cisco Live!

Let’s go

cisco The bridge to possible

ACI Multi-Site Architecture and Deployment

Max Ardica, Distinguished Engineer
@maxardica

Hall of Fame
Elite Speaker
cisco Live!

cisco Live!

BRKDCN-2980

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-Site architecture
- **Initial assumption:**
  - ✓ The audience already has a good knowledge of ACI main concepts (Tenant, BD, EPG, L2Out, L3Out, etc.)

Agenda

- Introduction
- Inter-Site Connectivity Deployment Considerations
- Nexus Dashboard Orchestrator (NDO)
- ACI Multi-Site Control and Data Plane
- Provisioning Policies on NDO
- Connecting to the External L3 Domain
- Network Services Integration (Stretch Goal)

Introduction

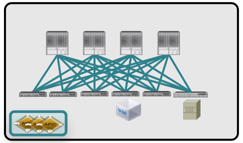
ACI Architectural Options

Fabric and Policy Domain Evolution

Single Fabric, Single Controller Domain

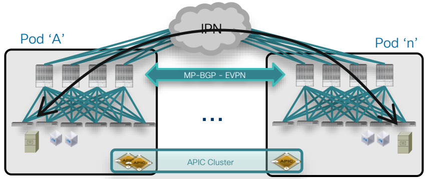
1

ACI Single Pod Fabric



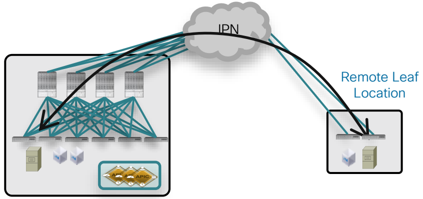
2

ACI Multi-Pod Fabric



3

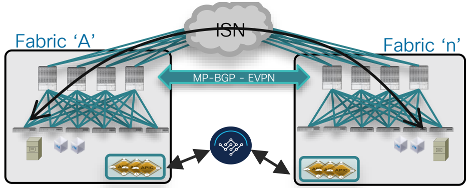
ACI Remote Leaf



Multiple Fabrics, Multiple Controller Domains

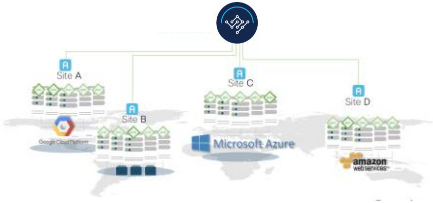
4

ACI Multi-Site



5

ACI Hybrid Cloud/Multi-Cloud



Multi-Pod or Multi-Site?

Where to Get More Information

- 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

Want to know how to provision Multi-Pod and Multi-Site from scratch? Come to BRKDCN-2919 (Wed @ 10.30 am)

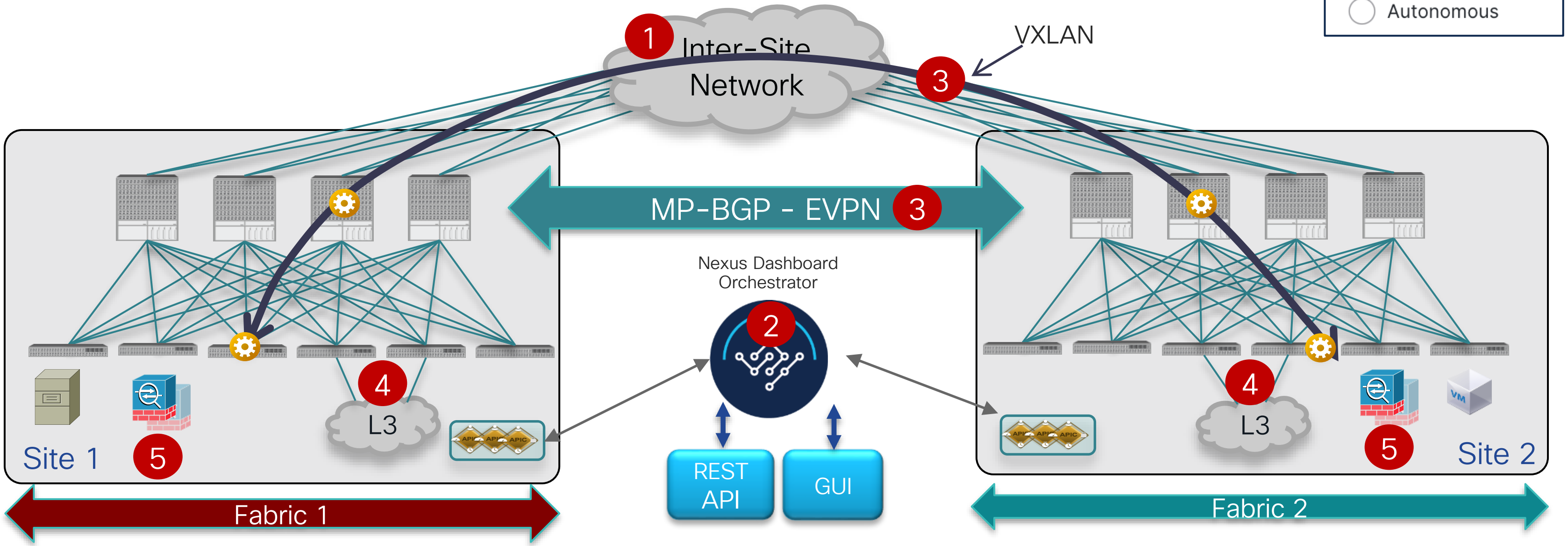
ACI Multi-Site

The Ideal Architecture for “Loosely Coupled” DCs

Deployment Mode ⓘ

● Multi-Site

○ Autonomous



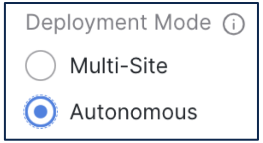
- Separate ACI Fabrics with independent APIC clusters
- No latency limitation between Fabrics
- ACI Multi-Site Orchestrator pushes cross-fabric configuration to multiple APIC clusters providing scoping of all configuration changes

- 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”

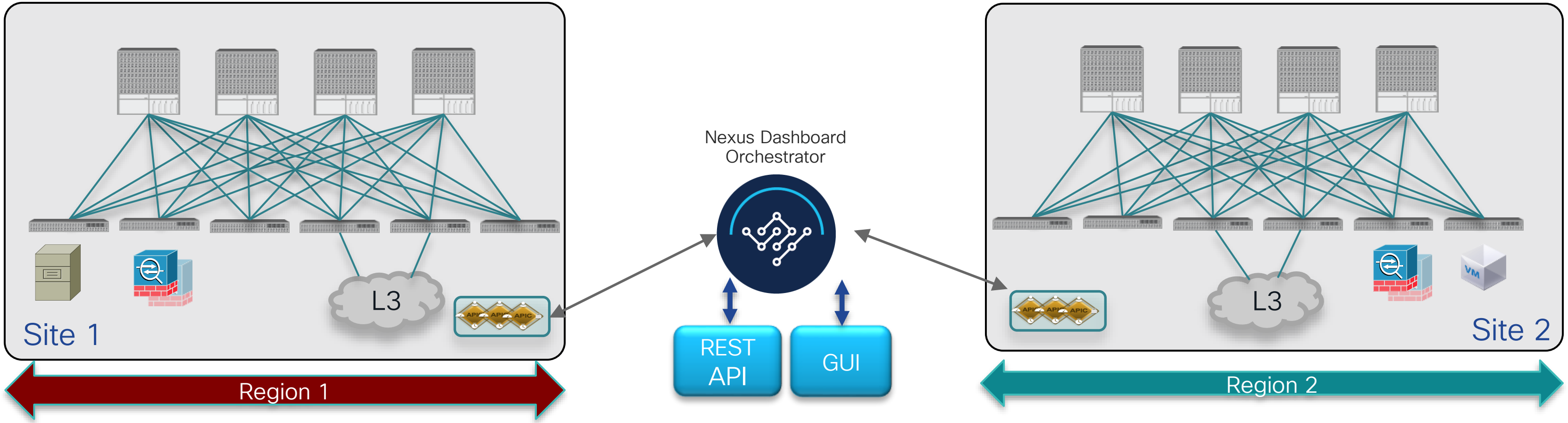
Cisco Nexus Dashboard Orchestrator 4.x



Deployment Mode ⓘ

○ Multi-Site

● Autonomous



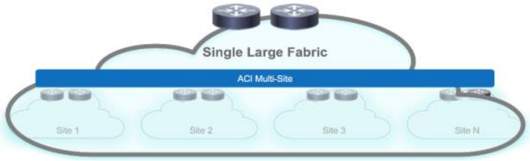
- If the fabrics are operated as independent (“autonomous”) sites, NDO could still be used as a single point of provisioning
- No use of ISN and VXLAN EVPN for east-west communication
- Intersite Layer 3 communication still possible via the L3Out data path
- NDO can be used to “replicate” configuration across sites by associating the same “autonomous template” to up to 100 fabrics

ACI Multi-Site Architecture

Most Common Use Cases

- Compartmentalization/Scale

  Building Multiple Fabrics inside a single Data Center

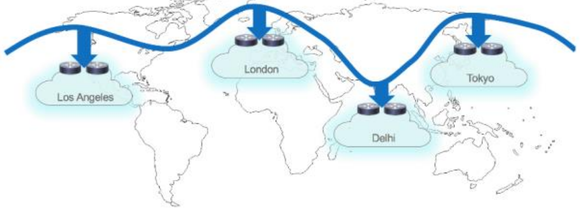


Optimized and controlled L2/L3 connectivity (including optimized/controlled BUM forwarding), scale out total number of leaf nodes (SP use case)

- Data Center Interconnect (DCI)

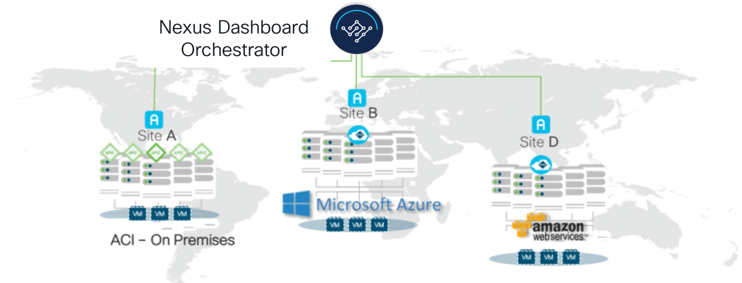
  Extend connectivity/policy between 'loosely coupled' DC sites

  Disaster Recovery and IP mobility use cases



- Hybrid-Cloud and Multi-Cloud

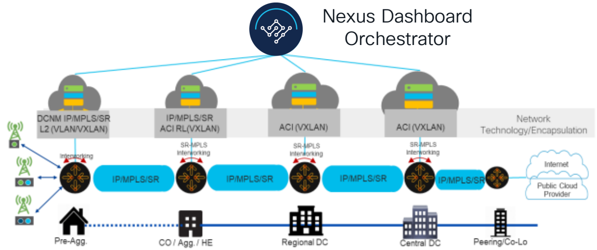
  Integration between on-prem and public clouds (AWS, Azure, GCP)



- SP 5G Telco DC/Cloud*

  Centralized DC Orchestration for "Autonomous Fabrics"

