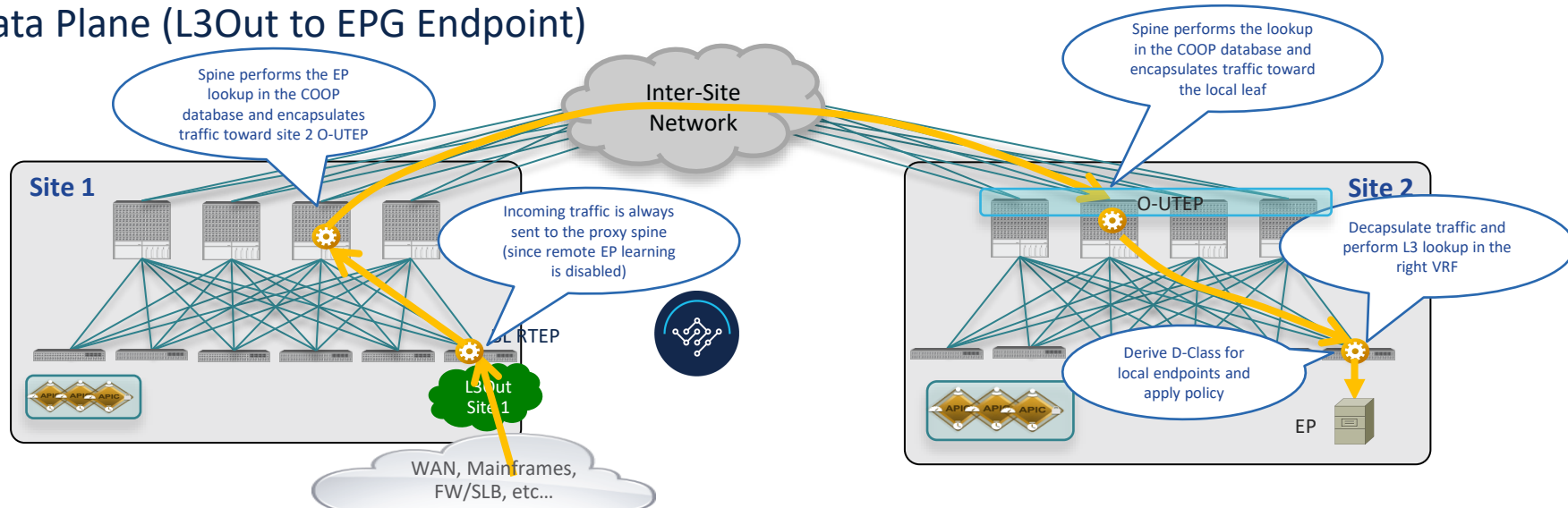
Data Plane (EPG Endpoint to L3Out)



- VXLAN tunnel is established directly between the leaf in site 2 and the BL in site 1
 - The spines in the source site still translate the SIP to be the site's O-UTEP
 - The spines in the destination site simply route the VXLAN packet toward the destination BL node
- The BL node uses the L3VNI info in the VXLAN header to perform the lookup in the right VRF
- Remote endpoint learning always disabled on the BL nodes to avoid learning the wrong sClass info (as there is no translation happening on the receiving spines)

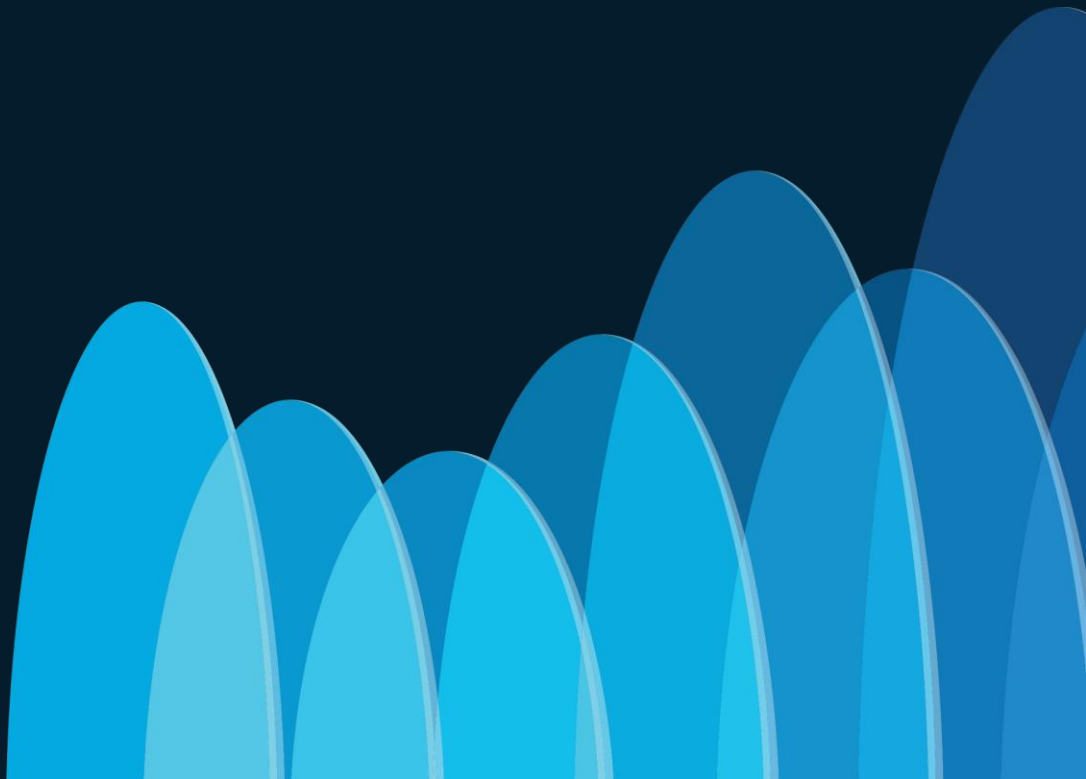
ACI Multi-Site and Intersite L3Out

Data Plane (L3Out to EPG Endpoint)



- Traffic received from the WAN hit the BL node and is always sent to the spine proxy
- The local spine performs the lookup for the destination endpoint and encapsulates to the O-UTEP of the destination site (after changing the SIP in the VXLAN header to match the local O-UTEP)
- The receiving spine performs the lookup in the COOP DB and S-Class/VNID translations (as in regular Multi-Site data-plane) → the traffic is encapsulated to the local leaf
- The local leaf decapsulates the packet, performs the L3 lookup, applies the policy and sends traffic to the endpoint (if allowed)

Intersite L3Out Deployment Considerations



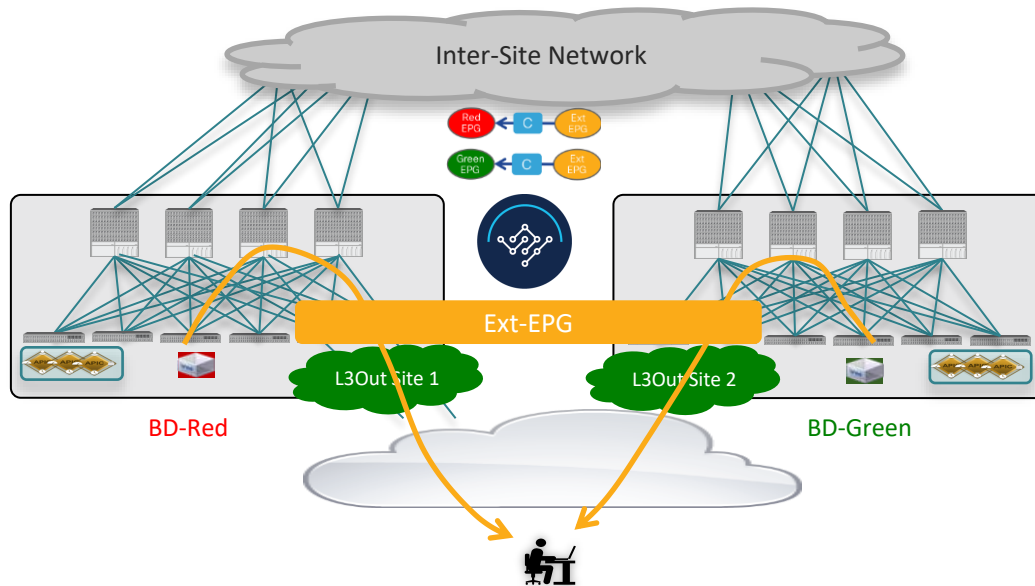
ACI Multi-Site and Intersite L3Out

Deployment Considerations

- Before ACI release 4.2(1), the outbound and inbound traffic flows take always a deterministic path
 - For BDs that are only locally defined in a site (i.e. not stretched), outbound communication is only possible via local L3Outs
 - Inbound communication is also only possible via the local L3Out, as it is not possible to advertise the BD subnet(s) out of a remote L3Out connection
 - For stretched BDs, the option of enabling host-based routing advertisement has been made available from ACI release 4.0(1) to ensure inbound flows take always an optimal path
- The enablement of the Inter-site L3Out functionality may change this behavior, so it is important to keep in mind some specific deployment considerations discussed in the following slide

ACI Multi-Site and Intersite L3Out

Stretched Ext-EPG - Outbound Flows



- When peering BGP with the external router, if all BGP parameters are the same, the local L3Out is preferred by default for outbound flows

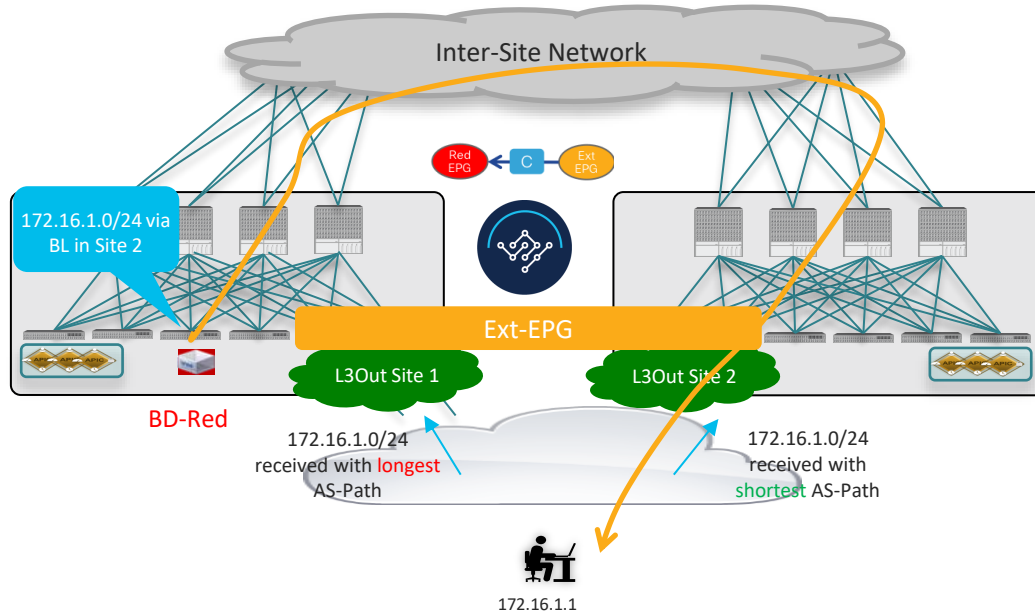
The default behavior can be modified by tuning a BGP parameter (Local-Preference, MED, ...) for the received external prefixes

Notice that Local-Preference is only propagated inside a BGS AS, so can't be used if separate sites are peering EBGP

- When peering OSPF or EIGRP with the external router, local L3Out is preferred by default (best IS-IS metric to the local BL nodes)

ACI Multi-Site and Intersite L3Out

Stretched Ext-EPG - Outbound Flows

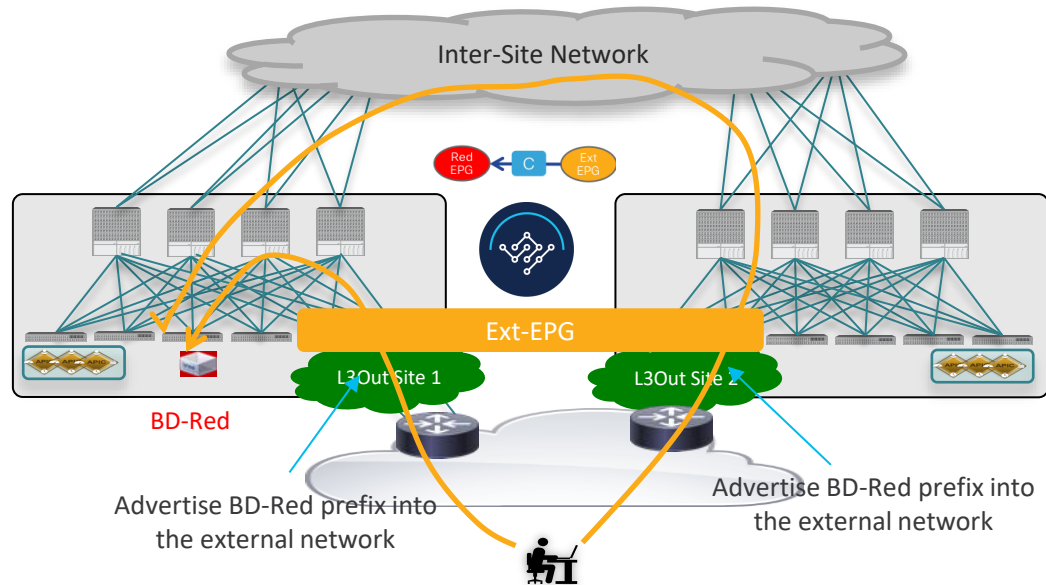


- When peering BGP with the external router, if all BGP parameters are the same, local L3Out is preferred by default
- This is not the case if the BGP attributes (like AS-Path for example) are better for a prefix received in a given site
- A user may not be able to modify this behavior by applying inbound route-map on the BL nodes and would lose control on outbound traffic path

AS-Path is considered before Local-Preference and MED in the BGP route selection algorithm

ACI Multi-Site and Intersite L3Out

Stretched Ext-EPG - Outbound Flows



- If Intersite L3Out is enabled for the WAN isolation use case, it is required to advertise the BD Red subnet out of the local and remote L3Outs
- Without additional tuning it may happen that inbound traffic is steered toward the site where the BD is not deployed
 - This may cause asymmetric traffic path also for not stretched BD
- Enabling host-based advertisement is a possible solution, if there are not scalability concerns on the amount of host routes advertised externally
- Policies can also be applied on the L3Out of each site to make one path preferable compared to the other

ACI Multi-Site and Intersite L3Out

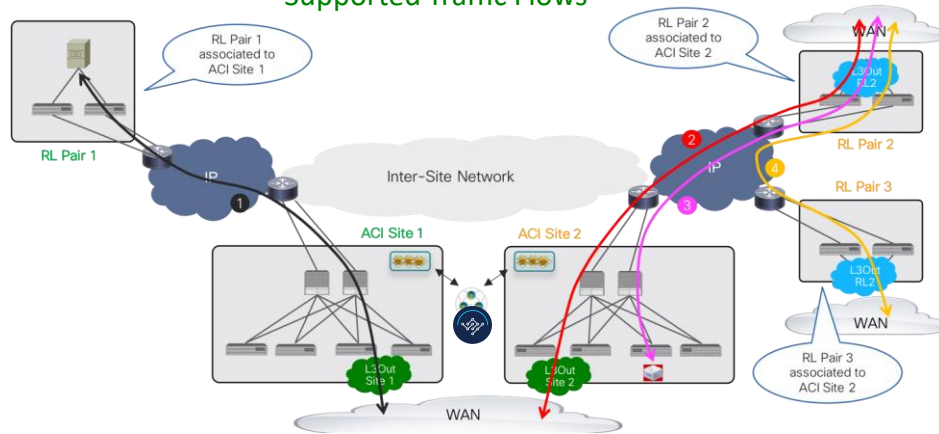
Current Restrictions

- The following restrictions currently apply to the Inter-Site L3Out functionality:
 - Not supported when deploying GOLF L3Outs: this implies that a local GOLF L3Out connection is always required to enable outbound connectivity for endpoints deployed in a given site
 - Not supported between Remote Leaf pairs associated to separate sites
 - Not supported for some specific L3 multicast deployment scenarios (external source with PIM ASM/SSM or external receiver with PIM ASM)
 - Use of a remote L3Out connection after redirecting traffic through a local service node via PBR is not supported as of ACI 5.0(1) release
 - Currently CloudSec and the configuration of Routable TEP pools are mutually exclusive
 - No support for Remote Leaf and/or Intersite L3Out deployments

ACI Multi-Site and Intersite L3Out

Integration with Remote Leaf Nodes

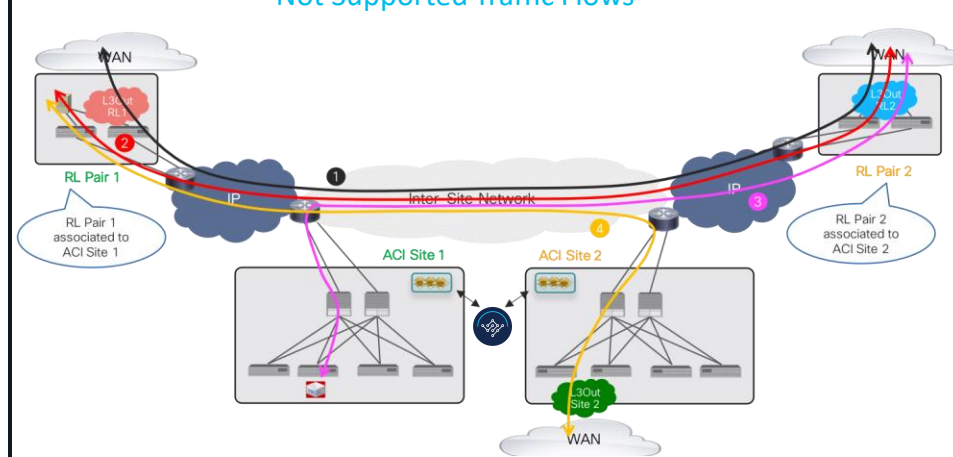
Supported Traffic Flows



- The traffic flows working with ACI release 4.1(2) will continue to be supported in 4.2(1)

- Endpoint connected to the RL pair associated to a site communicating with L3Out(s) deployed in the same site
- Transit routing between L3Outs deployed in the main site and the RL pair associated to the same site
- Endpoint connected to the main site communicating with the L3Out deployed on the RL pair associated to the same site
- Transit routing between L3Outs deployed on RL pairs associated to the same site

Not Supported Traffic Flows

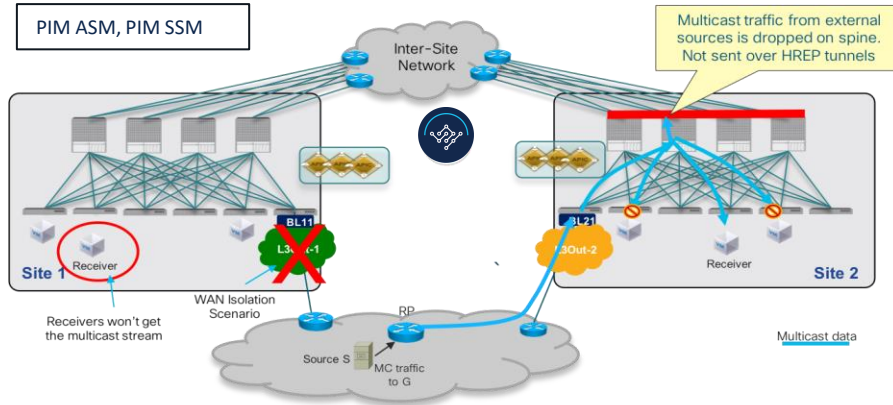


- The following traffic flows are not supported in ACI release 4.2(1)

- Transit routing between L3Outs deployed on RL pairs associated to separate sites
- Endpoint connected to a RL pair associated to a site communicating with the L3Out deployed on the RL pair associated to a remote site
- Endpoint connected to the local site communicating with the L3Out deployed on the RL pair associated to a remote site
- Endpoint connected to a RL pair associated to a site communicating with the L3Out deployed on a remote site

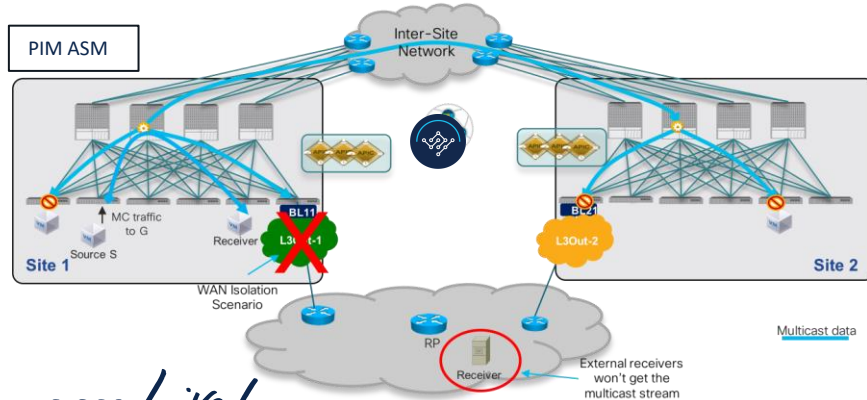
ACI Multi-Site and Intersite L3Out

Integration with L3 Multicast – Not Supported Scenarios



- Each site with local receivers must receive on a local L3Out connection the multicast stream originated by an external source

Multicast traffic received in a site from an external source will never be sent toward remote site via the Inter-Site network

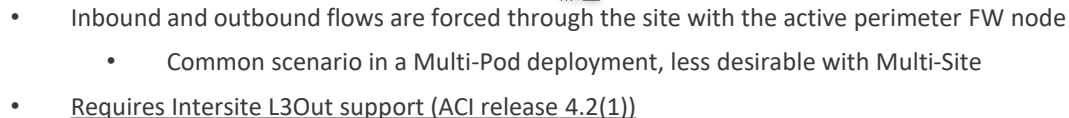


- The site with the local source must have connectivity to the external RP via a local L3Out

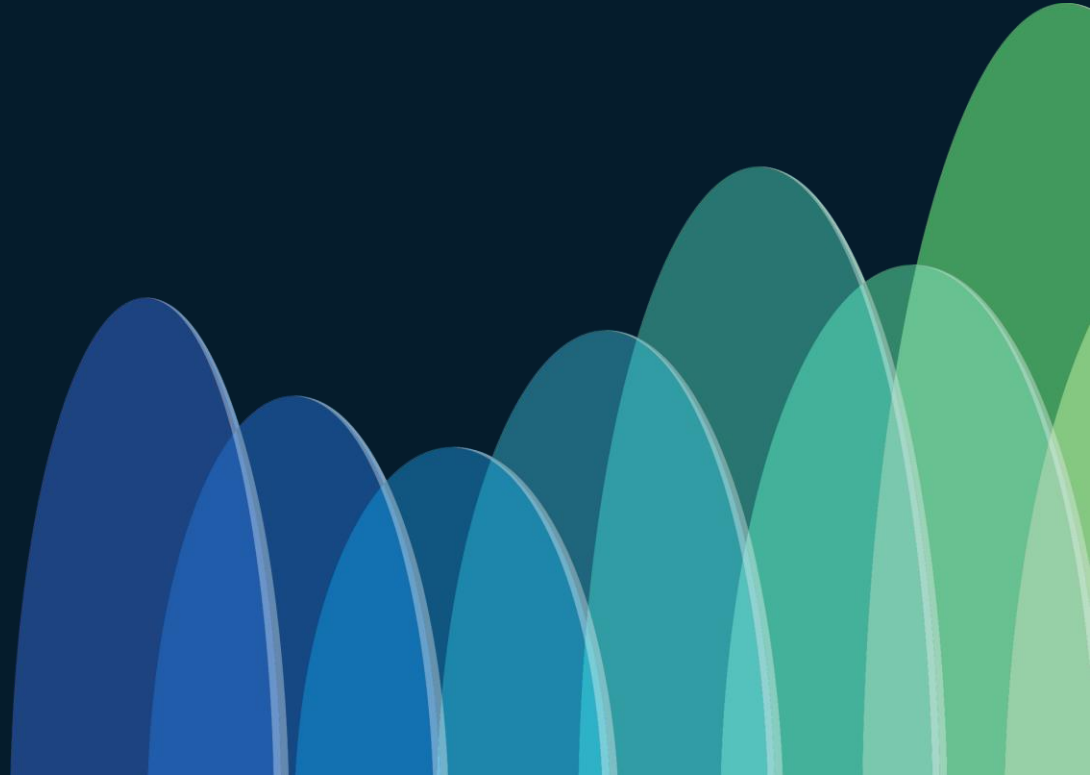
In a WAN isolation scenario, external receivers won't be able to receive the multicast stream originated in a site that is isolated from the WAN until the local L3Out connection is recovered

Use of Active/Standby FW Pair Deployed across Sites

CISCO *Live!*



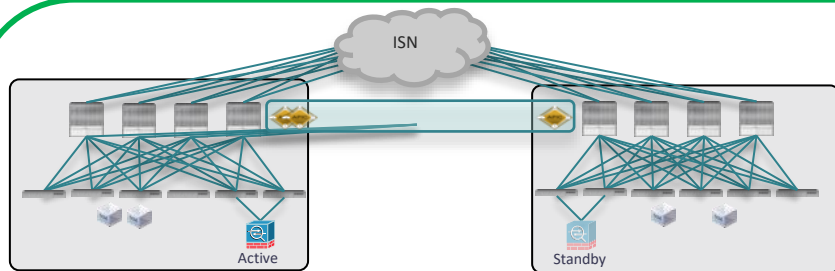
Network Services Integration



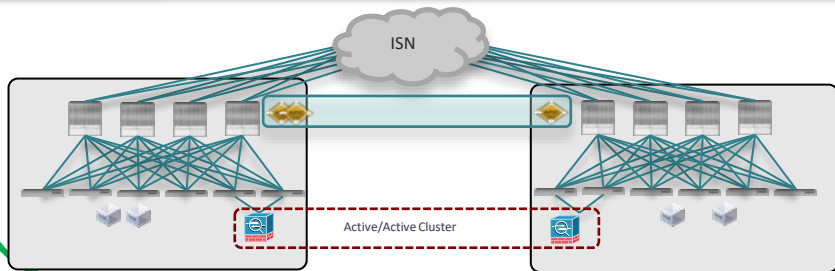
Integration Models

Multi-Pod and Network Services Integration Models

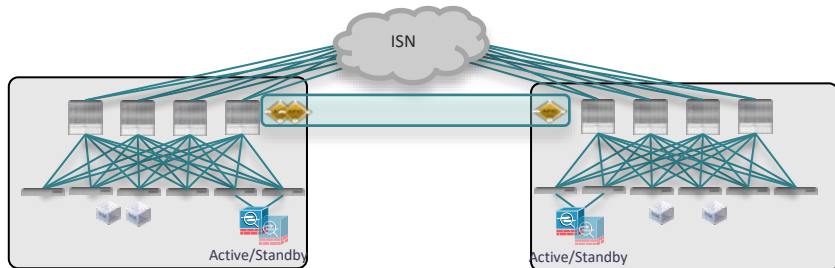
Typical options for an Active/Active DC use case



- 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

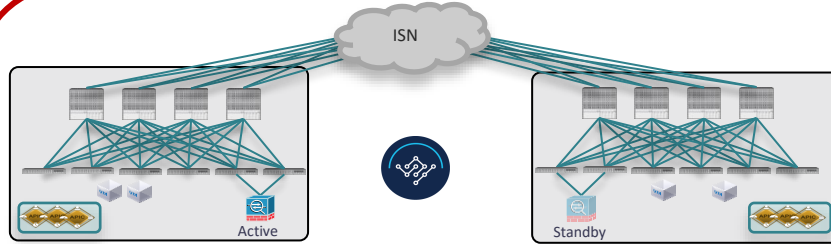


- Independent Active/Standby pairs deployed in separate Pods
- Use of Symmetric PBR to avoid the creation of asymmetric paths crossing different active FW nodes

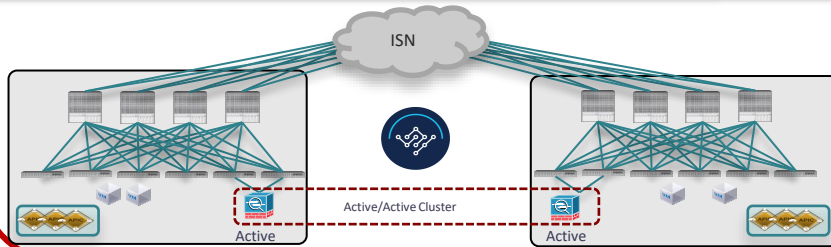
ACI Multi-Site and Network Services

Integration Models

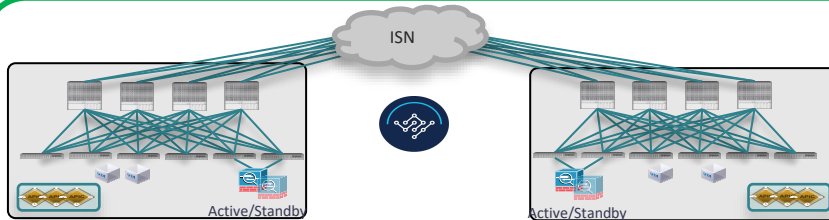
Deployment options fully supported
with ACI Multi-Pod



- 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

Independent Service Node Instances across Sites

Use of Service Graph and Policy Based Redirection

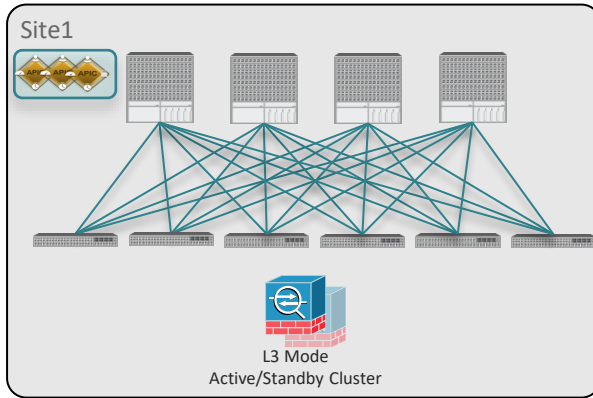
- The PBR policy applied on a leaf switch can only redirect traffic to a service node deployed in the local site
 - Requires the deployment of independent service node function in each site
 - Various design options to increase resiliency for the service node function: per site Active/Standby pair, per site Active/Active cluster, per site multiple independent Active nodes
- HW dependencies:
 - Mandates the use of EX/FX or newer leaf nodes (both for compute and service leaf switches)
- SW dependencies:
 - ACI release 6.0(4)F: Introduction of new vzAny PBR and L3Out-to-L3Out PBR use cases

Use of Service Graph and PBR

Resilient Service Node Deployment in Each Site

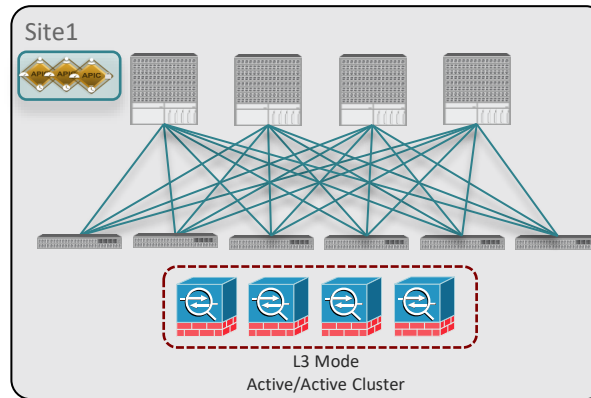
PBR redirection only supported to a local service function, hence it is important to deploy such function in a resilient way