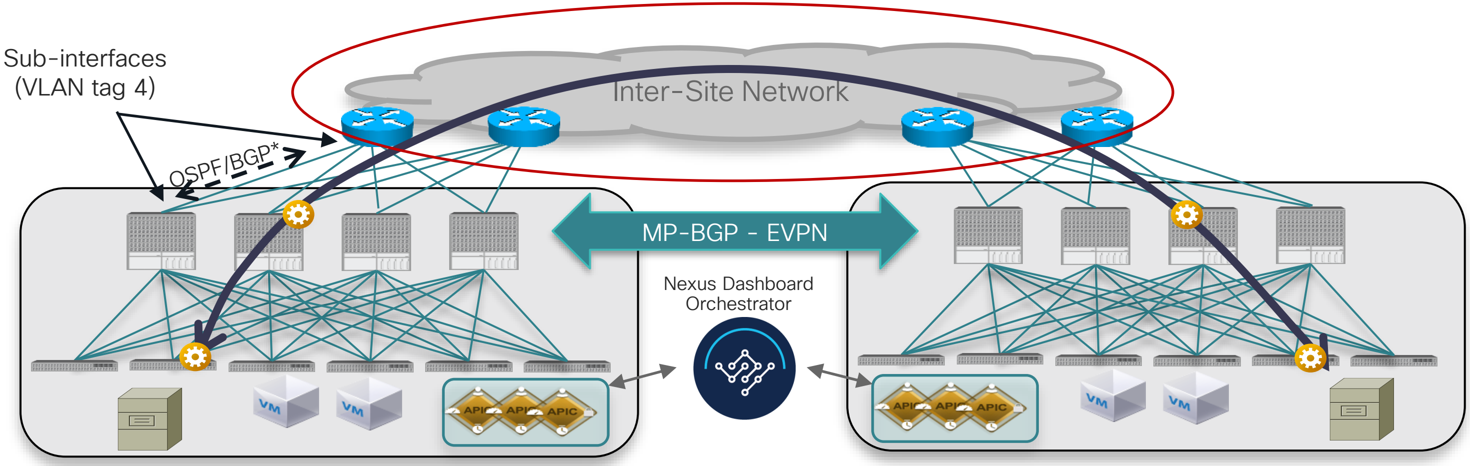
  Optional SR-MPLS/MPLS Handoff on Border Leaf nodes



*May also apply to Enterprise deployments

Inter-Site Connectivity Deployment Considerations

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)

*Requires ACI 5.2(1)

ACI Multi-Site and MTU Size

Different MTU Meanings

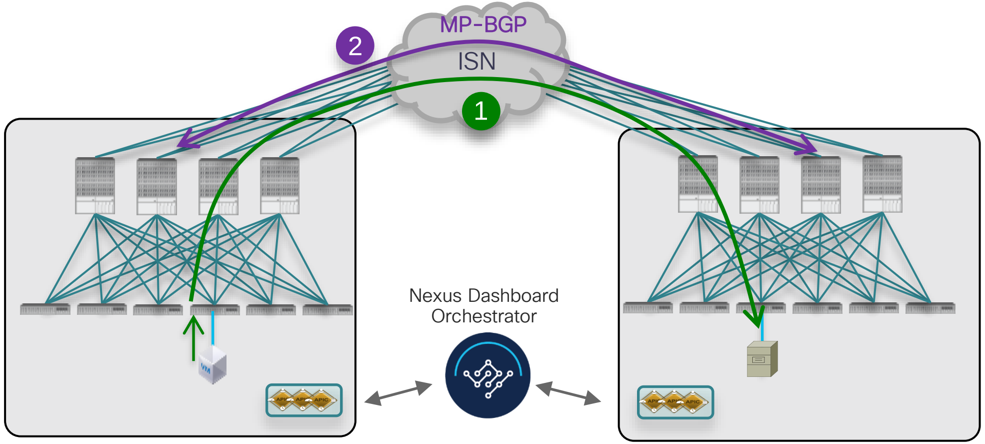
1. Data-Plane MTU: MTU of the traffic generate by endpoints (servers, routers, service nodes, etc.) connected to ACI leaf nodes

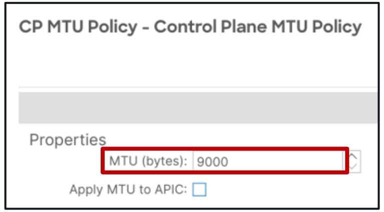
   Need to account for 50B of overhead (VXLAN encapsulation) for inter-site communication

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





CP MTU Policy - Control Plane MTU Policy

Properties

MTU (bytes): 9000

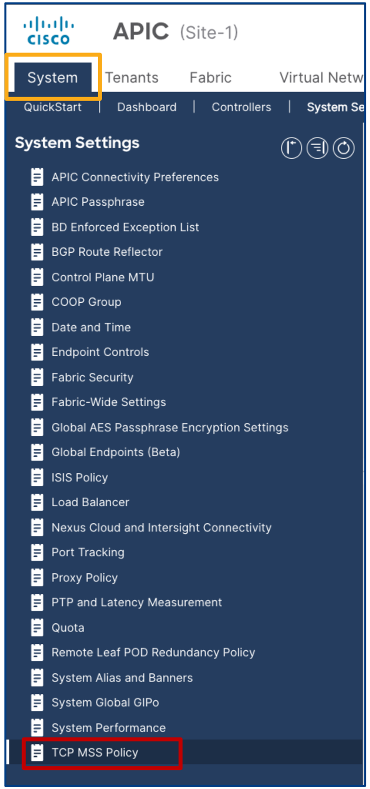
Apply MTU to APIC: □

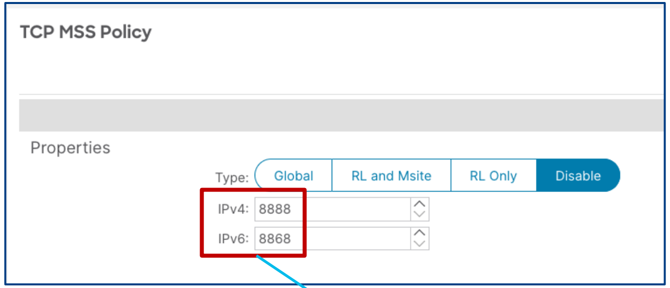
What if the ISN Supports Only 1500B MTU Size?

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

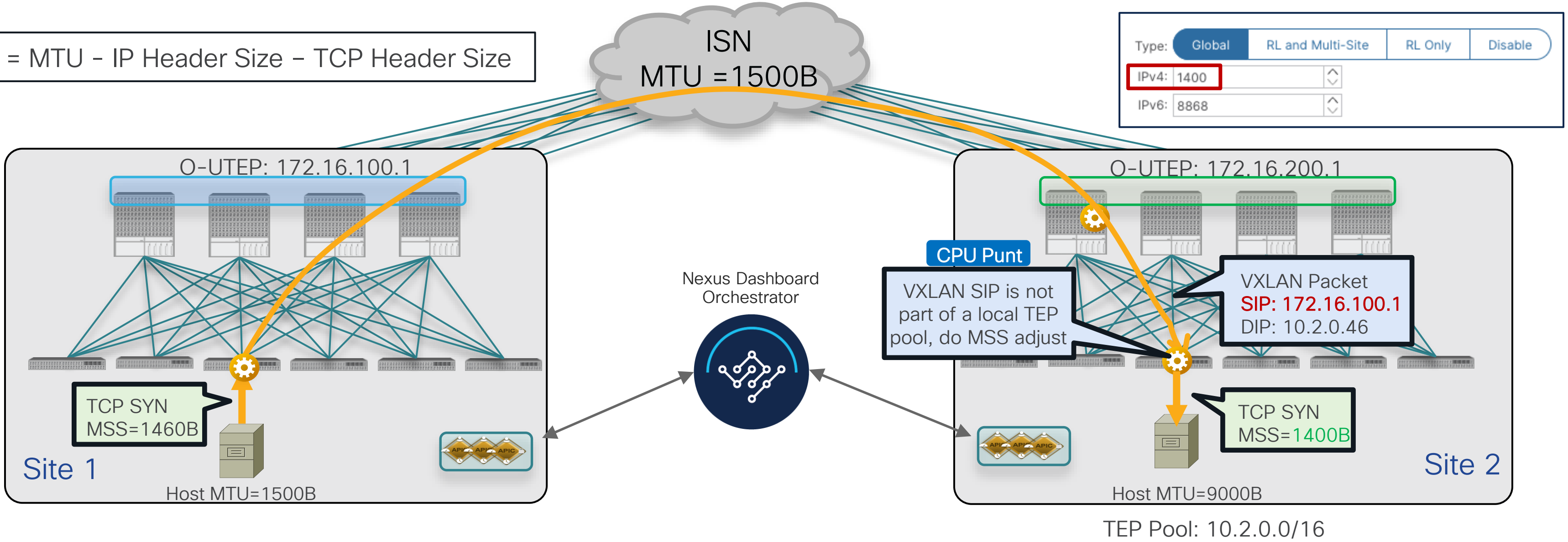
- TCP MSS adjust policy is enabled at System Settings level
- Supports different TCP MSS adjust setting for IPv4 and IPv6
- Supports three different options:
  1. **Global**: applies to all flows (Multi-Pod, Multi-Site, Remote Leafs)
  2. **RL and Msite**: applies to Multi-Site and Remote Leafs flows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