Active/Standby Cluster



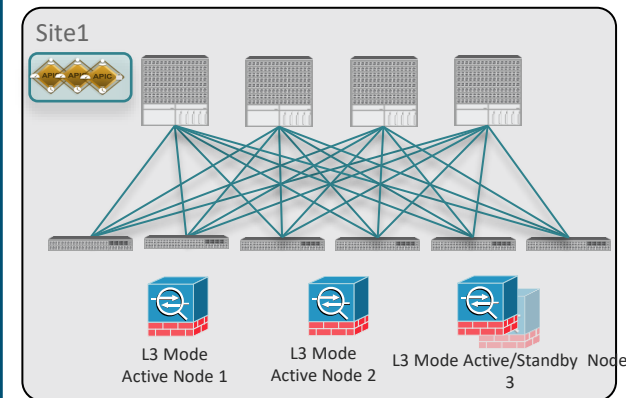
- The Active/Standby pair represents a single MAC/IP entry in the PBR policy

Active/Active Cluster



- The Active/Active cluster represents a single MAC/IP entry in the PBR policy
- Spanned EtherChannel Mode supported with Cisco ASA/FTD platforms

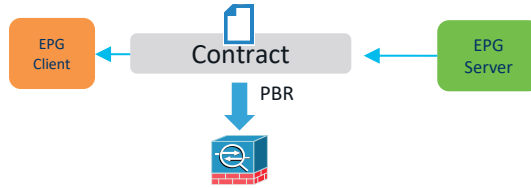
Independent Active Nodes



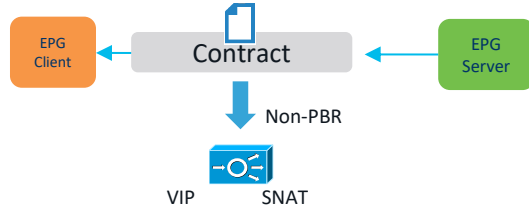
- Each Active node represent a unique MAC/IP entry in the PBR policy
- Use of Symmetric PBR to ensure each flow is handled by the same Active node in both directions

Multi-Site and Network Services Integration

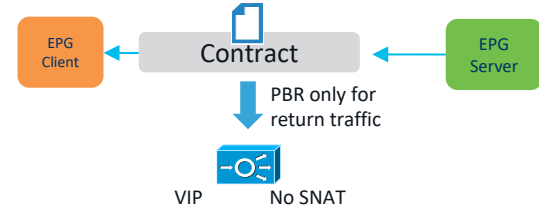
Single-Node Service Graph with PBR



- Single-Node FW with PBR (used for both traffic directions)
- FW connected in one-arm or two-arms mode
- ACI 3.2(1): PBR policy applied on the compute leaf (N-S flows) or on the consumer leaf (E-W flows)
- ACI 4.0(1): PBR policy applied on the compute leaf (N-S flows) or on the provider leaf (E-W flows)



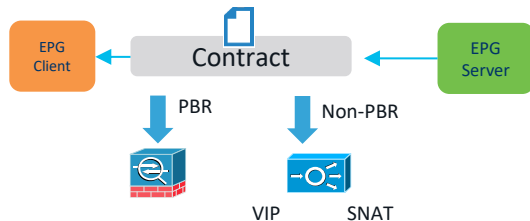
- LB connected in one-arm or two-arms mode
- No need for PBR at all, LB connected to BDs/EPGs as a regular endpoint
- LB could also be connected to an L3Out (may require intersite L3Out support)



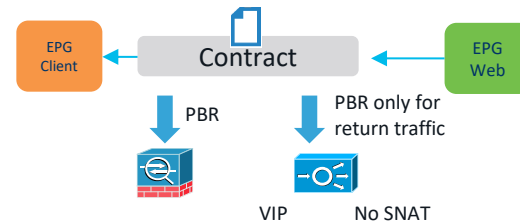
- LB with PBR for return traffic supported from ACI 4.0(1) release
- LB connected in one-arm or two-arms mode
- PBR policy applied on the compute leaf (N-S flows) or on the provider leaf (E-W flows)

Multi-Site and Network Services Integration

Dual-Node Service-Graph with PBR (Main Use Cases)

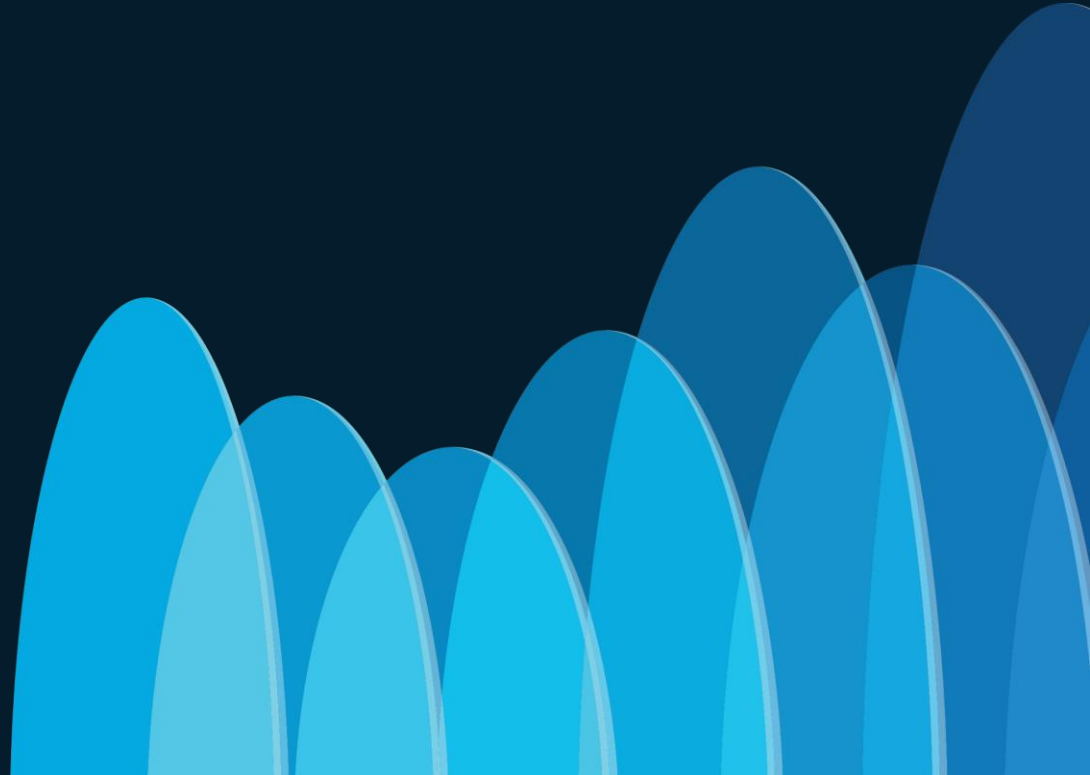


- PBR policy only to redirect traffic to the FW, no need for PBR for the LB (LB connected to BDs/EPGs as a regular endpoint)
- FW and LB in single-arm or dual-arms mode
- Supported from ACI release 4.0(1)



- PBR policy to redirect traffic to the FW, second PBR policies for the LB (for return traffic)
- FW and LB in single-arm or dual-arms mode
- Supported from ACI release 4.0(1)

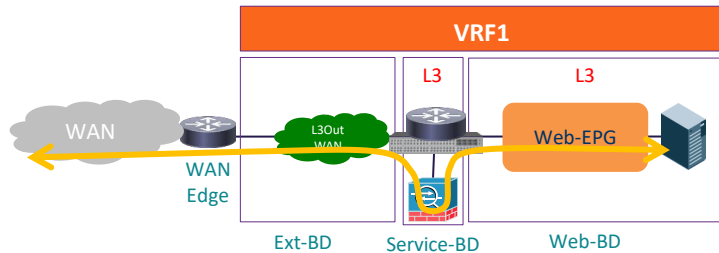
Use of Service Graph and PBR North-South and East-West



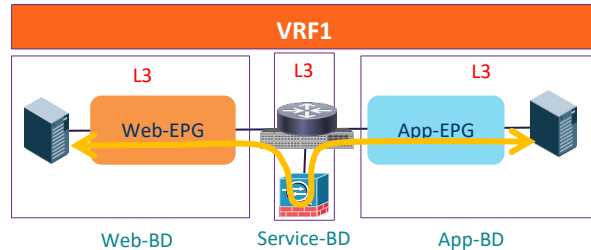
Use of Service Graph and PBR

North-South and East-West Use Cases

North-South



East-West



- Best practice recommendations for both North-South and East-West use cases:

Service Node deployed in 'one arm' mode ('two-arms' mode also supported but not preferred)

Service-BD must be stretched across sites (BUM flooding can/should be disabled)

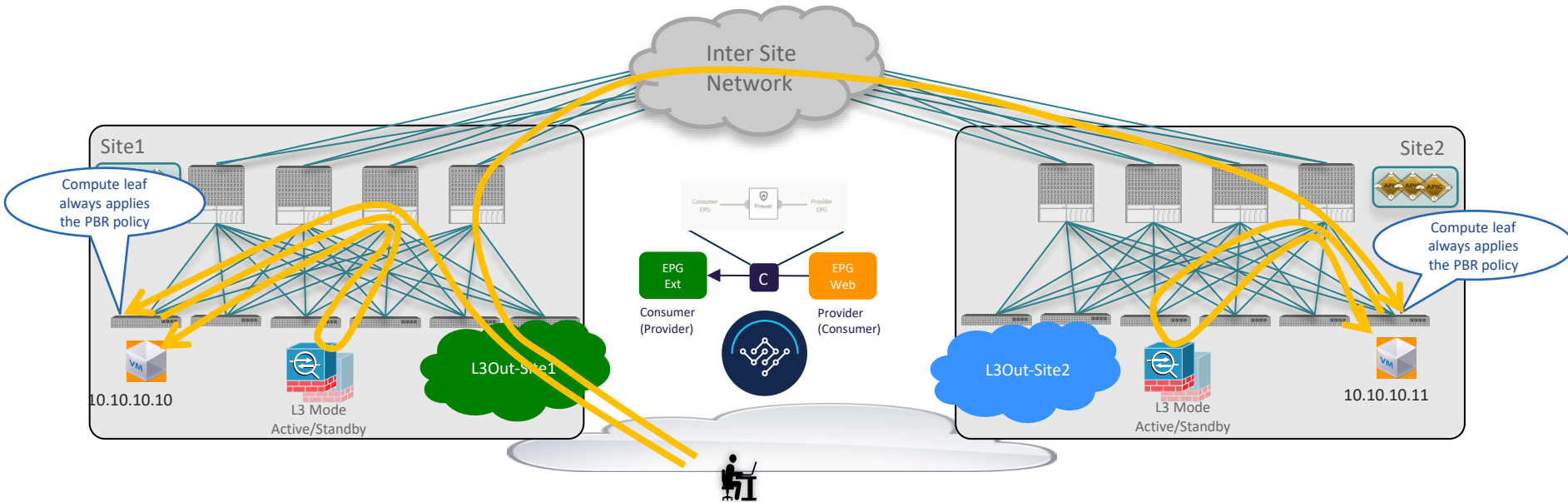
Ext-EPG must also be a stretched object, mapped to the individual L3Outs defined in each site

Web-BD and App-BD can be stretched across sites or locally defined in each site

- North-South use case
 - Intra-VRF only support as of ACI 5.0(1) release
- East-West use case
 - Supported intra-VRF or inter-VRFs/Tenants
 - Requires to configure the subnet under the Consumer EPG from ACI release 4.0(1) (under the Provider EPG in 3.2(x) releases)

North-South Communication

Inbound Traffic



- Inbound traffic can enter any site when destined to a stretched subnet (if ingress optimisation is not deployed or possible)
- 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