TCP-MSS Adjust Functionality

SYN Packet

ACI Release 6.0(3)F

MSS = MTU – IP Header Size – TCP Header Size



Type: ● Global | RL and Multi-Site | RL Only | Disable

IPv4: 1400

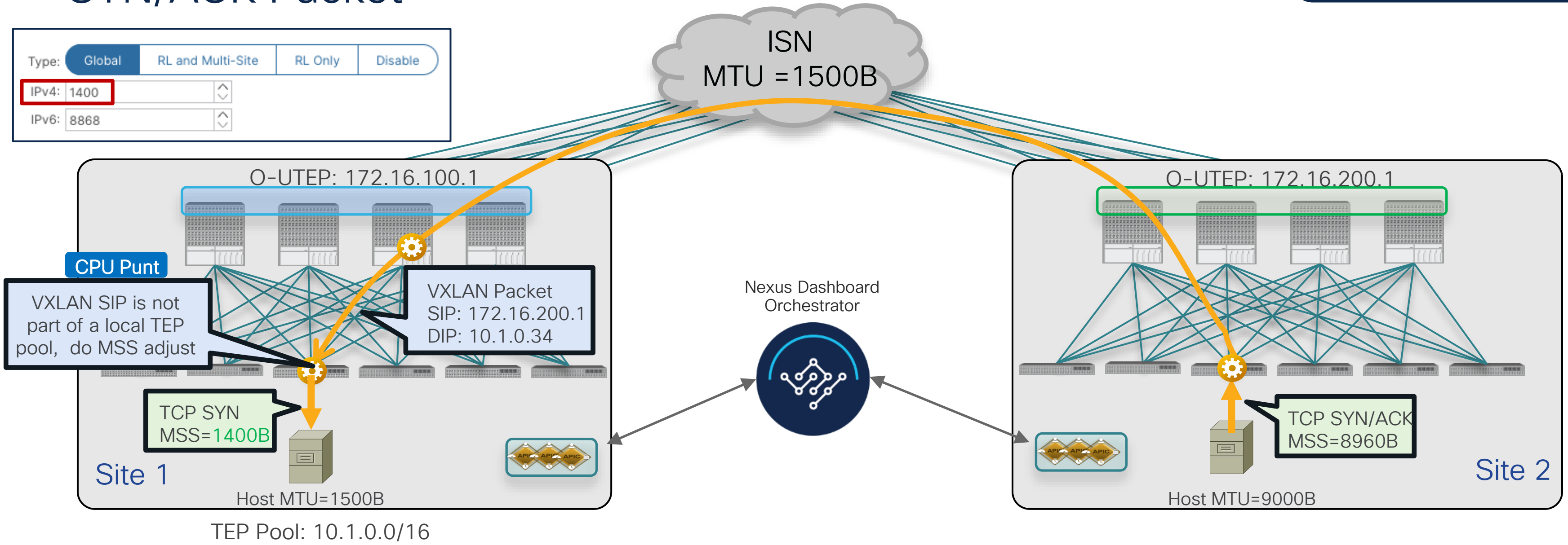
IPv6: 8868

- 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



Type: ● Global ○ RL and Multi-Site ○ RL Only ○ Disable

IPv4: 1400

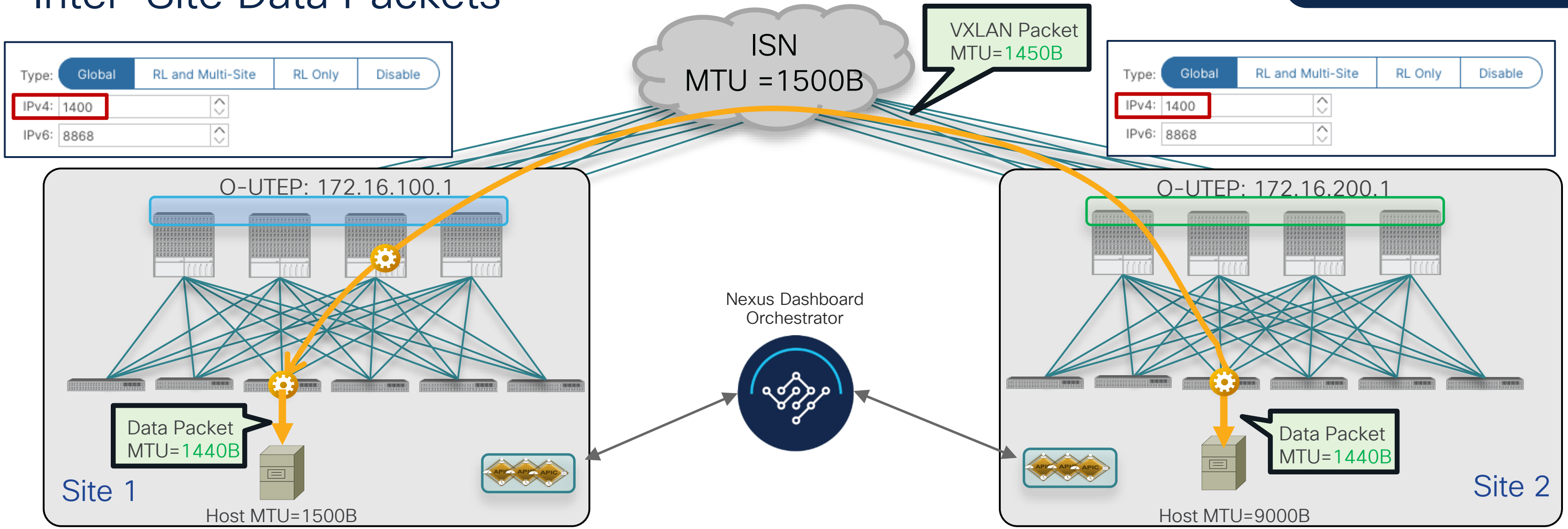
IPv6: 8868

- 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

Inter-Site Data Packets

ACI Release 6.0(3)F



Type: ● Global ○ RL and Multi-Site ○ RL Only ○ Disable
IPv4: 1400
IPv6: 8868

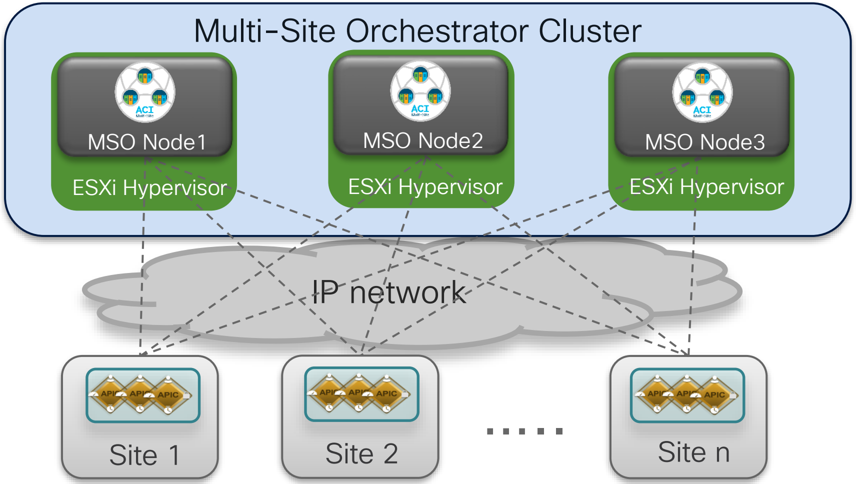
Type: ● Global ○ RL and Multi-Site ○ RL Only ○ Disable
IPv4: 1400
IPv6: 8868

- 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

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

ACI
Multi-Site

Cisco Multi-Site Orchestrator

Cisco Nexus Dashboard Orchestrator

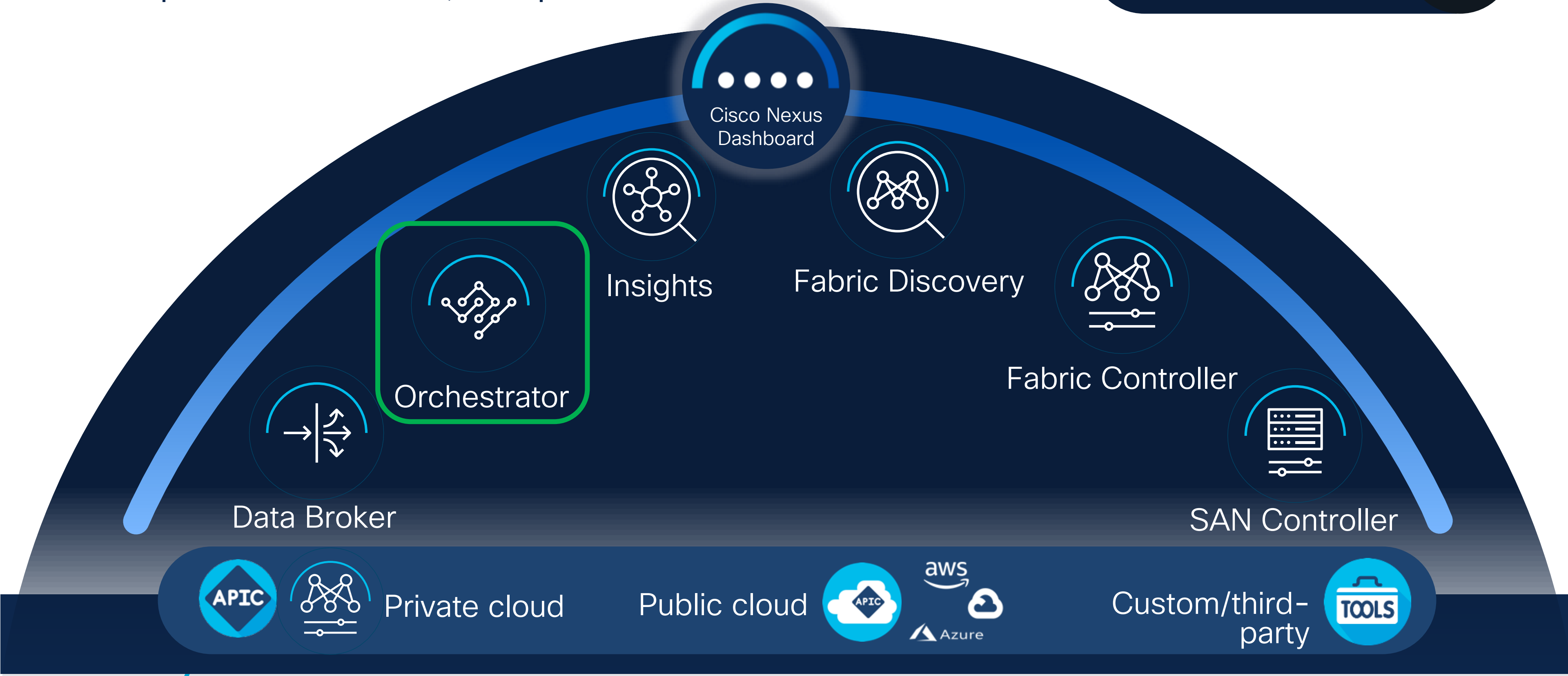
Up to release 3.1(1)

From release 3.2(1)

Cisco Nexus Dashboard

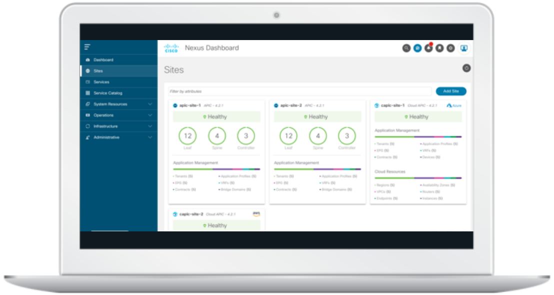
Simple to Automate, Simple to Consume

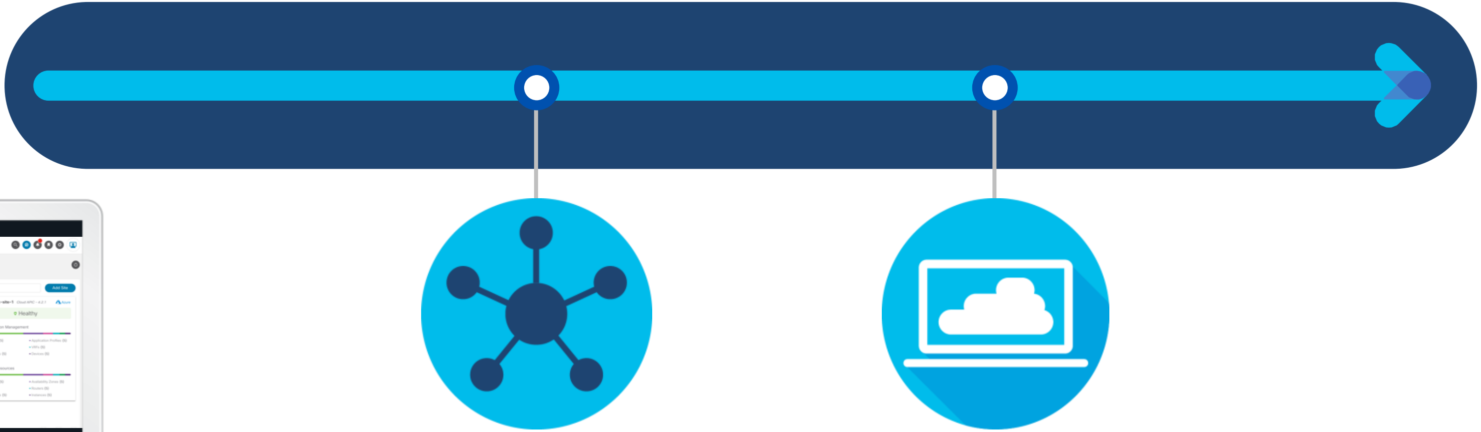
Powering automation
Unified agile platform