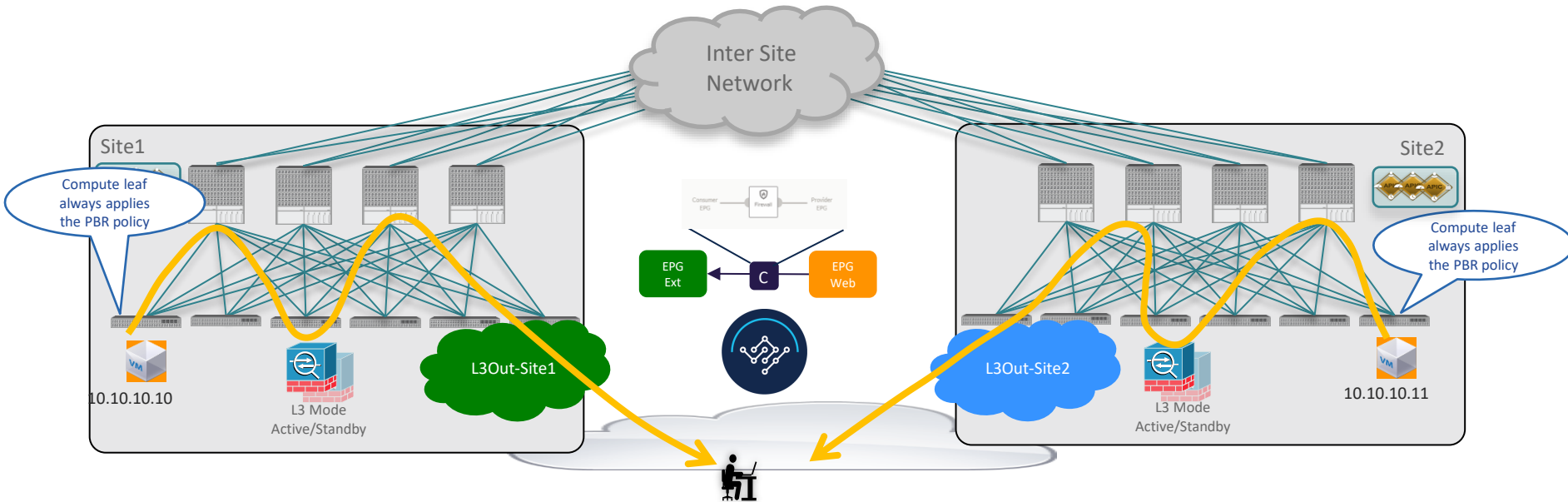


Policy Control Enforcement Preference: **Enforced** Unenforced

Policy Control Enforcement Direction: Egress **Ingress**

North-South Communication

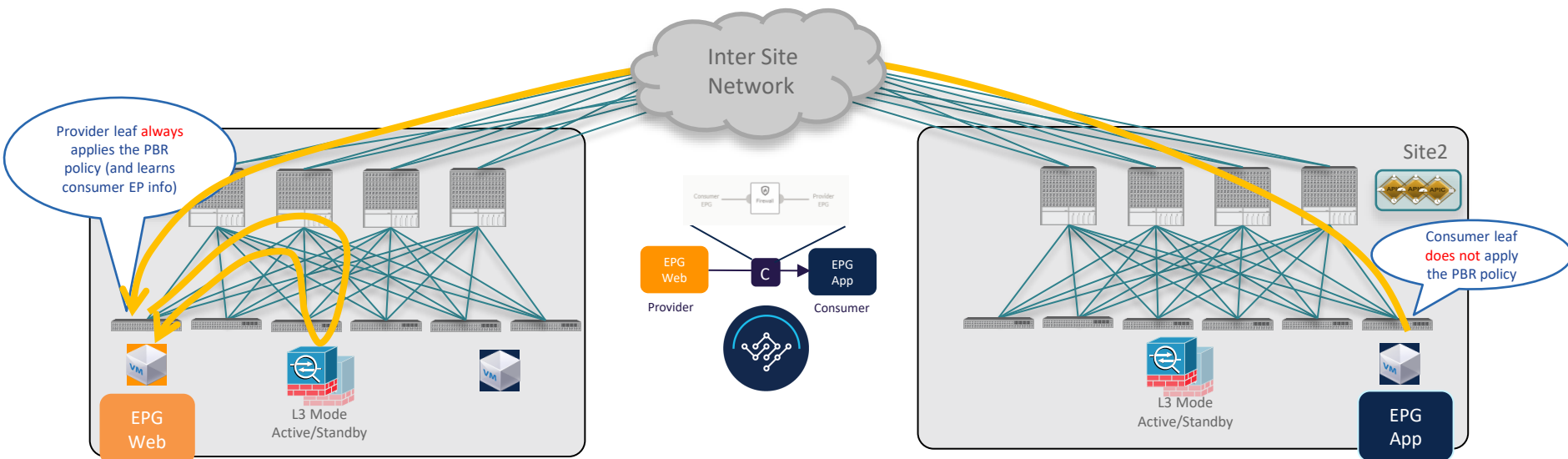
Outbound Traffic



- 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

East-West Communication

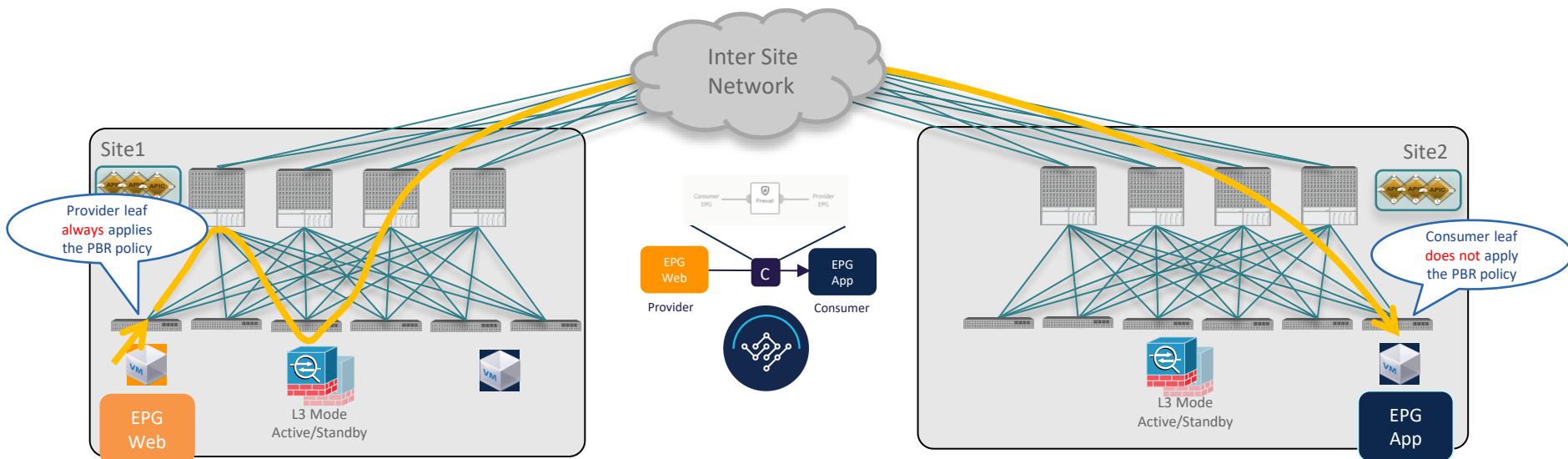
Consumer to Provider Flow



- 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

East-West Communication

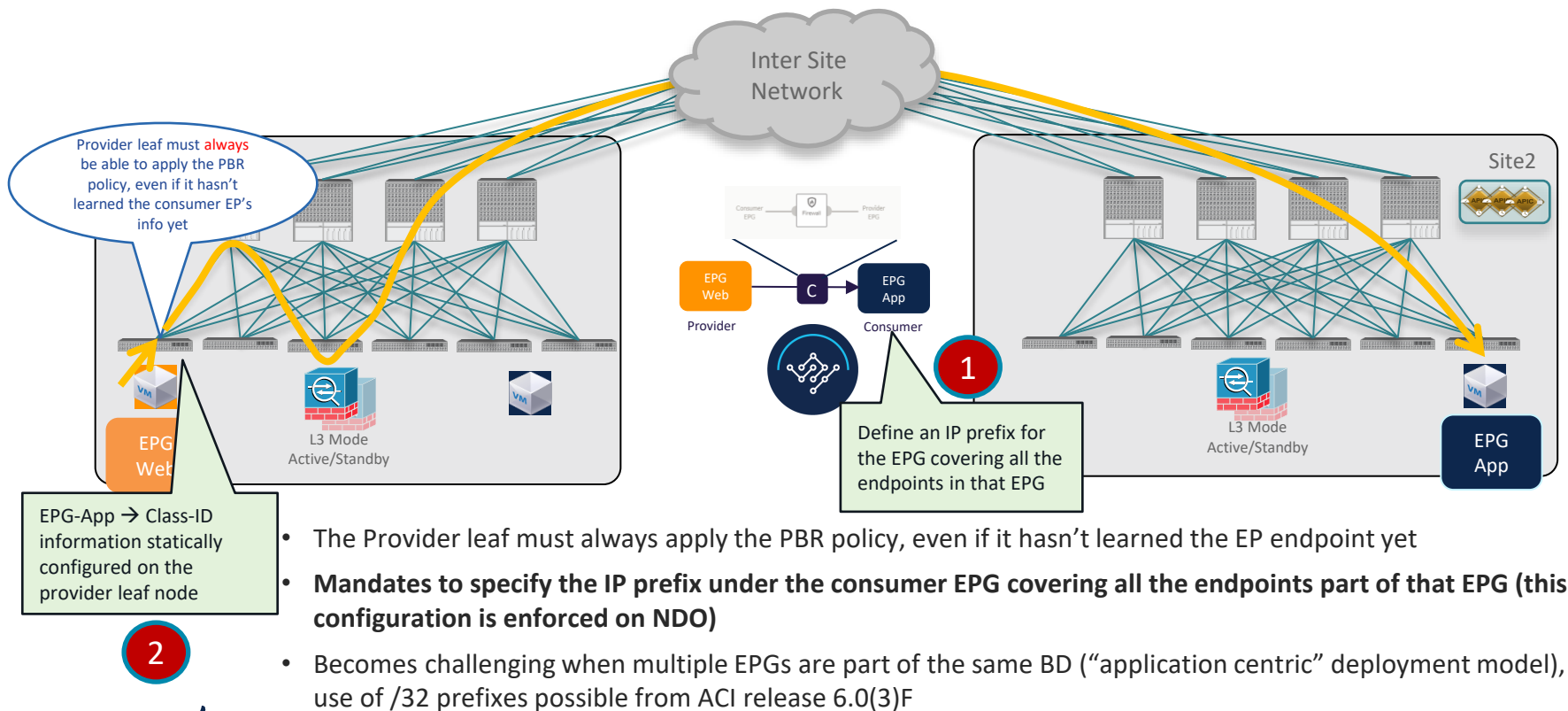
Provider to Consumer Return Flow



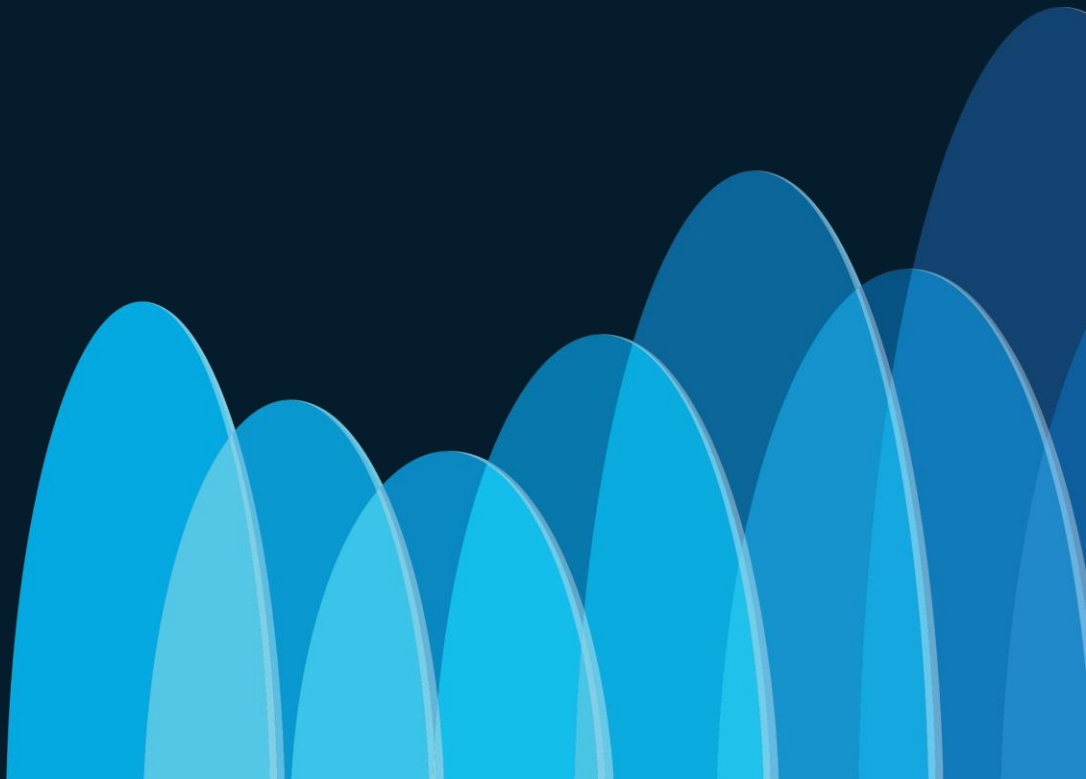
- EPGs can be locally defined or stretched across sites and can be part of the same VRF or in different VRFs (and/or Tenants)
- PBR policy is always applied only on the leaf switch where the **Provider** endpoint is connected
 - The Provider leaf always redirects traffic to a local service node

East-West Communication

What if the Communication is Initiated by the Provider?



New PBR Supported Use Cases

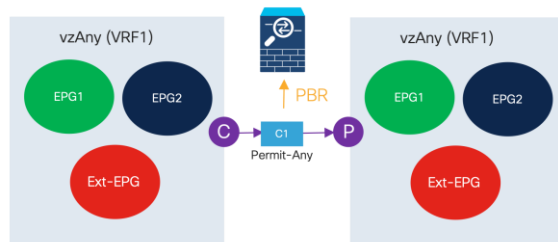


ACI Multi-Site and PBR Enhancements

New Supported Use Cases

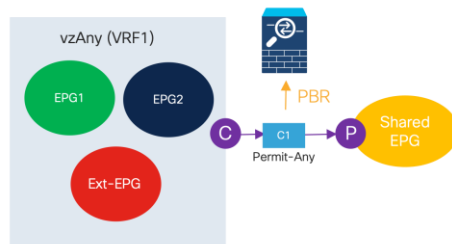
NDO 4.2(1)/ACI 6.0(3F): Beta
NDO 4.2(3)/ACI 6.0(4F): GA

Any-to-Any



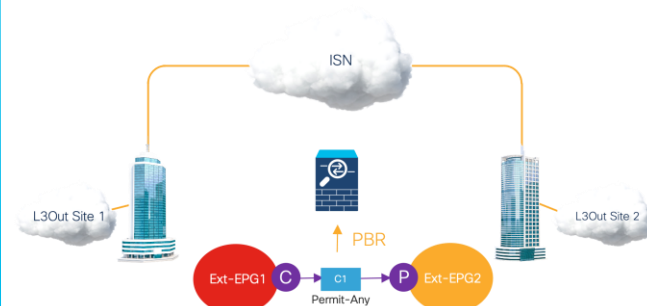
- Support only for single service node insertion (one-arm)
- Distributed deployment model (traffic is redirected via both local and remote service node)
- Intra-VRF only
- Works for both “network centric” and “app centric” designs

Many-to-One



- Support only for single service node insertion (one-arm)
- Intra-VRF only
- Two scenarios:
 1. vzAny-to-EPG
 2. vzAny-to-L3Out
- Works for both “network centric” and “app centric” designs

Transit Intersite L3Out



- Support only for single service node insertion (one-arm)
- Redirect intersite transit routing traffic flows
- Traffic is redirected via both local and remote service node
- Intra-VRF and inter-VRF

<https://www.cisco.com/c/en/us/td/docs/dcn/ndo/4x/configuration/cisco-nexus-dashboard-orchestrator-configuration-guide-aci-421/ndo-configuration-aci-use-case-vzany-pbr-42x.html>
<https://www.cisco.com/c/en/us/td/docs/dcn/ndo/4x/configuration/cisco-nexus-dashboard-orchestrator-configuration-guide-aci-421/ndo-configuration-aci-use-case-intersite-l3out-pbr-42x.html>

ACI Multi-Site and PBR Enhancements

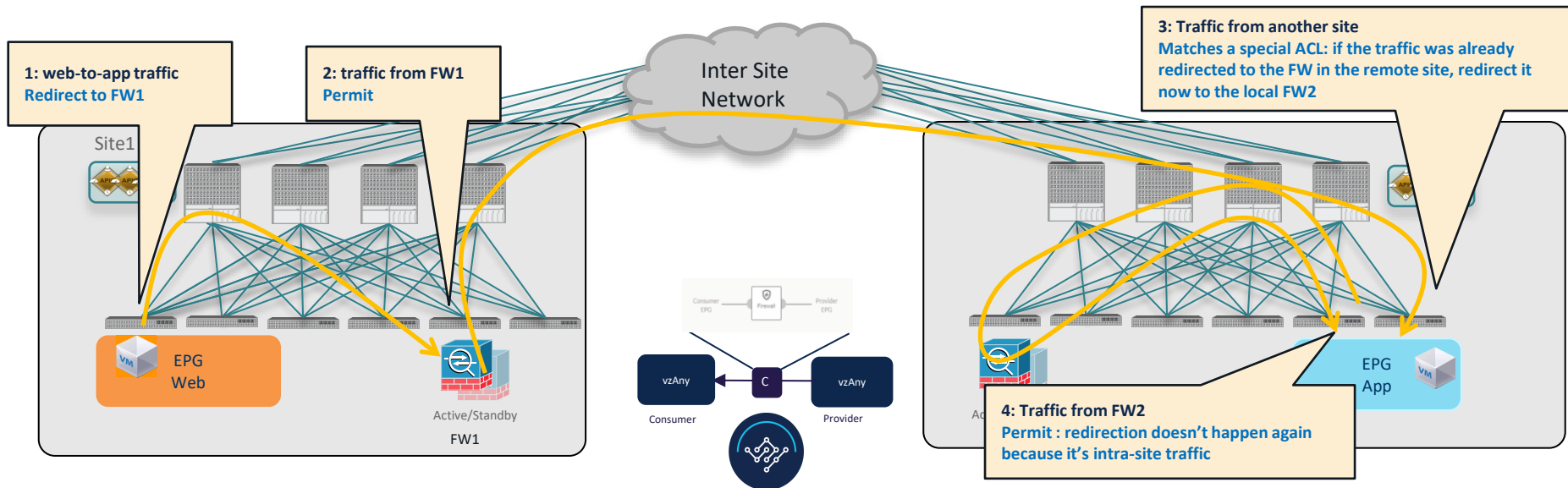
How to Keep Traffic Symmetric

- vzAny-to-vzAny, vzAny-to-L3OutEPG, L3OutEPG-to-L3OutEPG
 - Redirect inter-site traffic in both ingress and egress sites
 - Note:** If it's intra-site traffic, redirection only happens to the service node in the local site
- vzAny-to-EPG
 - Redirect inter-site traffic always on the provider side
- For all the 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 in the vzAny-to-EPG use case) can always resolve the destination class-ID without requiring any static IP prefix configuration under the consumer EPG (fully supporting “app centric” designs)

How to Keep Traffic Symmetric

vzAny-to-vzAny, vzAny-to-L3OutEPG, L3OutEPG-to-L3OutEPG

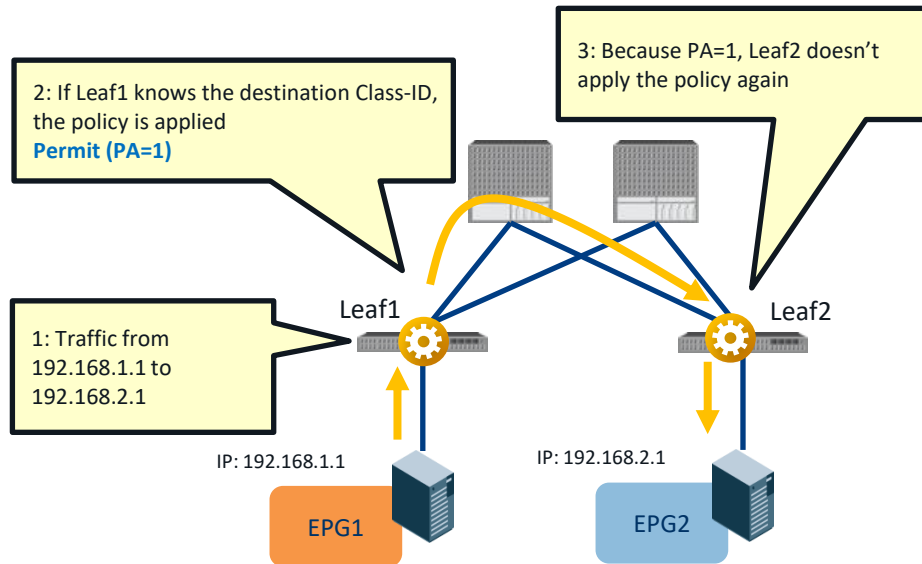
- Redirect “inter-site” traffic in both ingress and egress sites



How to Identify if Traffic Was Redirected?