Cisco Nexus Dashboard

Deployment Evolution





Physical Cisco ND Platform Cluster

Virtual/Cloud Cisco ND Platform Cluster

ND virtual cluster supported on ESXi and KVM hypervisors

Spec: 16 vCPUs, 64Gb RAM and 500Gb disk

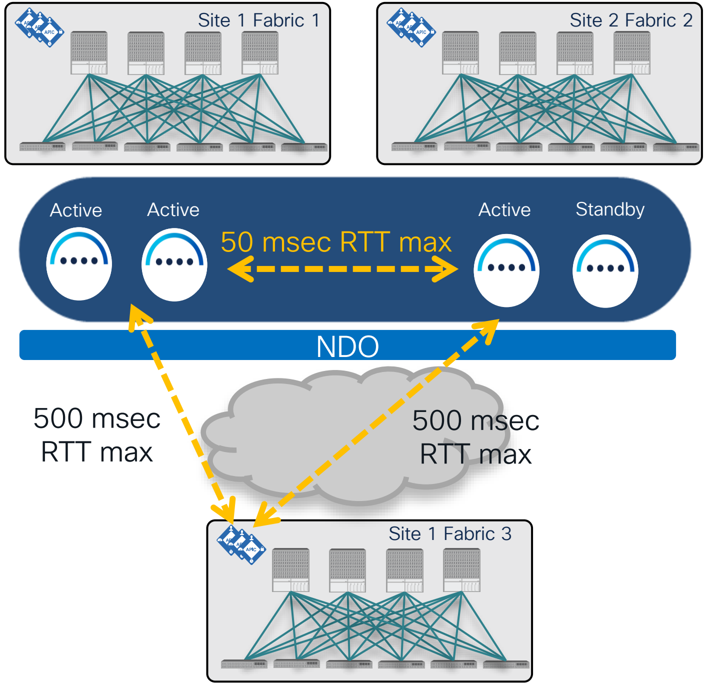
ND cloud cluster supported for AWS and Azure

Nexus Dashboard Orchestrator

Distributed ND Cluster Deployment for NDO

- At least 2 ND active nodes are needed to keep the cluster operational
- When distributing an ND cluster across DC locations, deployment of a standby is recommended

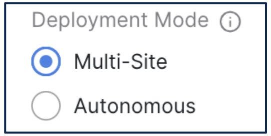
  In case of concurrent failure of 2 ND active nodes, the standby node can be activated to replace a failed node and restore the cluster's health
- Maximum supported latency values
  - ➢ 50 msec RTT: between ND nodes
  - ➢ 500 msec RTT: between an ND node and an APIC node



ACI Multi-Site Control- and Data-Plane

ACI Multi-Site

Network and Identity Extended between Fabrics



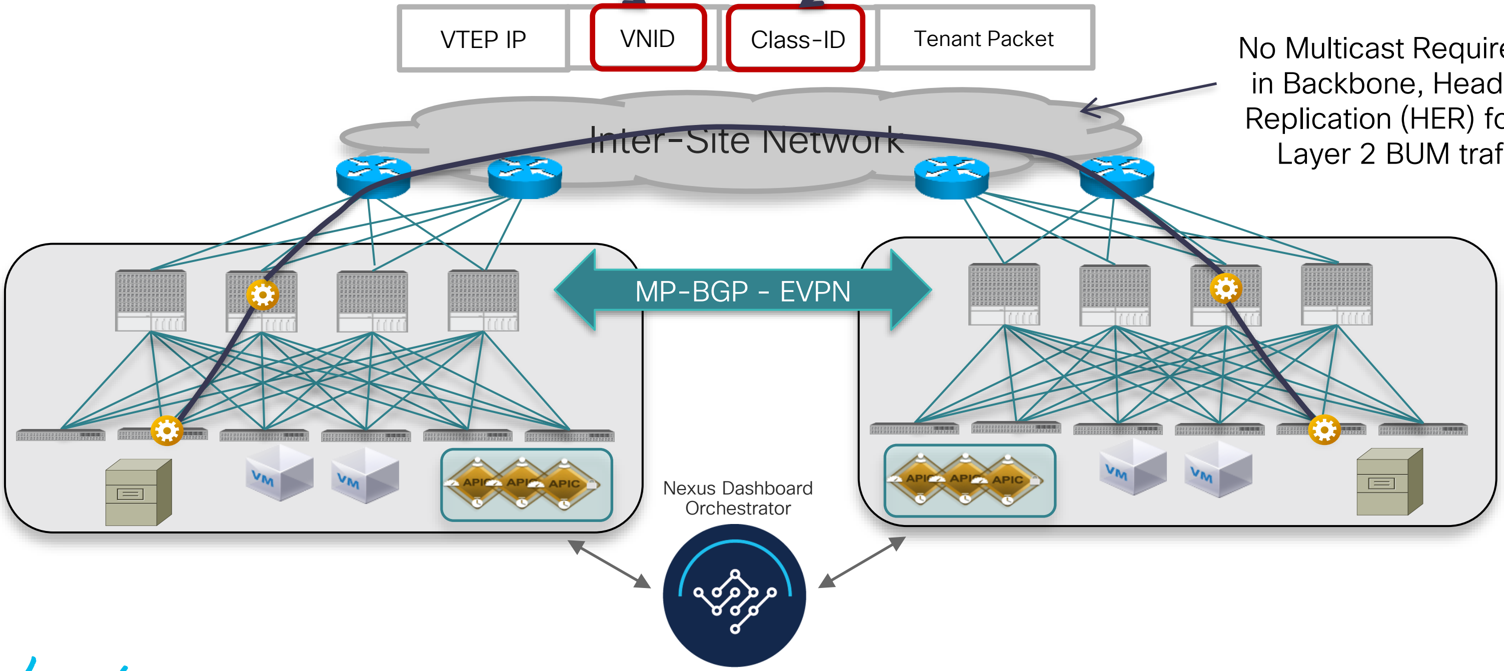
Deployment Mode ⓘ

● Multi-Site

○ Autonomous

Network information carried across Fabrics (Availability Zones)

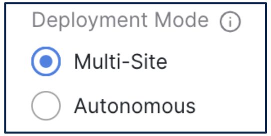
Identity information carried across Fabrics (Availability Zones)



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

ACI Multi-Site

Inter-Site Policies and Spines' Translation Tables



Deployment Mode ⓘ

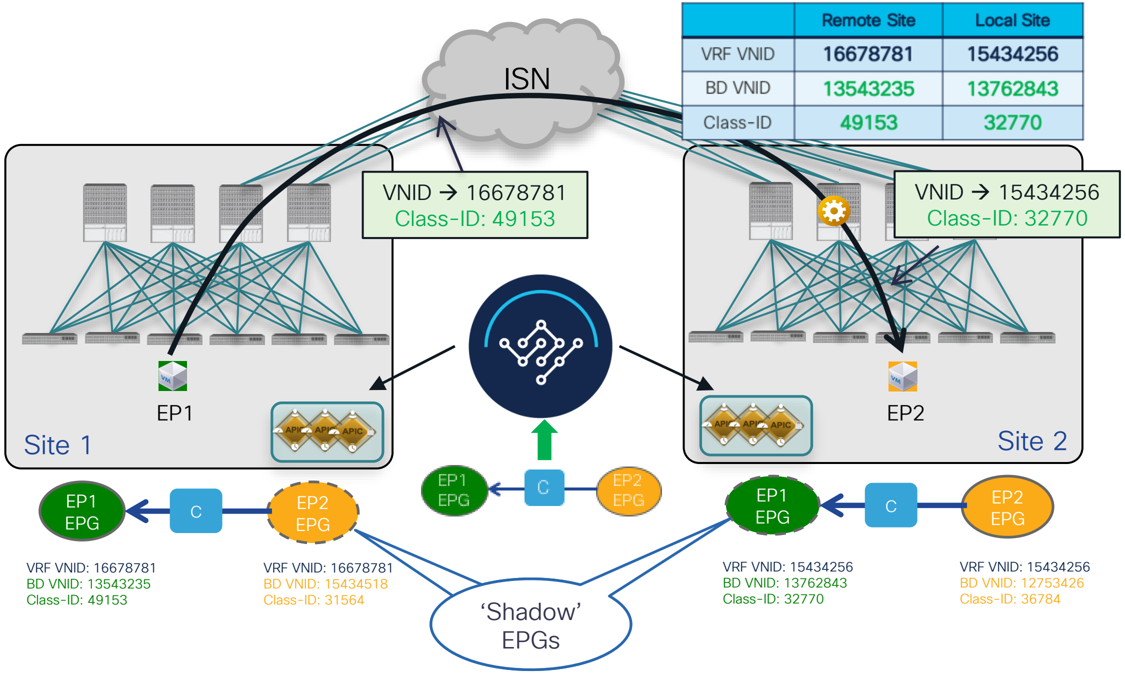
● Multi-Site

○ Autonomous

- 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**

Site 2 Spines Translation Table

	Remote Site	Local Site
VRF VNID	16678781	15434256
BD VNID	13543235	13762843
Class-ID	49153	32770



VRF VNID: 16678781
BD VNID: 13543235
Class-ID: 49153

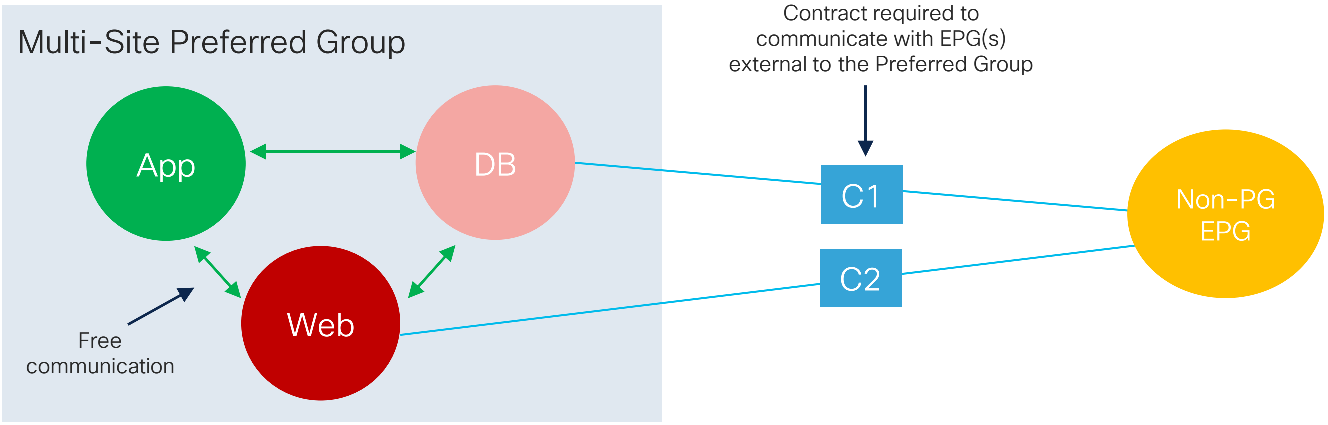
VRF VNID: 16678781
BD VNID: 15434518
Class-ID: 31564

VRF VNID: 15434256
BD VNID: 13762843
Class-ID: 32770

VRF VNID: 15434256
BD VNID: 12753426
Class-ID: 36784

ACI Multi-Site

Simplify Policy Enforcement: Preferred Groups

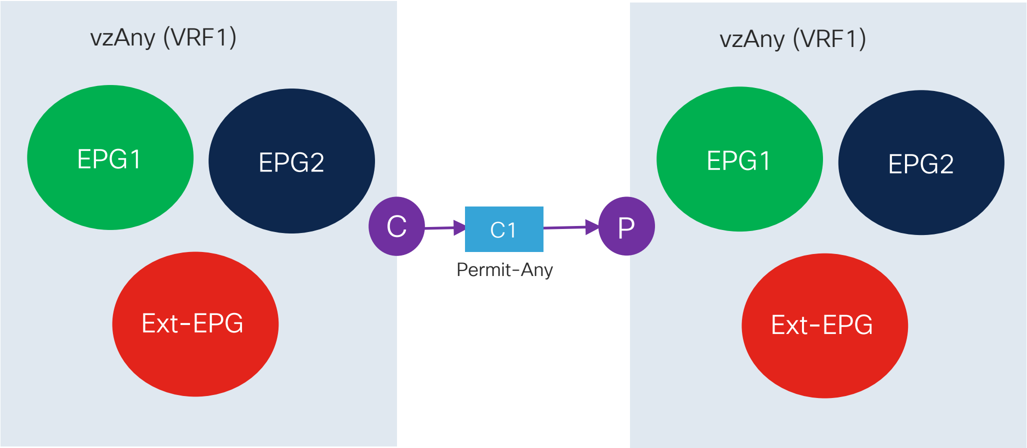

Multi-Site Preferred Group
App
DB
Web
Free communication
Contract required to communicate with EPG(s) external to the Preferred Group
C1
C2
Non-PG EPG