Use of the Policy Applied (PA) Bits

- PA bits (2 bits) in the VXLAN Header: Source Policy (SP) bit and Destination Policy (DP) bit

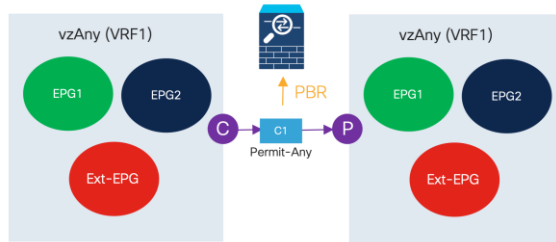


	SP	DP	Behavior
PA=1	1	1	The egress leaf doesn't apply policy because it was already applied in the ingress leaf
PA=0	0	0	The egress leaf should apply the policy because it has not been applied yet



“SP=0, DP=1” combination: will be set for traffic received from the service EPG to indicate that it was redirected to a service node

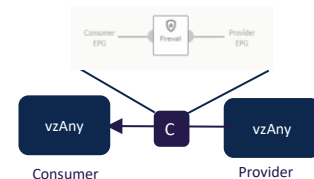
1. Any-to-Any PBR Use Case



vzAny-to-vzAny PBR Use Case

Deployment Considerations

NDO 4.2(3)/ACI 6.0(4)



- The goal is redirecting to a Firewall Service all the intra-VRF traffic flows (north-south and east-west)
- Support only for single service node insertion in NDO 4.2(3)/ACI 6.0(4)
- How to avoid asymmetric traffic paths through different FW nodes?
 - Redirecting “inter-site” traffic to the Firewall services in the source and destination fabrics
 - 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 EPG?
 - Use of conversational learning
- Full “application centric” support, no need to configure any IP prefix under any EPG

vzAny-to-vzAny PBR Use Case

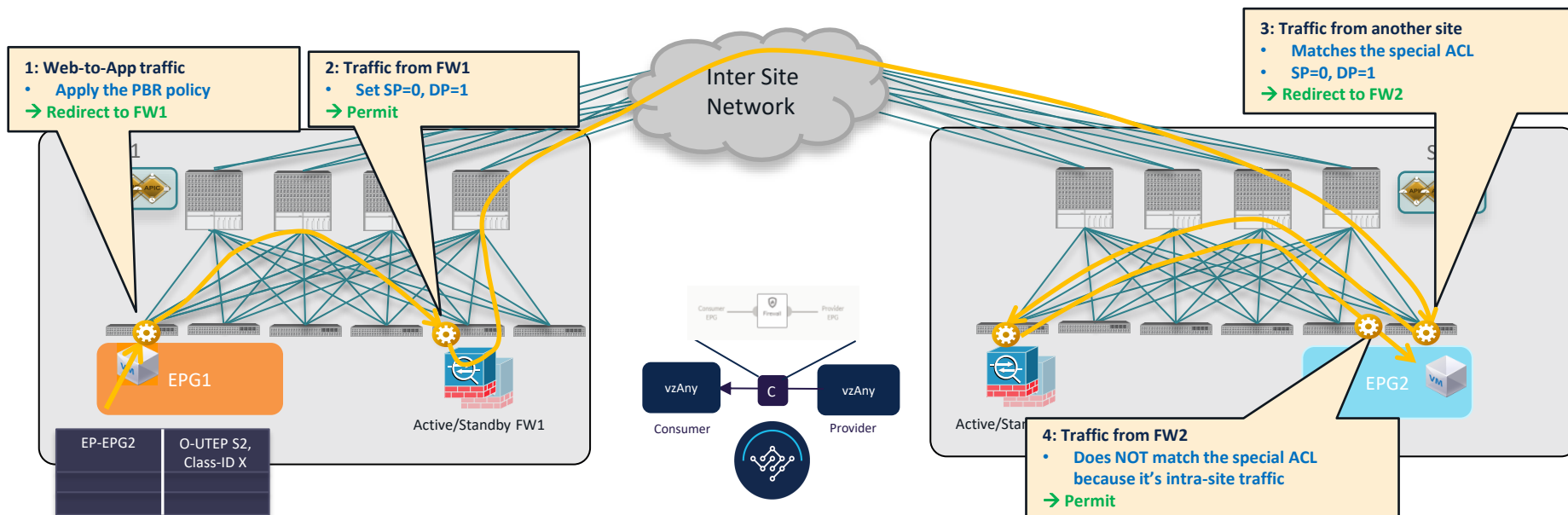
NDO 4.2(3)/ACI 6.0(4)

Consumer to Provider Direction



SP=0, DP=1
for traffic from
the service EPG

Assumption: the class-ID for the provider endpoint **is known** on the consumer leaf



vzAny-to-vzAny PBR Use Case

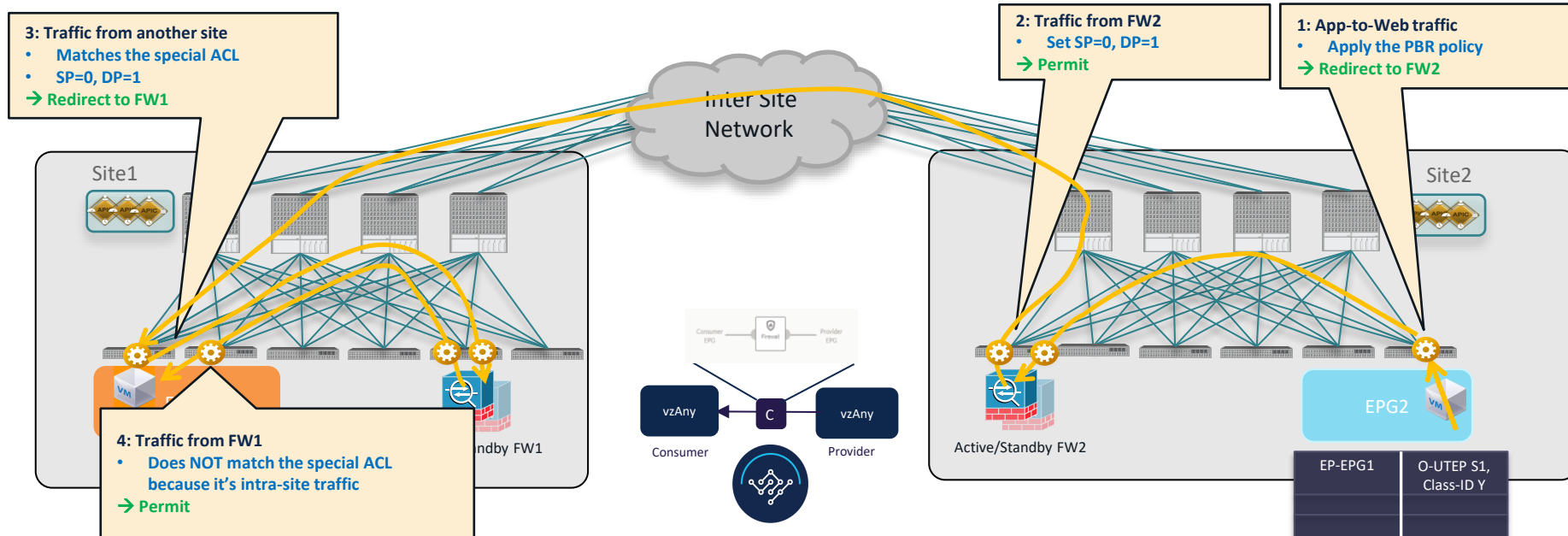
Provider to Consumer Direction

NDO 4.2(3)/ACI 6.0(4)



SP=0, DP=1
for traffic from
the service EPG

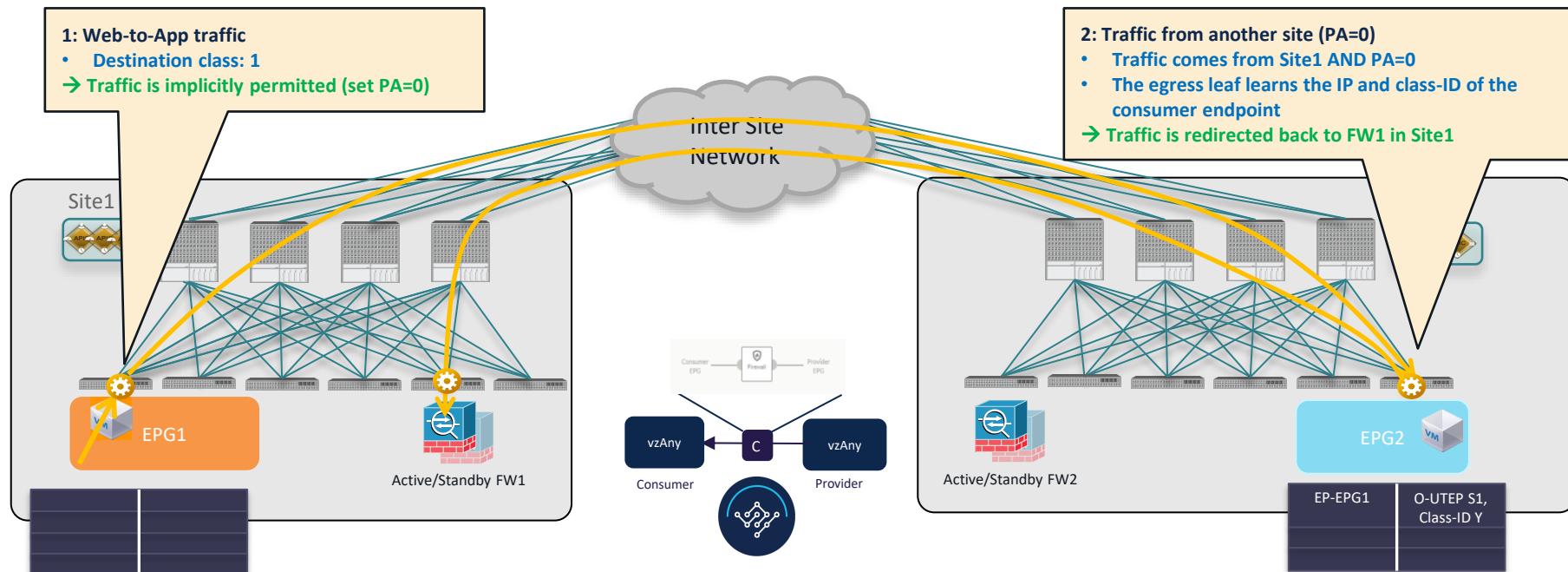
Assumption: the class-ID for the provider endpoint **is known** on the consumer leaf



vzAny-to-vzAny PBR Use Case

What if the Ingress Leaf doesn't Know the Destination Class-ID? (1/3)

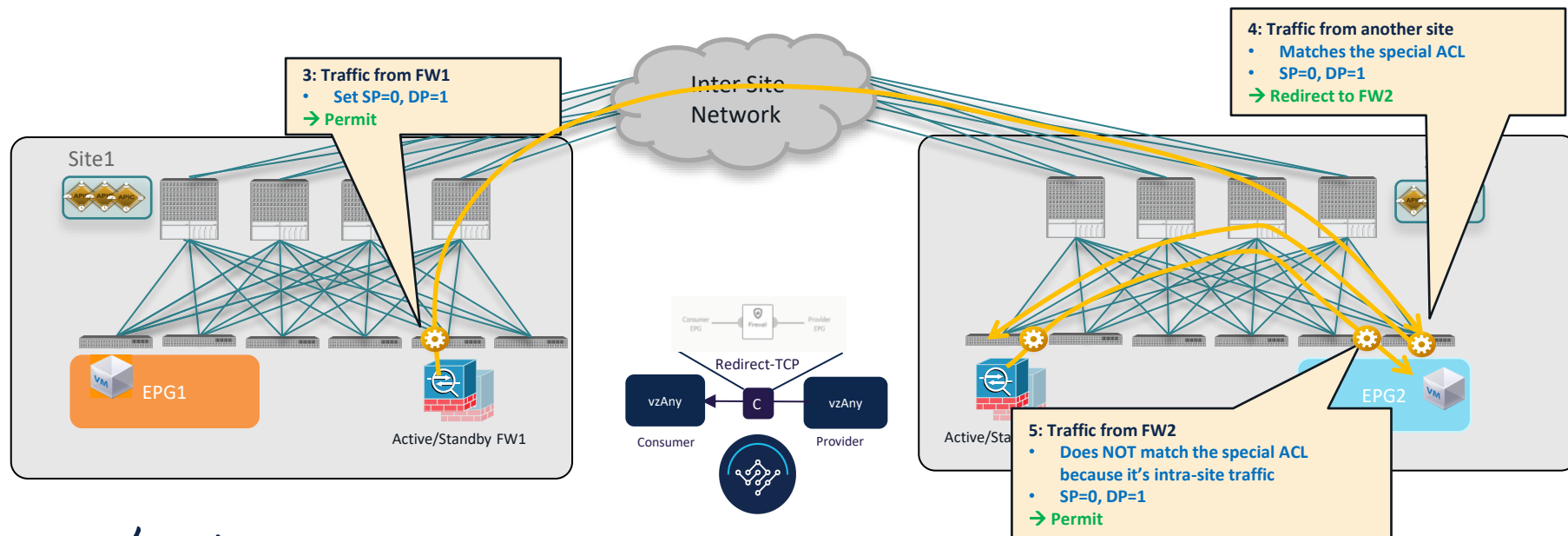
- The destination leaf steers the traffic back to the source site to be inspected by the service device there



vzAny-to-vzAny PBR Use Case

What if the Ingress Leaf doesn't Know the Destination Class-ID? (2/3)