- “VRF unenforced” not supported with Multi-Site
- Multi-Site Preferred Group configuration can be provision 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

Enabling Free Communication inside a VRF

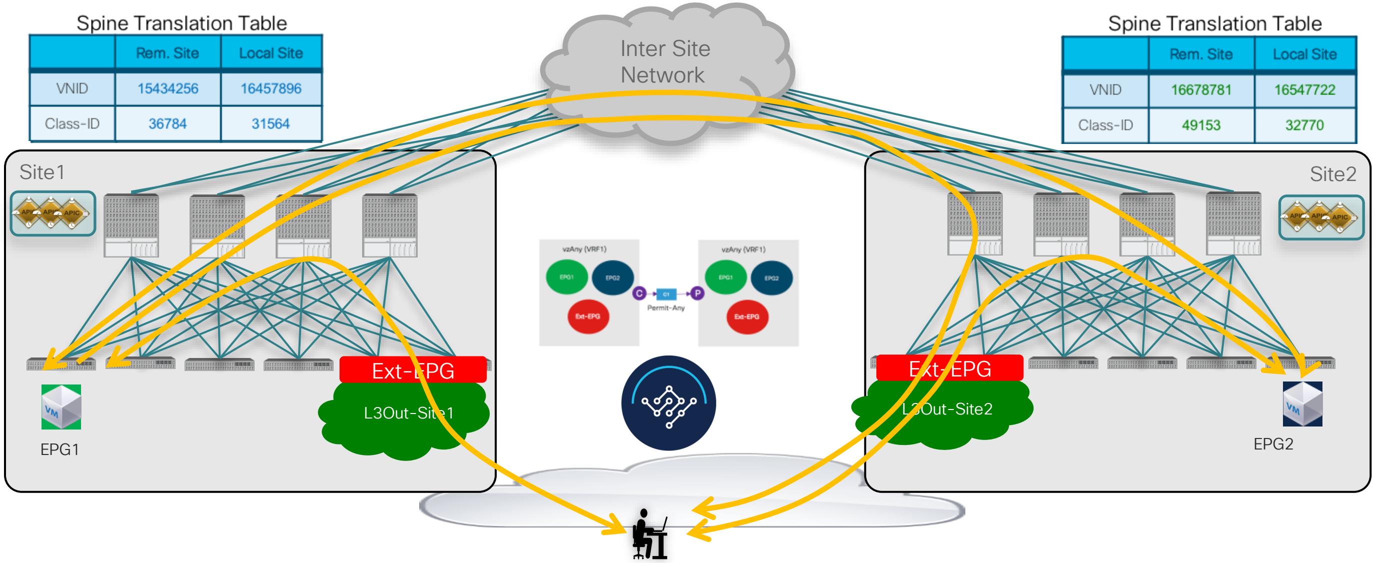
What is vzAny? Logical object representing all the EPGs/Ext-EPGs in a VRF



- 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



Spine Translation Table

	Rem. Site	Local Site
VNID	15434256	16457896
Class-ID	36784	31564

Spine Translation Table

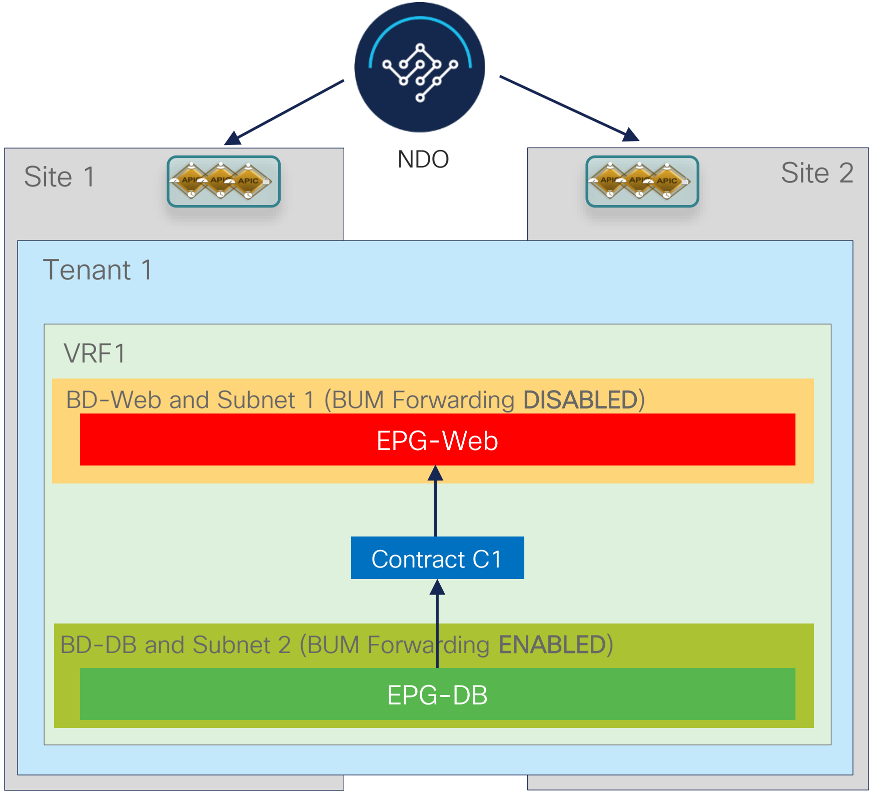
	Rem. Site	Local Site
VNID	16678781	16547722
Class-ID	49153	32770

- 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
- vzAny + PBR support available from NDO release 4.2(3) and ACI release 6.0(4)

Per Bridge Domain Behavior

ACI Multi-Site

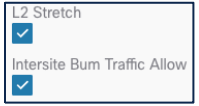
Layer 2 Extension across Sites



NDO
Site 1
Site 2
APIC APIC APIC
APIC APIC APIC
Tenant 1
VRF1
BD-Web and Subnet 1 (BUM Forwarding **DISABLED**)
EPG-Web
Contract C1
BD-DB and Subnet 2 (BUM Forwarding **ENABLED**)
EPG-DB

- Stretch tenant/VRF but also BDs/EPGs across ACI fabrics
- BUM forwarding can be controlled on a BD basis

Required only for establishing pure L2 communication across sites (DB clustering using L2 multicast or broadcast, for example)

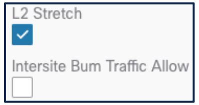


L2 Stretch ☑

Intersite Bum Traffic Allow ☑

BD-DB

IP mobility (and live migration) can be supported **without** enabling BUM forwarding



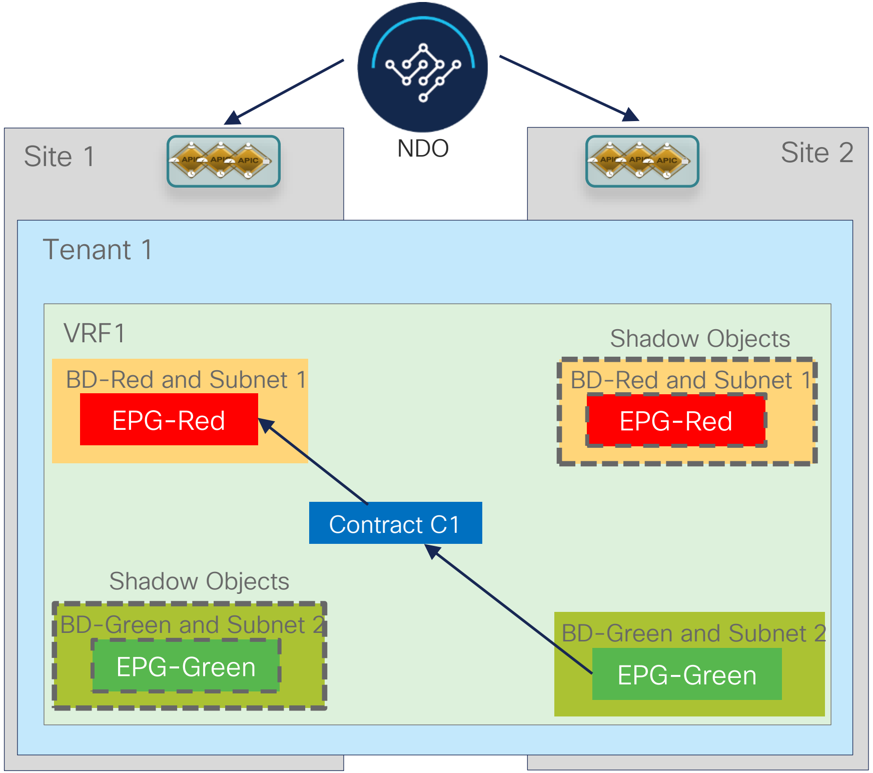
L2 Stretch ☑

Intersite Bum Traffic Allow ☐

BD-Web

ACI Multi-Site

Intra-VRF Layer 3 Communication across Sites



NDO

Site 1

Site 2

Tenant 1

VRF1

BD-Red and Subnet 1

EPG-Red

Shadow Objects

BD-Red and Subnet 1

EPG-Red

Contract C1

Shadow Objects

BD-Green and Subnet 2

EPG-Green

BD-Green and Subnet 2

EPG-Green

- Stretch tenant/VRF across ACI fabrics
- BDs/EPGs defined as site local objects

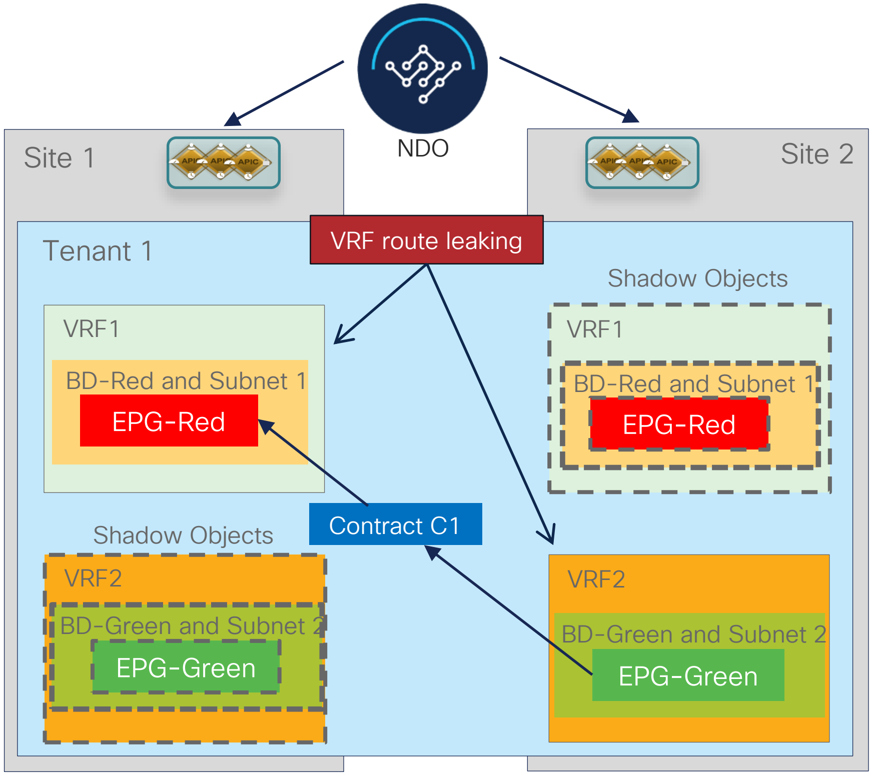
L2 Stretch □

BD-Red and BD-Green

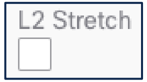
- Configuration of policy between EPGs in separate fabrics to enable intra-VRF Layer 3 inter-site connectivity
- Creation of shadow BDs/EPGs in remote site(s)