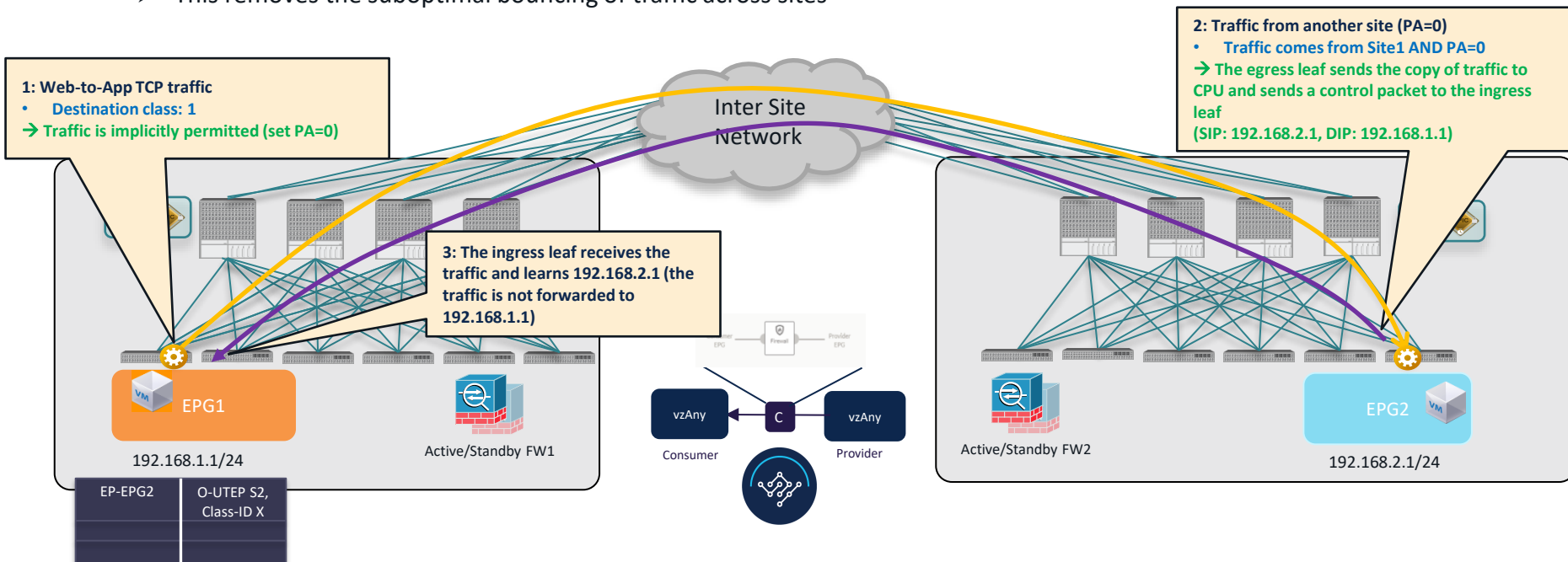
- When the destination leaf receives the flow from the service device in site 1, it can now redirect it to the local service node



vzAny-to-vzAny PBR Use Case

What if the Ingress Leaf doesn't Know the Destination Class-ID? (3/3)

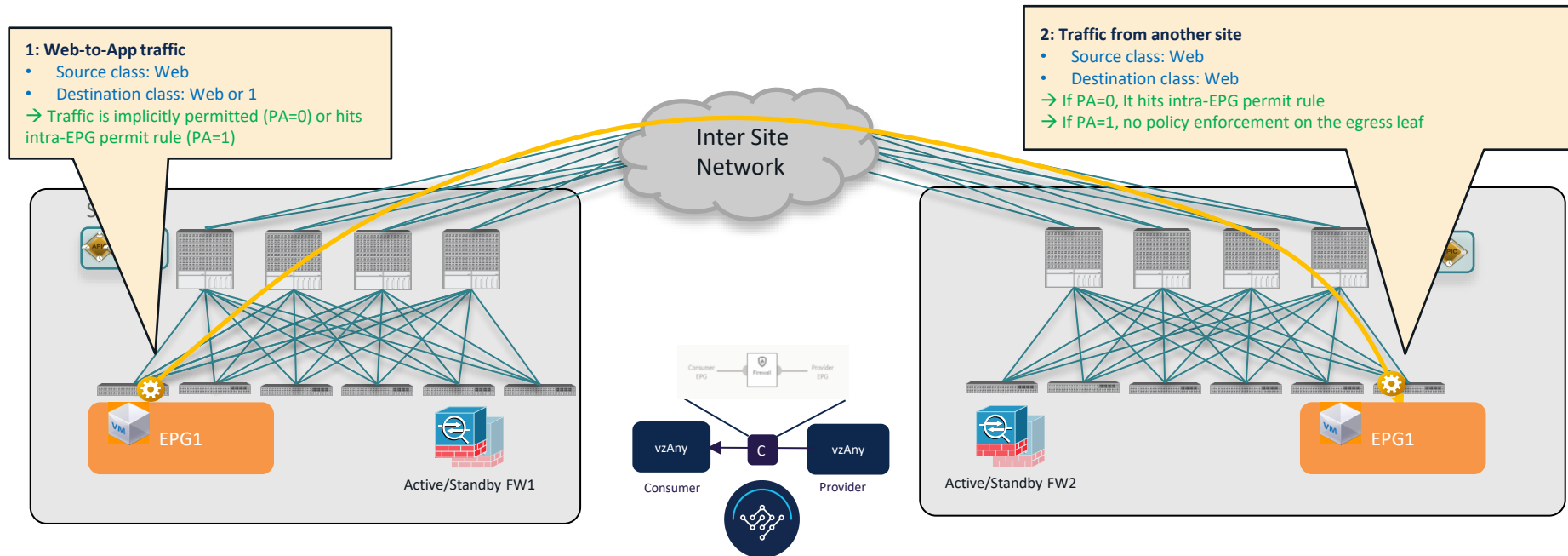
- Conversational Learning is activated to ensure that the ingress leaf can learn the destination EP's information
 - This removes the suboptimal bouncing of traffic across sites



vzAny-to-vzAny PBR Use Case

Intra-EPG Traffic

- Intra-EPG permit rule (priority 3) wins over vzAny-to-vzAny rule (priority 17)
 - This ensures that intersite intra-EPG communication happens freely without any PBR redirection



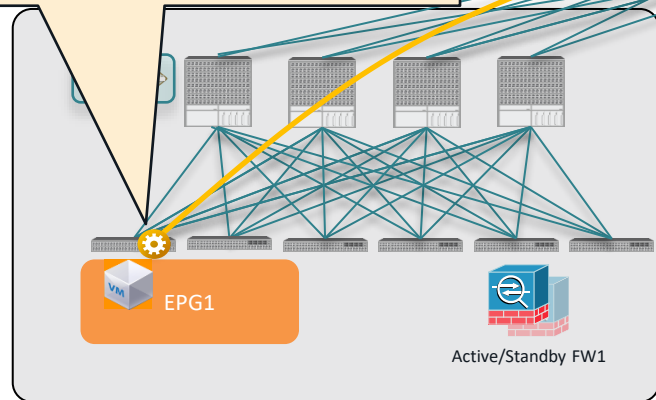
vzAny-to-vzAny PBR Use Case

Bypass the Firewall for Specific EPG-to-EPG Traffic

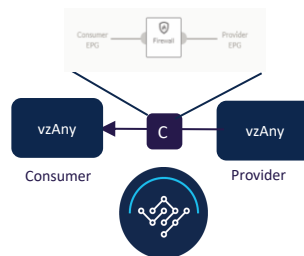
- EPG-to-EPG specific rule (priority 7 or 9) wins over vzAny-to-vzAny rule (priority 17)
 - This ensures that intersite communication between EPG can happen based on the specific contract that is applied between those EPGs

1: Web-to-App SSH traffic

- Source class: Web
 - Destination class: App or 1
- Traffic is implicitly permitted (set PA=0)
or hits Web-to-App permit SSH rule (set PA=1)

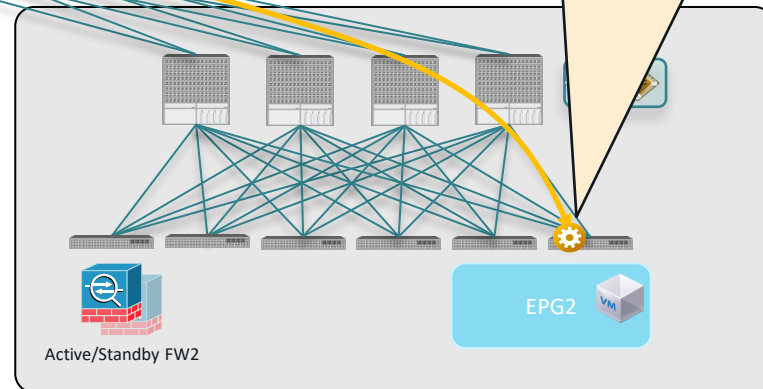


Inter Site
Network



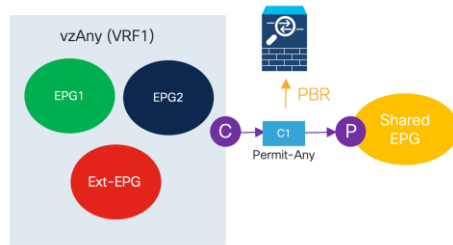
2: TCP traffic from another site

- Source class: Web
 - Destination class: App
- If PA=0, apply EPG-to-EPG permit SSH rule
→ If PA=1, policy is not applied on the leaf



2. Many-to-One PBR Use Case

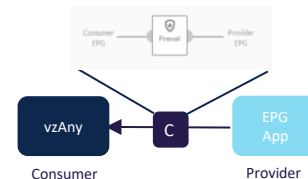
2.1. vzAny-to-EPG



vzAny-to-EPG PBR Use Case

Deployment Considerations

NDO 4.2(3)/ACI 6.0(4)

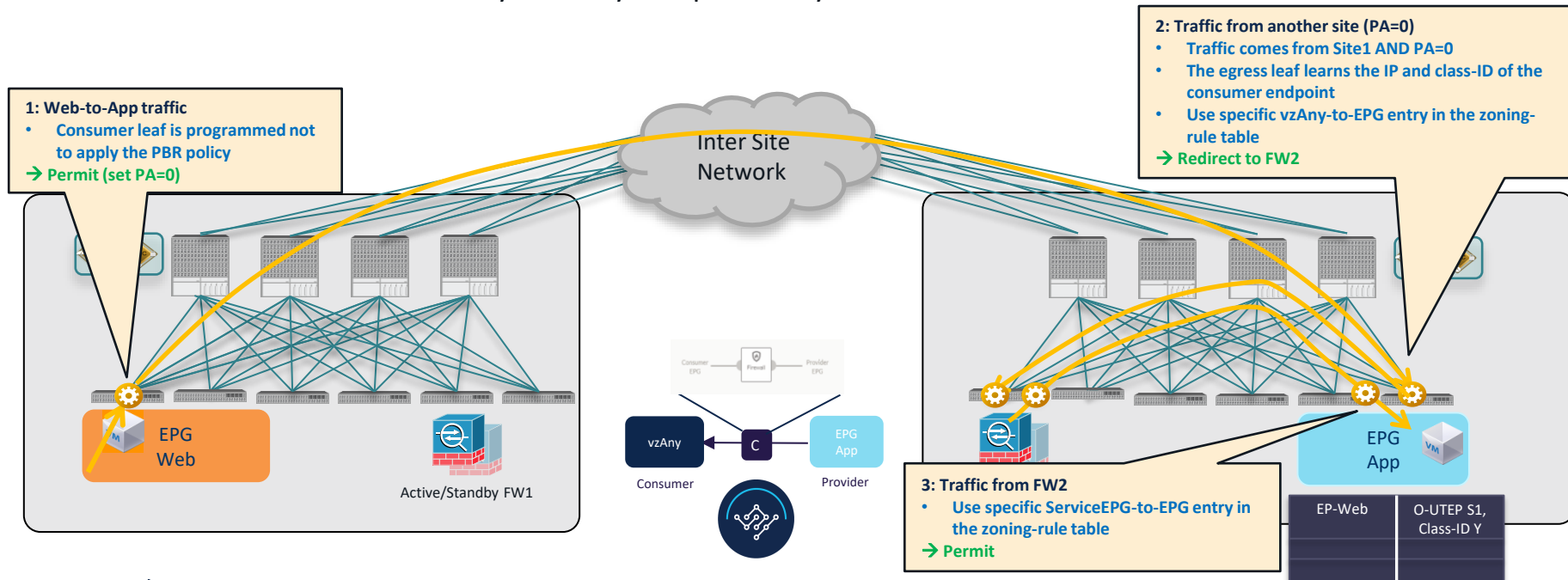


- The goal is redirecting to a Firewall Service all the flows originated from an EPG/L3OutEPG part of a VRF and destined to a shared EPG
- The shared EPG must be part of the same VRF as the source endpoint
- How to avoid asymmetric traffic paths through different FW nodes?
 - PBR enforcement is always and only applied on the provider leaf, redirecting the flows (both directions) to the Firewall service deployed in the provider site
- How to ensure the Provider leaf nodes can resolve the destination class-ID without requiring any static IP prefix configuration under the EPG?
 - Use of conversational learning
- Full “application centric” support, no need to configure any IP prefix under any EPG

vzAny-to-EPG PBR Use Case

Consumer to Provider Direction

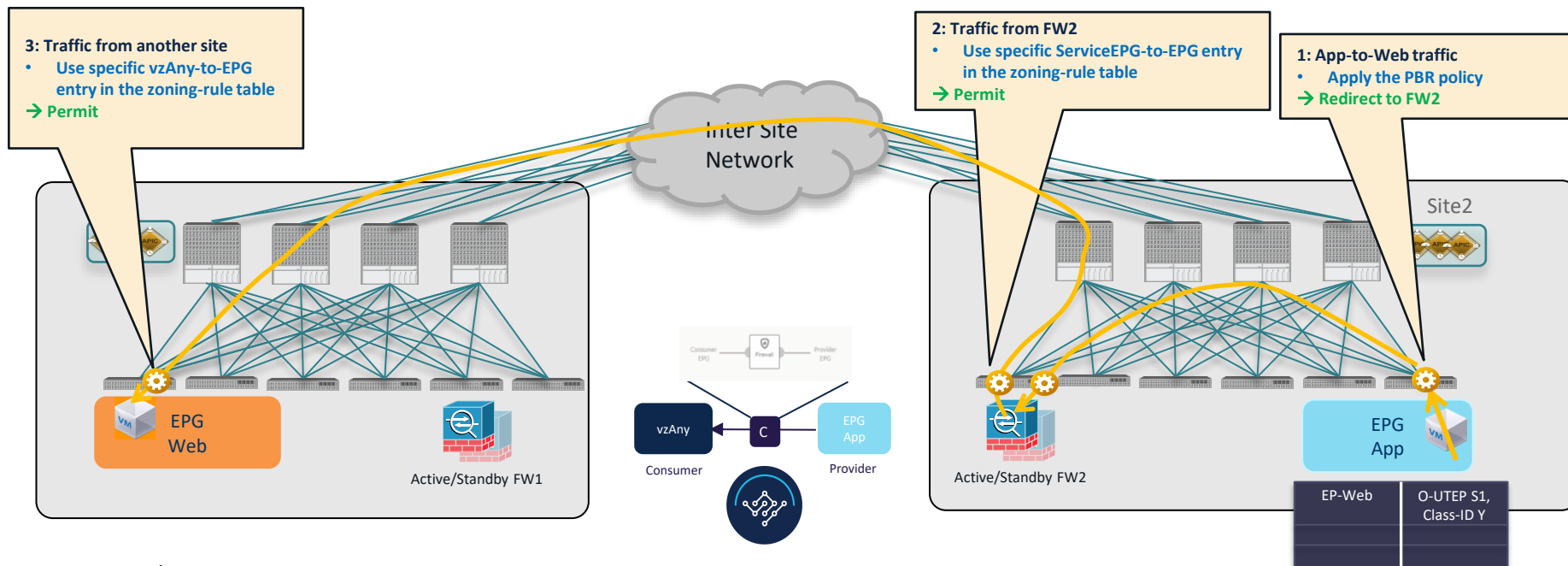
- Provider leaf enforcement only to always keep traffic symmetric



vzAny-to-EPG PBR Use Case

Provider to Consumer Direction

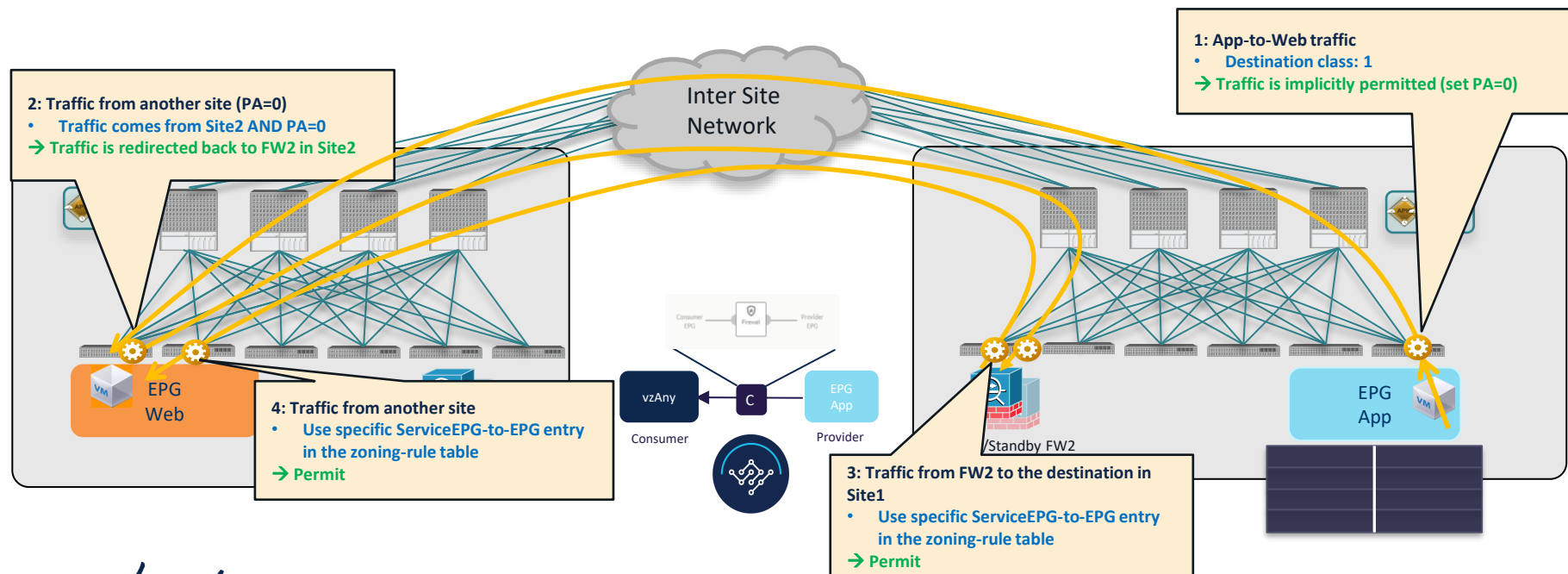
- Provider leaf enforcement only to always keep traffic symmetric



vzAny-to-EPG PBR Use Case

What if the Provider Leaf doesn't Know the Consumer Endpoint? (1/2)