ACI Multi-Site

Inter-VRF Layer Communication across Sites (Shared Services)



- VRF/BD/EPG locally defined in each site



L2 Stretch

☐

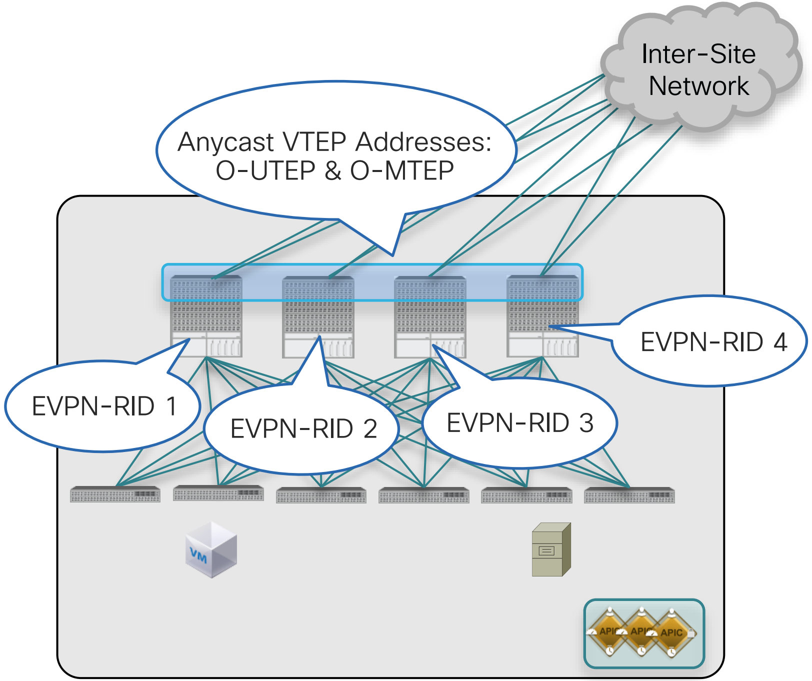
BD-Red and BD-Green

- Inter-VRF communication across sites (shared services)
- Route leaking between VRFs (requires subnet configured under the provider EPG)
- Supported within the same stretched tenant but also between different tenants
- Creation of shadow VRFs/BDs/EPGs in remote site(s)

Underlay and Overlay Control-Plane Considerations

ACI Multi-Site

BGP Inter-Site Peers

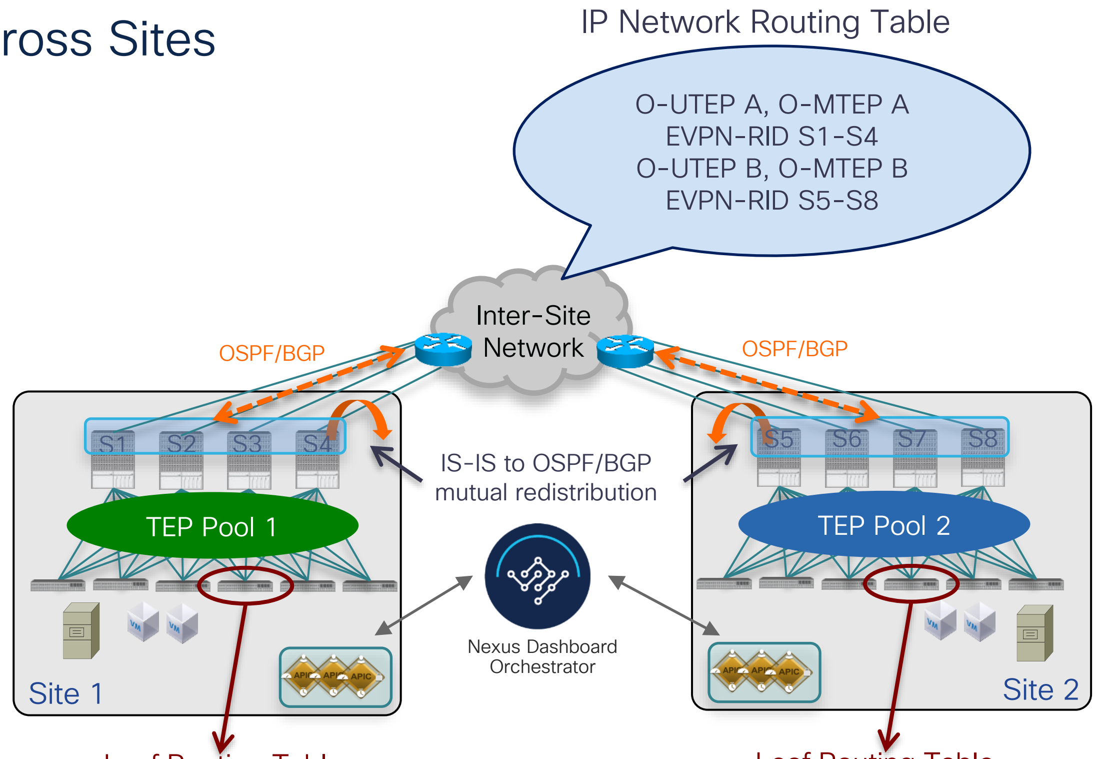


- Spines connected to the Inter-Site Network perform two main functions:
  1. Establishment of MP-BGP EVPN peerings with spines in remote sites
     - One dedicated Control-Plane address (EVPN-RID) is assigned to each spine running MP-BGP EVPN
  2. Forwarding of inter-sites data-plane traffic
     - 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
- 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 and O-MTEP) across sites
  - Use of overlapping internal TEP Pools across sites possible and fully supported



Leaf Routing Table

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

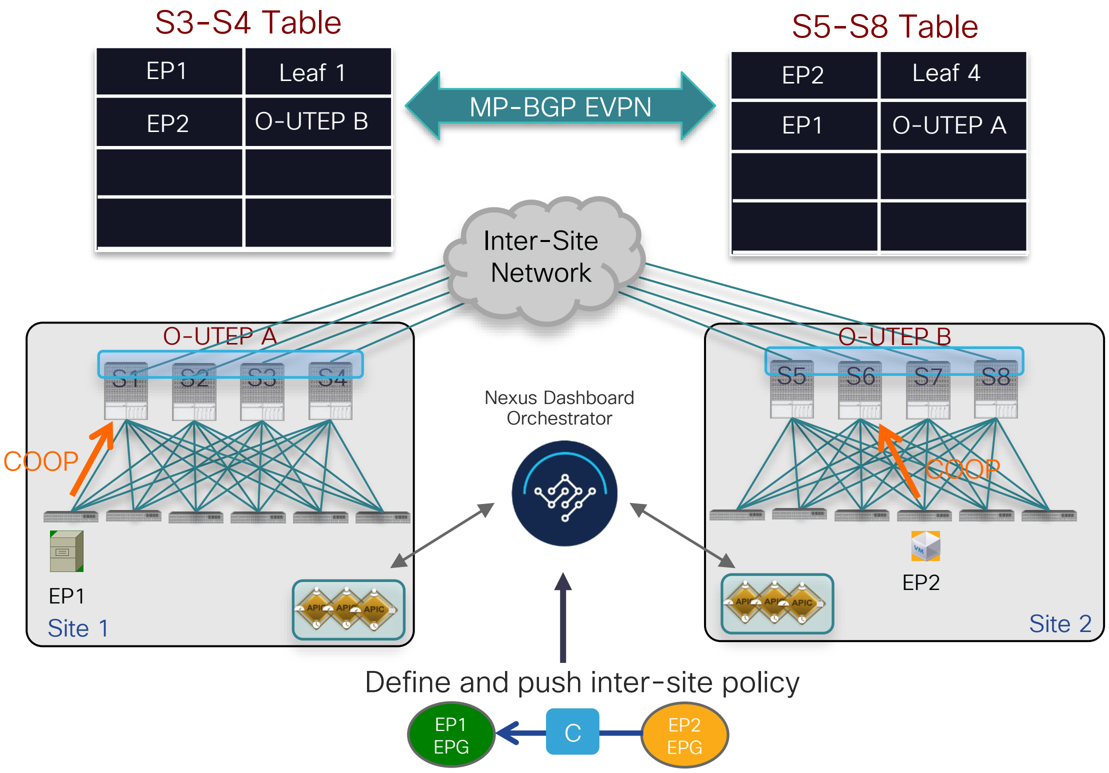
Leaf Routing Table

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

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



S3-S4 Table

EP1	Leaf 1
EP2	O-UTEP B

S5-S8 Table

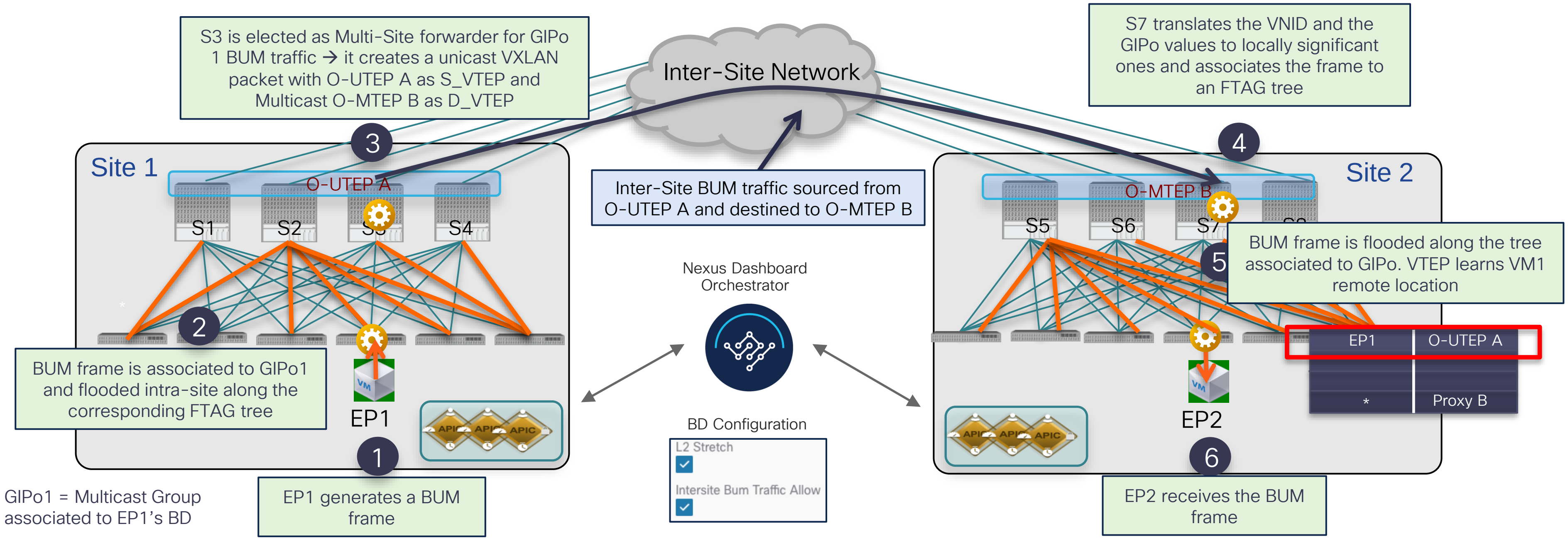
EP2	Leaf 4
EP1	O-UTEP A

Data-Plane Communication across Sites

ACI Multi-Site

Inter-Site Layer 2 BUM* Forwarding

*BUM – Broadcast, Unknown Unicast, Multicast



1 EP1 generates a BUM frame

2 BUM frame is associated to GIPo1 and flooded intra-site along the corresponding FTAG tree

3 S3 is elected as Multi-Site forwarder for GIPo 1 BUM traffic → it creates a unicast VXLAN packet with O-UTEP A as S_VTEP and Multicast O-MTEP B as D_VTEP

Inter-Site BUM traffic sourced from O-UTEP A and destined to O-MTEP B

4 S7 translates the VNID and the GIPo values to locally significant ones and associates the frame to an FTAG tree

5 BUM frame is flooded along the tree associated to GIPo. VTEP learns VM1 remote location

EP1	O-UTEP A
*	Proxy B

6 EP2 receives the BUM frame

BD Configuration

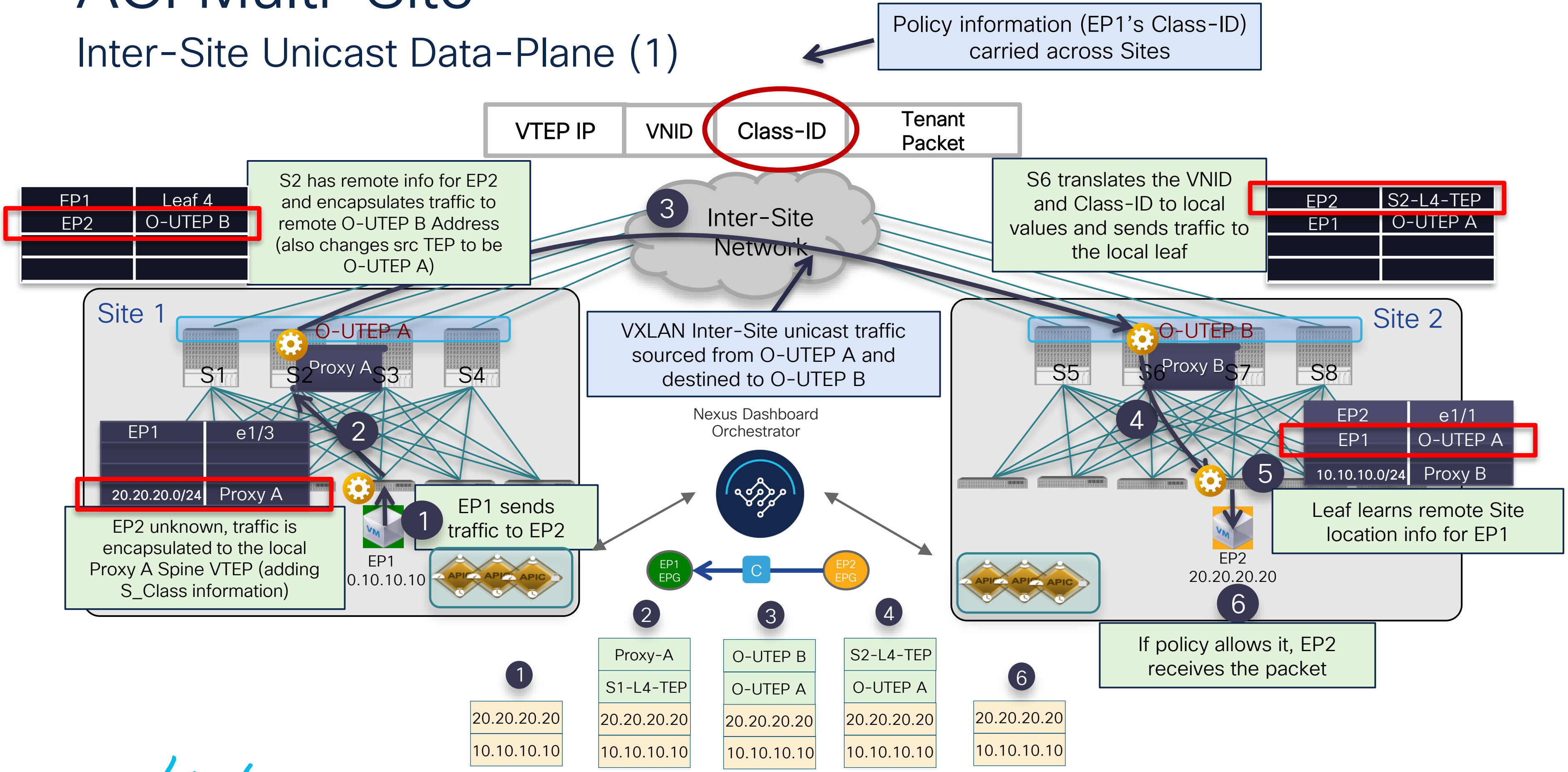
L2 Stretch
☑

Intersite Bum Traffic Allow
☑

GIPo1 = Multicast Group associated to EP1's BD

ACI Multi-Site

Inter-Site Unicast Data-Plane (1)



Policy information (EP1's Class-ID) carried across Sites

VTEP IP	VNID	Class-ID	Tenant Packet

EP1	Leaf 4
EP2	O-UTEP B

S2 has remote info for EP2 and encapsulates traffic to remote O-UTEP B Address (also changes src TEP to be O-UTEP A)

S6 translates the VNID and Class-ID to local values and sends traffic to the local leaf

EP2	S2-L4-TEP
EP1	O-UTEP A

VXLAN Inter-Site unicast traffic sourced from O-UTEP A and destined to O-UTEP B

EP1	e1/3
20.20.20.0/24	Proxy A

EP2 unknown, traffic is encapsulated to the local Proxy A Spine VTEP (adding S_Class information)

EP1 sends traffic to EP2

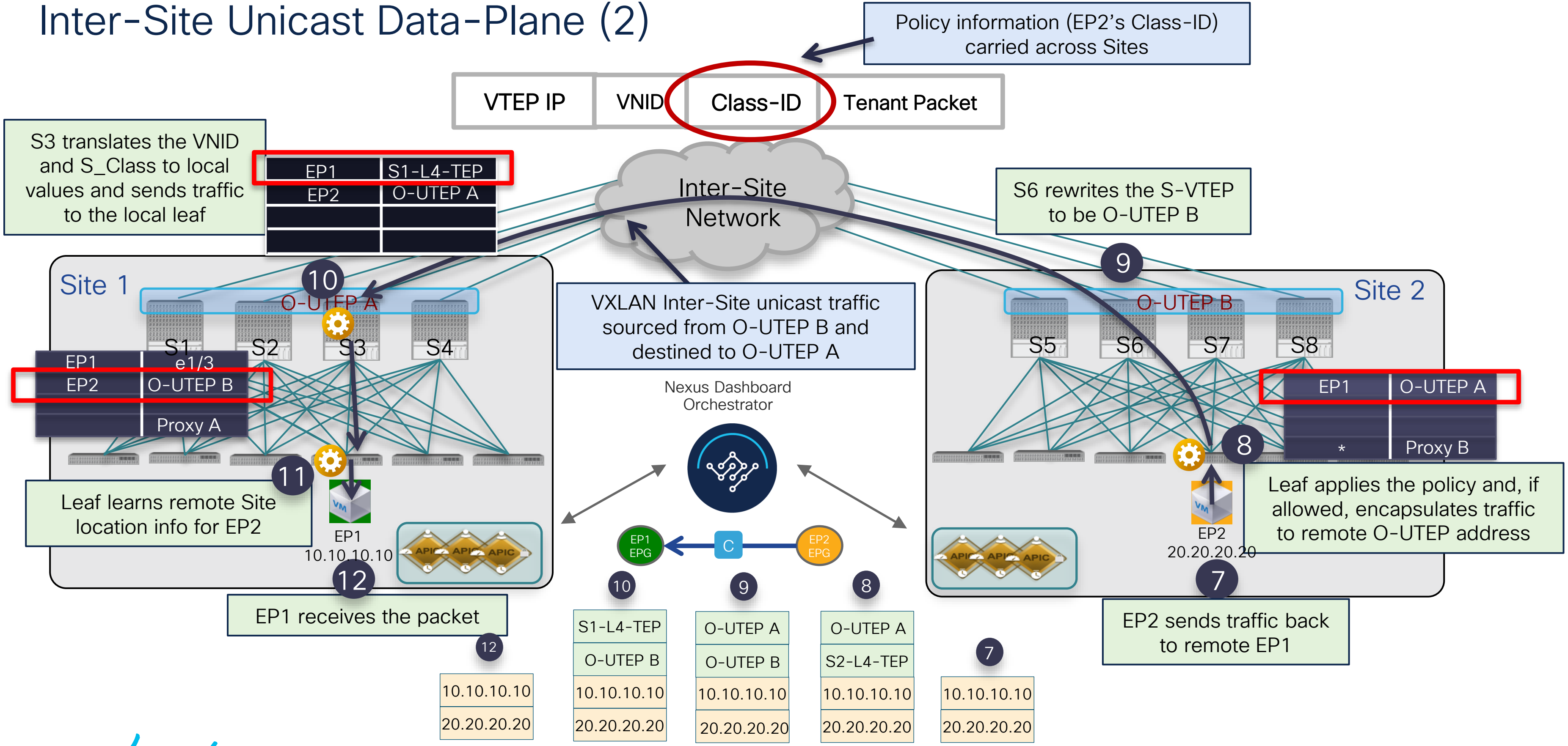
EP2	e1/1
EP1	O-UTEP A
10.10.10.0/24	Proxy B

Leaf learns remote Site location info for EP1

If policy allows it, EP2 receives the packet

ACI Multi-Site

Inter-Site Unicast Data-Plane (2)



Policy information (EP2's Class-ID) carried across Sites

VTEP IP	VNID	Class-ID	Tenant Packet

S3 translates the VNID and S_Class to local values and sends traffic to the local leaf

EP1	S1-L4-TEP
EP2	O-UTEP A

S6 rewrites the S-VTEP to be O-UTEP B

VXLAN Inter-Site unicast traffic sourced from O-UTEP B and destined to O-UTEP A

EP1	e1/3
EP2	O-UTEP B
	Proxy A

EP1	O-UTEP A
*	Proxy B

Leaf learns remote Site location info for EP2

Leaf applies the policy and, if allowed, encapsulates traffic to remote O-UTEP address