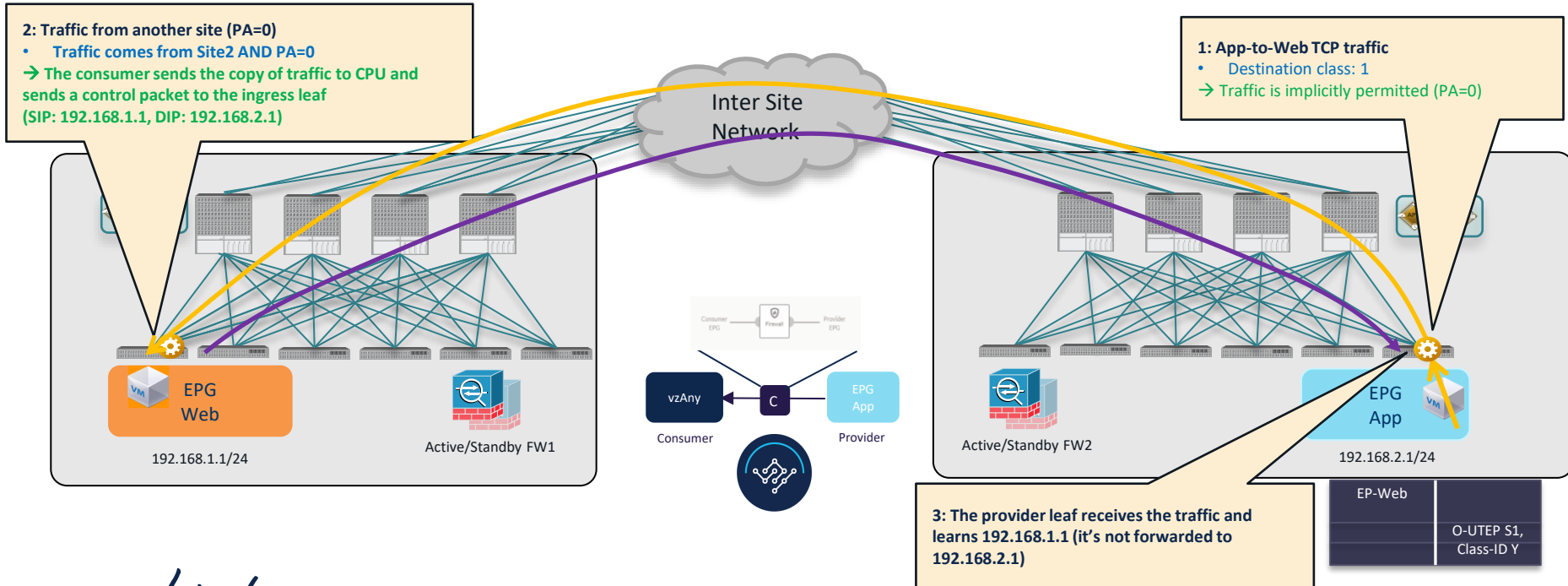
- The consumer leaf steers the traffic back to the provider's site to be inspected by the service device there



vzAny-to-EPG PBR Use Case

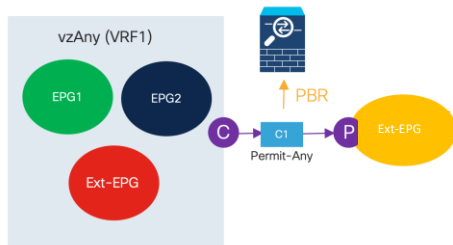
What if the Provider Leaf doesn't Know the Consumer Endpoint? (2/2)

- Conversational Learning is activated to ensure that the provider leaf can learn the consumer's EP information
 - This removes the suboptimal bouncing of traffic across sites



2. Many-to-One PBR Use Case

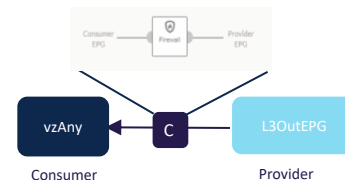
2.2. vzAny-to-L3Out



vzAny-to-L3OutEPG PBR Use Case

Deployment Considerations

NDO 4.2(3)/ACI 6.0(4)



- The goal is redirecting to a Firewall Service all the flows originated from an EPG/L3OutEPG part of a VRF and destined to an L3OutEPG
- The L3OutEPG (and the L3Out) must be part of the same VRF as the source endpoint
- How to avoid asymmetric traffic paths through different FW nodes?
 - Redirecting “inter-site” traffic to the Firewall services in the source and destination fabrics
 - Only redirection to the local Firewall service for intra-fabric flows
- Full “application centric” support, no need to configure any IP prefix under any EPG

vzAny-to-L3OutEPG PBR Use Case

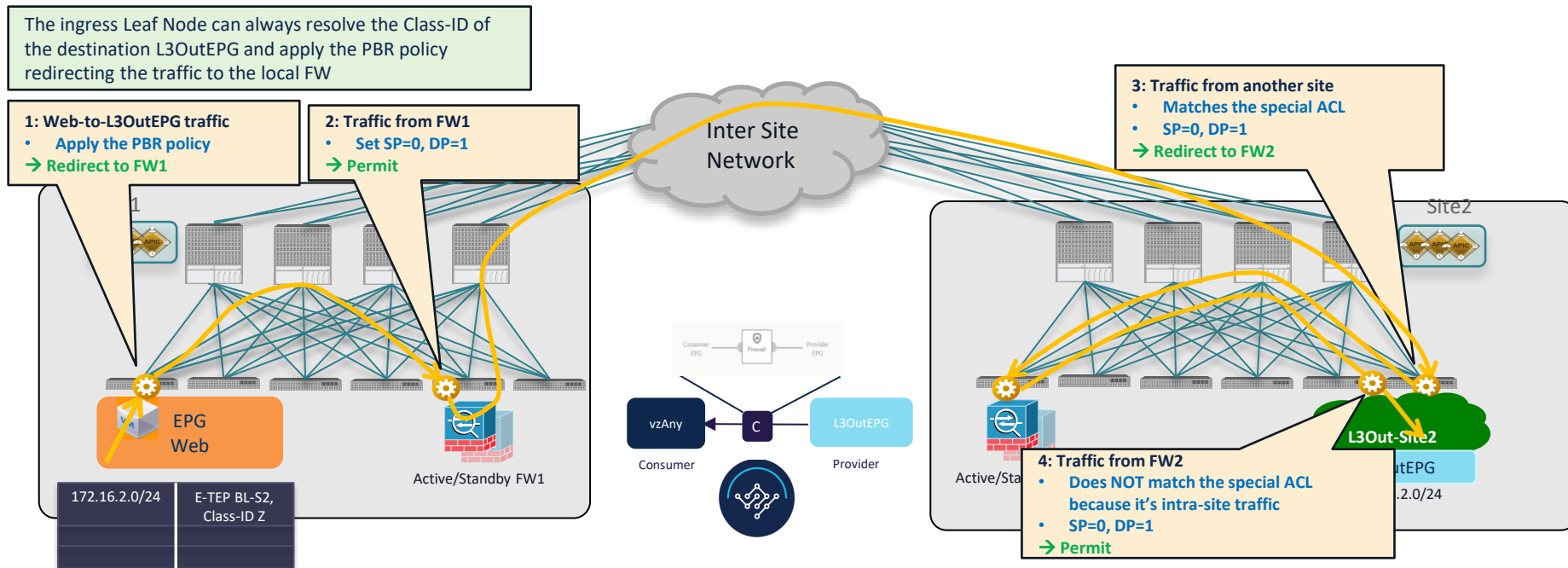
NDO 4.2(3)/ACI 6.0(4)

Consumer to Provider Direction



SP=0, DP=1
for traffic from
the service EPG

- Redirect “inter-site” traffic in both ingress and egress sites



vzAny-to-L3OutEPG PBR Use Case

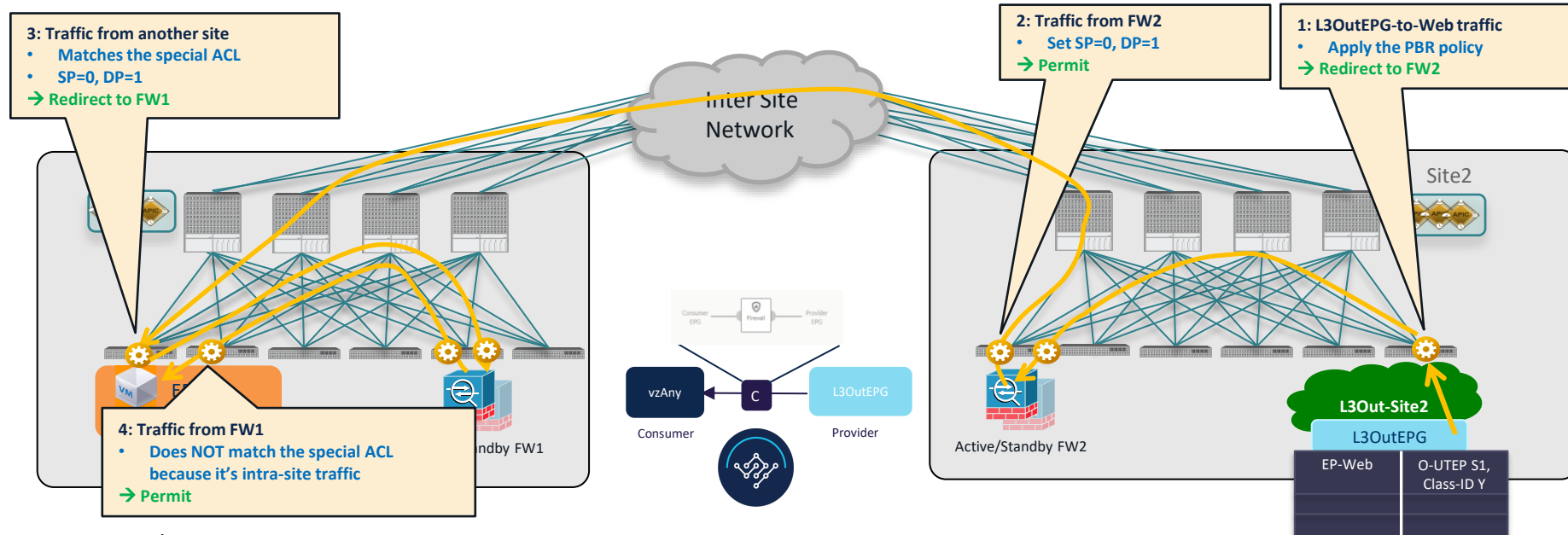
NDO 4.2(3)/ACI 6.0(4)

Provider to Consumer Direction



SP=0, DP=1
for traffic from
the service EPG

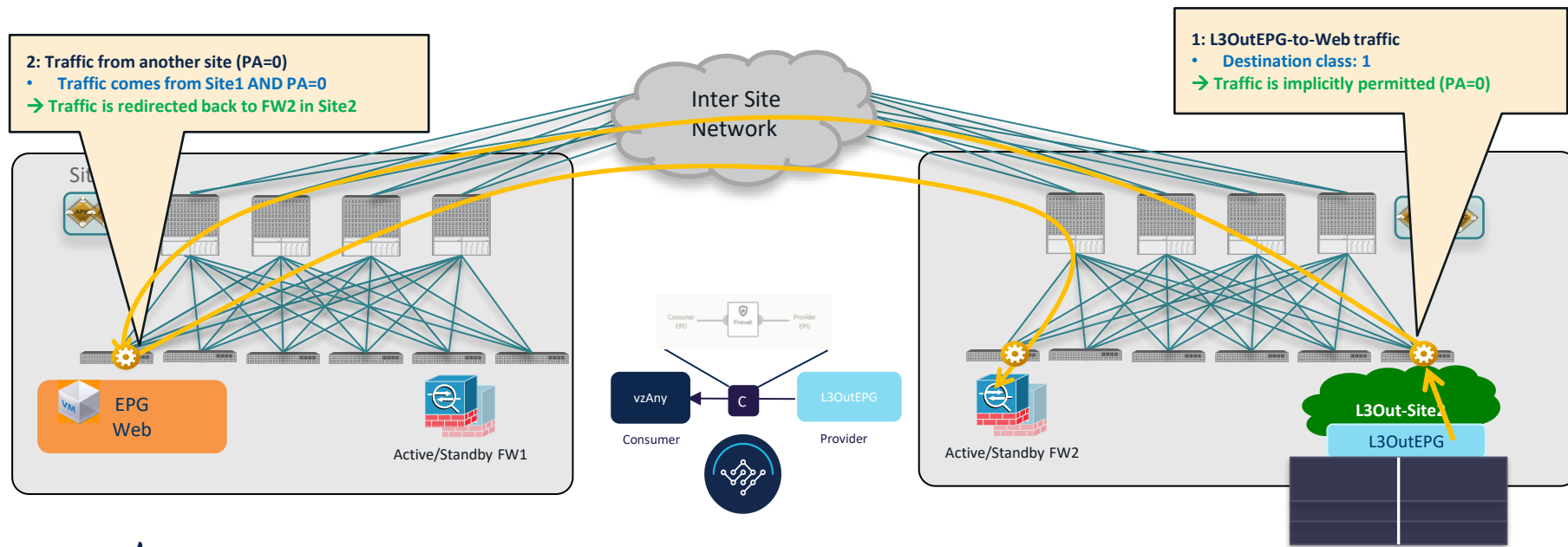
Assumption: the class-ID for the consumer endpoint **is known** on the provider leaf



vzAny-to-L3OutEPG PBR Use Case

What if the Provider Leaf doesn't Know the Consumer Endpoint? (1/3)

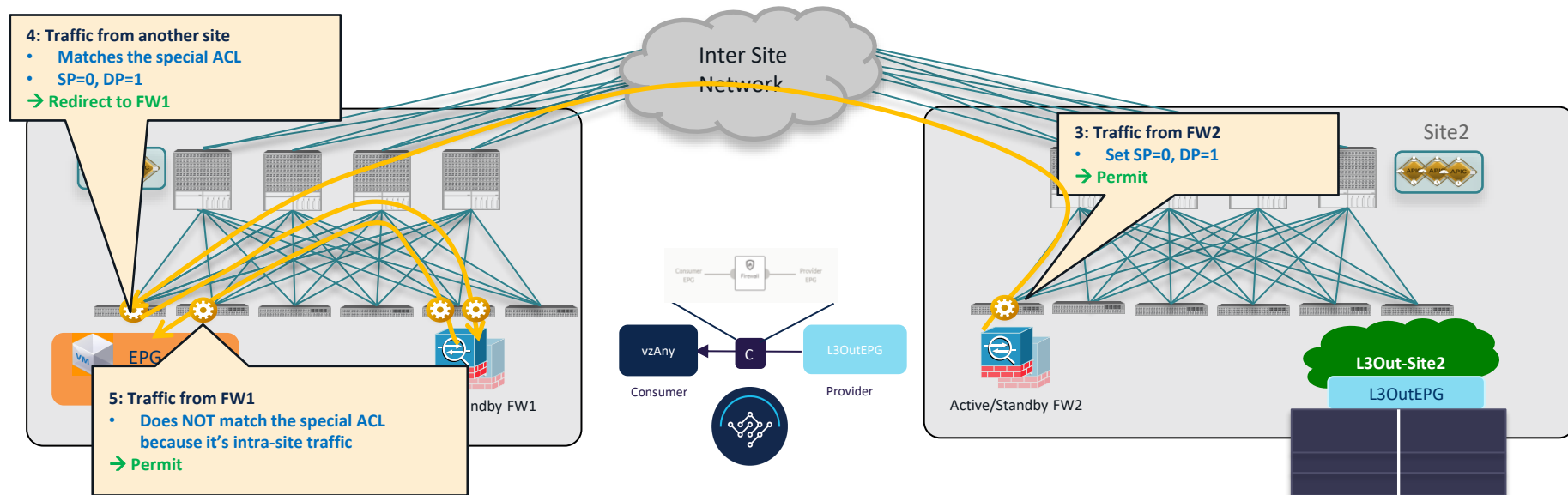
- The consumer leaf steers the traffic back to the provider's site to be inspected by the service device there
- When the destination leaf receives again the flow from the service device in Site2, it's just forwarded to the destination



vzAny-to-L3OutEPG PBR Use Case

What if the Provider Leaf doesn't Know the Consumer Endpoint? (2/3)

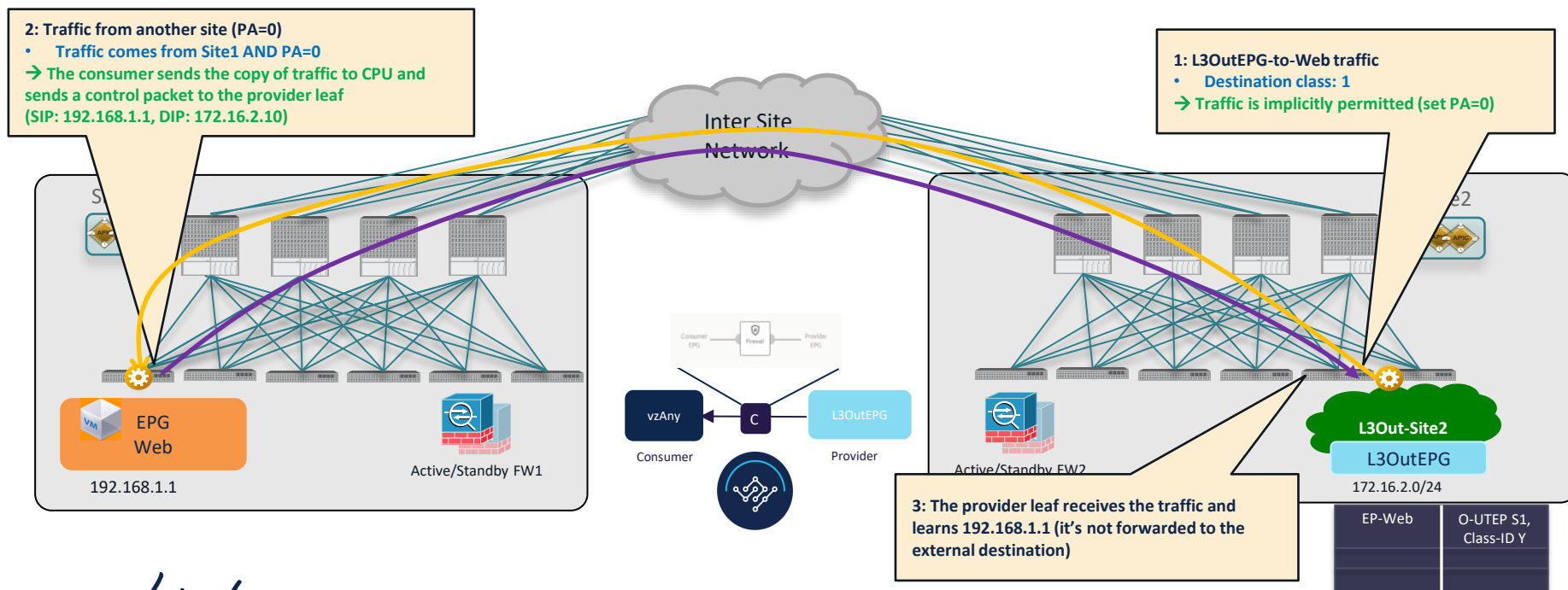
- The consumer leaf steers the traffic back to the provider's site to be inspected by the service device there
- When the destination leaf receives again the flow from the service device in Site2, it's just forwarded to the destination



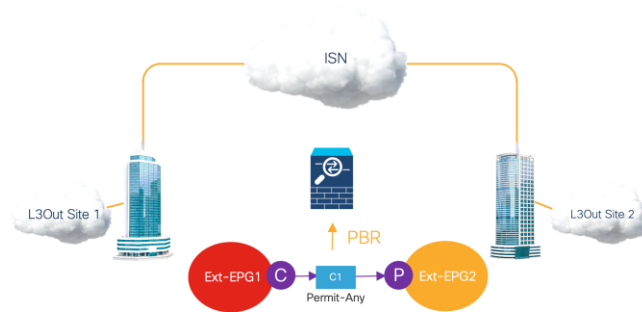
vzAny-to-L3OutEPG PBR Use Case

What if the Provider Leaf doesn't Know the Consumer Endpoint? (3/3)

- Conversational Learning is activated to ensure that the provider leaf can learn the consumer's EP information
 - This removes the suboptimal bouncing of traffic across sites



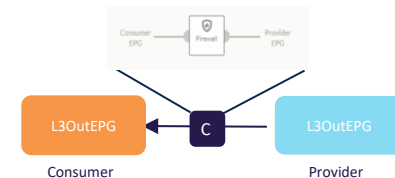
Intersite Transit Routing PBR Use Case



L3OutEPG-to-L3OutEPG PBR Use Case

NDO 4.2(3)/ACI 6.0(4)

Deployment Considerations



- The goal is redirecting to a Firewall Service all the intersite transit routing flows (L3Out-to-L3Out)
- Single service node insertion only in NDO 4.2(3)
- Support for intra-VRF and inter-VRF communication
- How to avoid asymmetric traffic paths through different FW nodes?
 - Redirecting “inter-site” traffic to the Firewall services in both the source and destination fabrics (same as for the any-to-any PBR use case)
 - Only redirection to the local Firewall service for intra-fabric transit routing flows
- The ingress leaf nodes can always resolve the destination class-ID based on L3OutEPG subnet configuration
 - Conversational learning is NOT required

L3OutEPG-to-L3OutEPG PBR Use Case

NDO 4.2(3)/ACI 6.0(4)

Consumer to Provider Direction

- Redirect “inter-site” traffic in both ingress and egress sites



SP=0, DP=1
for traffic from
the service EPG

The ingress Leaf Node can always resolve the class ID of the destination L3OutEPG and apply the PBR policy redirecting the traffic to the local FW

1: L3OutEPG1-to-L3OutEPG2 traffic

- Apply the PBR policy
→ Redirect to FW1

2: TCP traffic from FW1

- Set SP=0, DP=1
→ Permit

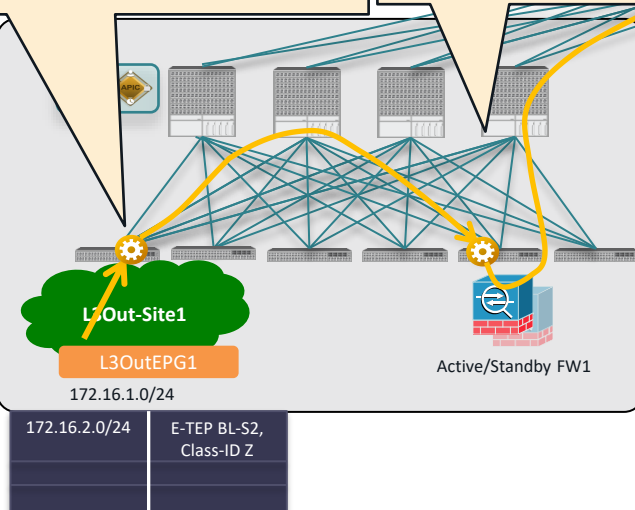
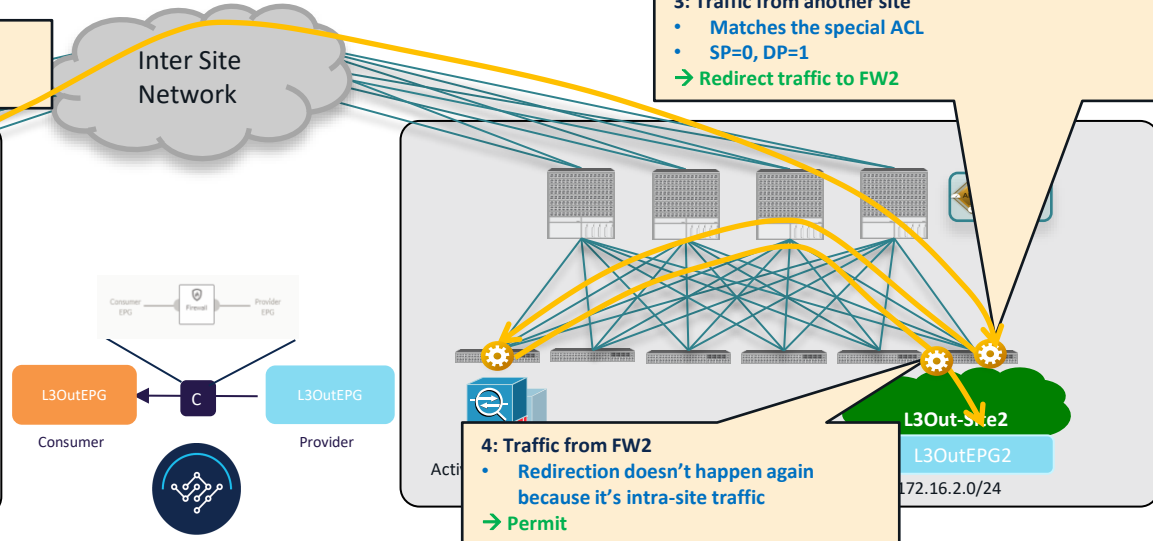
Inter Site
Network

3: Traffic from another site

- Matches the special ACL
- SP=0, DP=1
→ Redirect traffic to FW2

4: Traffic from FW2

- Redirection doesn't happen again
because it's intra-site traffic
→ Permit



L3OutEPG-to-L3OutEPG PBR Use Case

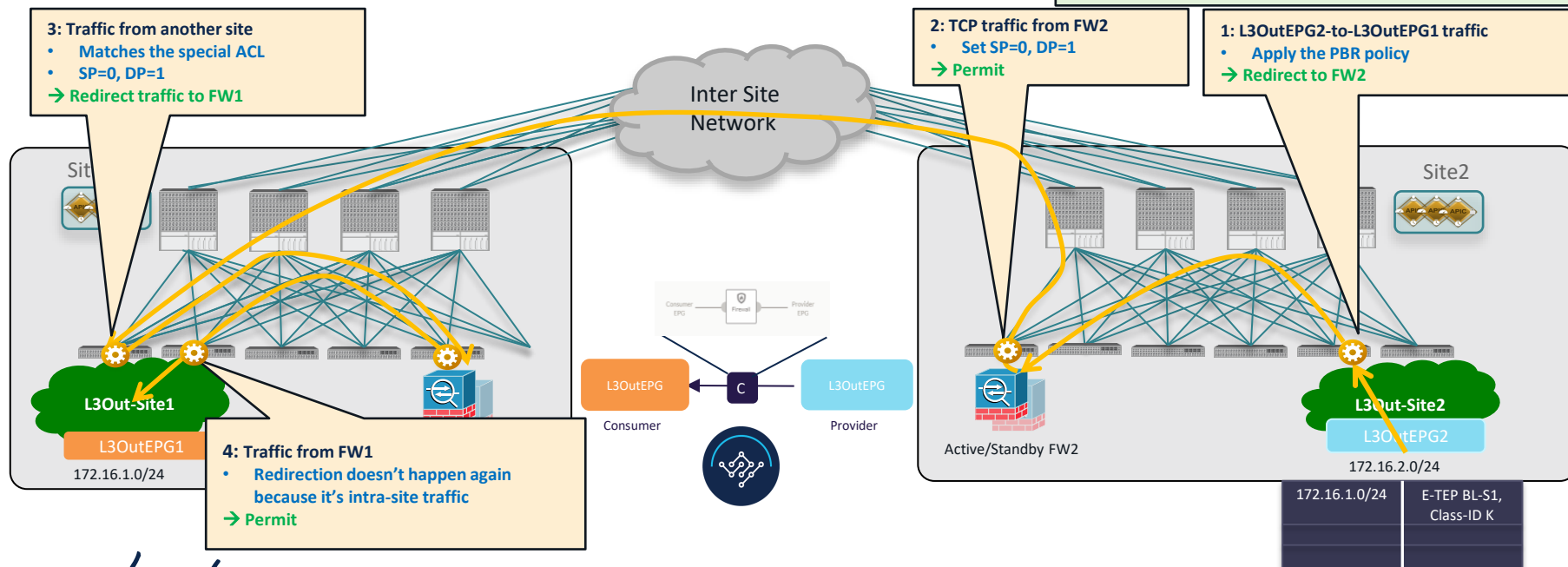
NDO 4.2(3)/ACI 6.0(4)

Provider to Consumer Direction



SP=0, DP=1
for traffic from
the service EPG

The ingress Leaf Node can always resolve the class ID of the destination L3OutEPG and apply the PBR policy redirecting the traffic to the local FW



Where to Go for More Information

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>

Complete Your Session Evaluations



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



Complete your surveys in the **Cisco Live mobile app**.

Keynote Deep Dives

Wednesday
10:30am – 11:30am



Experiences Amplified:
How AI Can Fuel Better Employee and Customer Experiences

Level 1
Room 106



Smart, Secure, Seamless:
Transforming Experiences with Next-Generation Networking

Level 2
Room 204



Harness a Bold New Era:
Transform Data Centre and Service Provider Connectivity

Level 2
Room 203



Securing User to Application and Everything in Between

Level 2
Melbourne Room 2



Unlocking Digital Resilience through Unified Observability

The HUB
Centre Stage

Complete Your Session Evaluations



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



Complete your surveys in the **Cisco Live mobile app**.



Continue your education



- Visit the Cisco Stand for related demos
- Book your one-on-one Meet the Expert meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand



Thank you

CISCO *Live!*

#CiscoLiveAPJC

CISCO *Live!*

GO BEYOND

#CiscoLiveAPJC