EP1 receives the packet

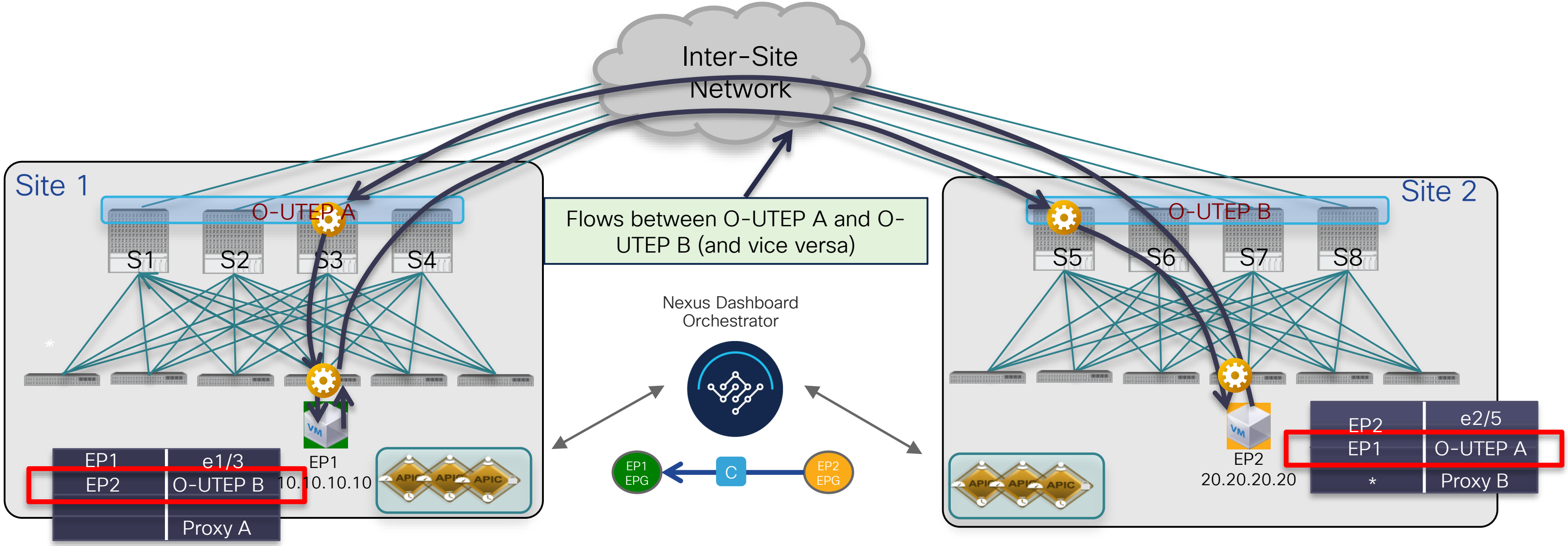
EP2 sends traffic back to remote EP1

12	10	9	8	7
	S1-L4-TEP	O-UTEP A	O-UTEP A	
	O-UTEP B	O-UTEP B	S2-L4-TEP	
10.10.10.10	10.10.10.10	10.10.10.10	10.10.10.10	10.10.10.10
20.20.20.20	20.20.20.20	20.20.20.20	20.20.20.20	20.20.20.20

ACI Multi-Site

Inter-Site Unicast Data-Plane (3)

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



EP1	e1/3
EP2	O-UTEP B
	Proxy A

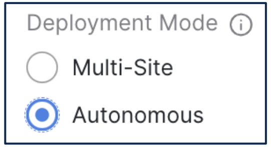
EP2	e2/5
EP1	O-UTEP A
*	Proxy B

cisco *Live!*

Layer 3 Only Communication between Autonomous Sites

ACI Multi-Site

L3 Only across Sites (“Autonomous Sites”)

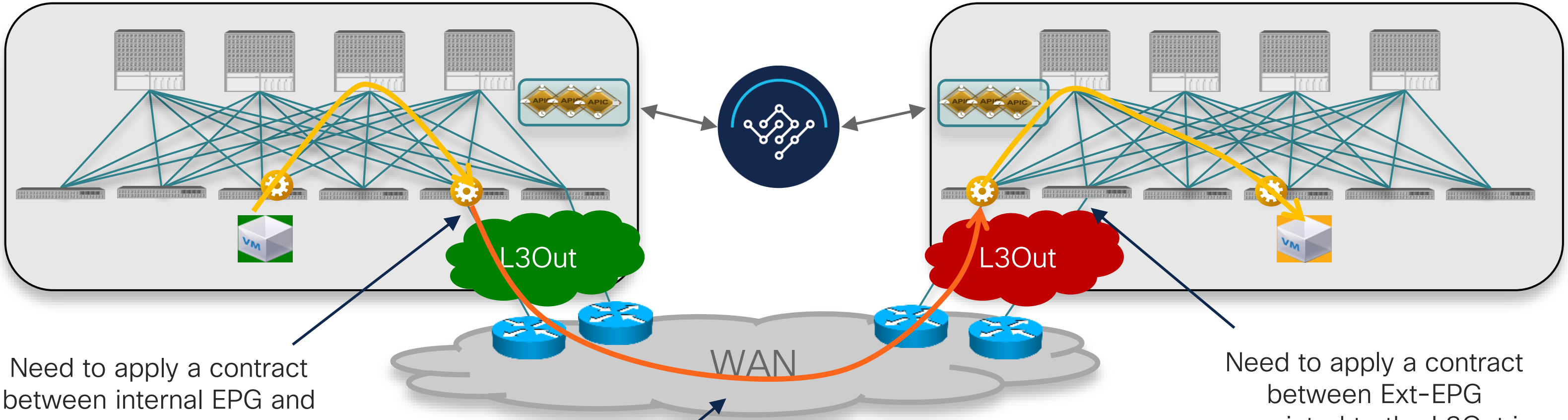


Deployment Mode ⓘ

○ Multi-Site

● Autonomous

- Autonomous deployment mode, NDO used as for ”configuration replication”
- Routing across sites via the WAN backbone



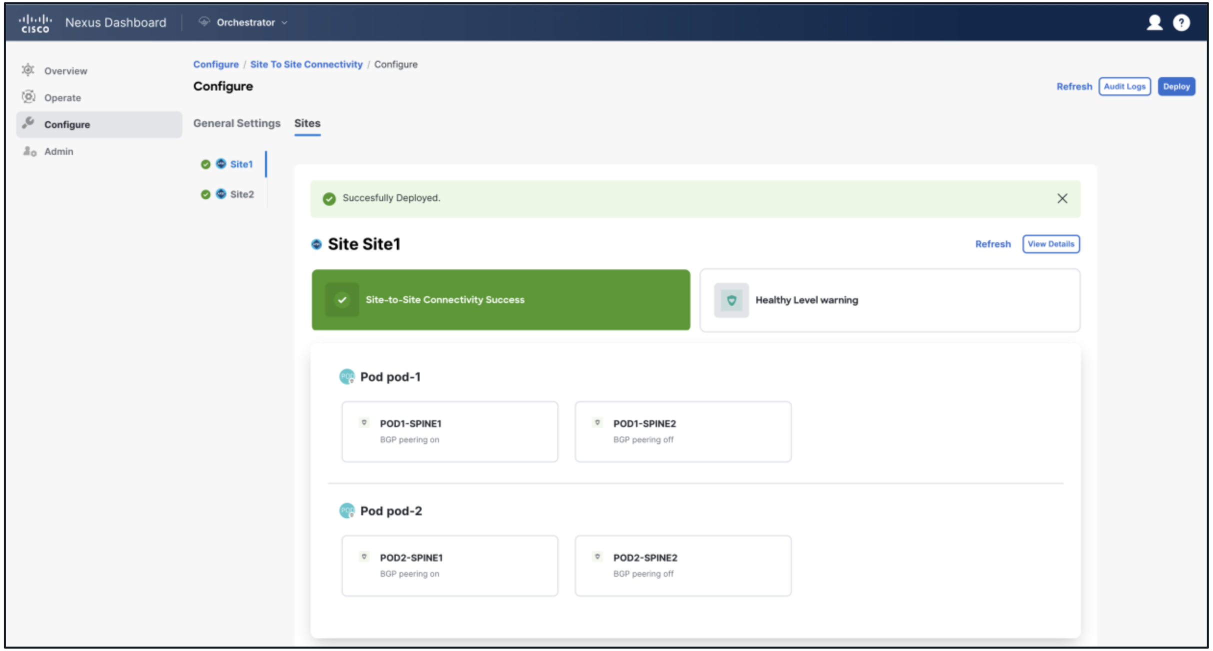
Need to apply a contract between internal EPG and Ext-EPG associated to the L3Out in Fabric 1

Mandates the use of a multi-VRF capable backbone network (VRF-Lite, MPLS-VPN, etc.) to extend multiple VRFs across fabrics

Need to apply a contract between Ext-EPG associated to the L3Out in Fabric 2 and internal EPG

Provisioning Policies on NDO

Provisioning Infra Configuration for the Fabrics



Succesfully Deployed.

Site Site1

Site-to-Site Connectivity Success

Healthy Level warning

Pod pod-1

POD1-SPINE1 BGP peering on

POD1-SPINE2 BGP peering off

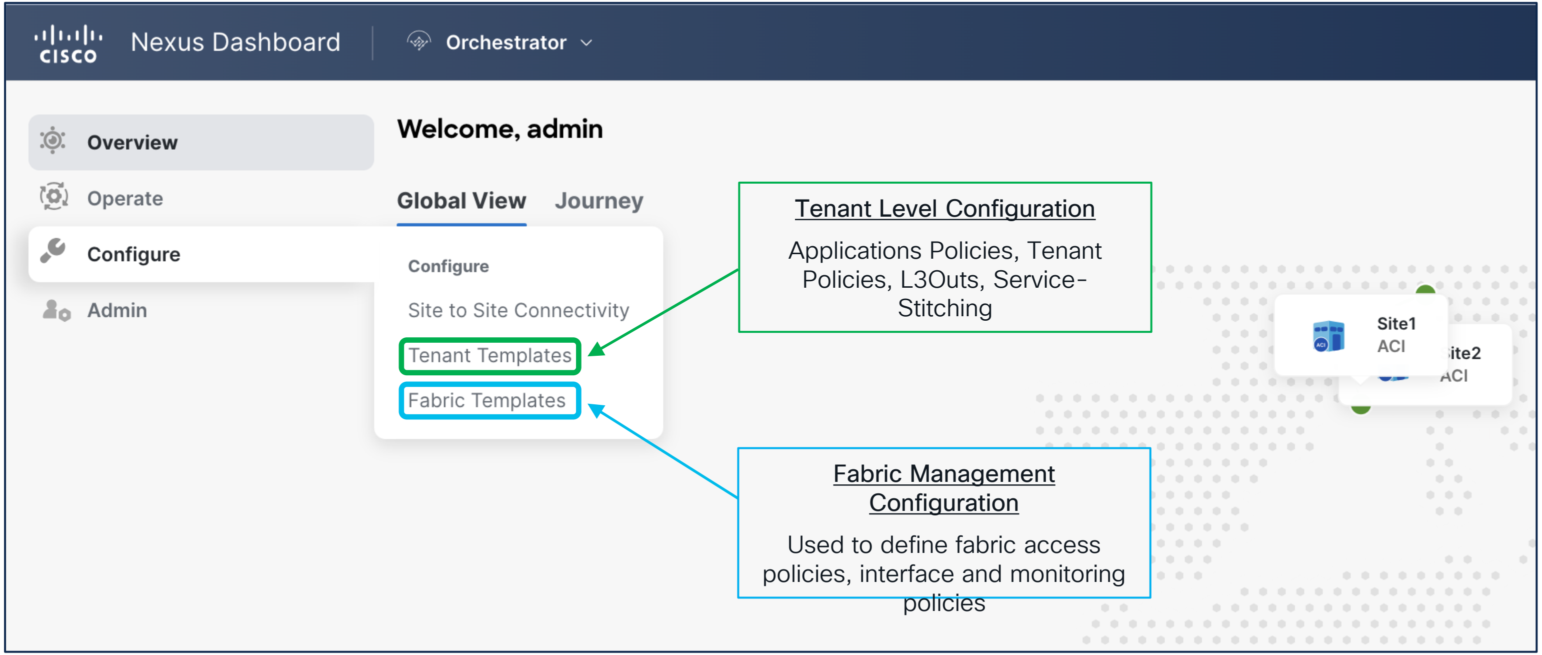
Pod pod-2

POD2-SPINE1 BGP peering on

POD2-SPINE2 BGP peering off

- 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



Tenant Level Configuration

Applications Policies, Tenant Policies, L3Outs, Service-Stitching

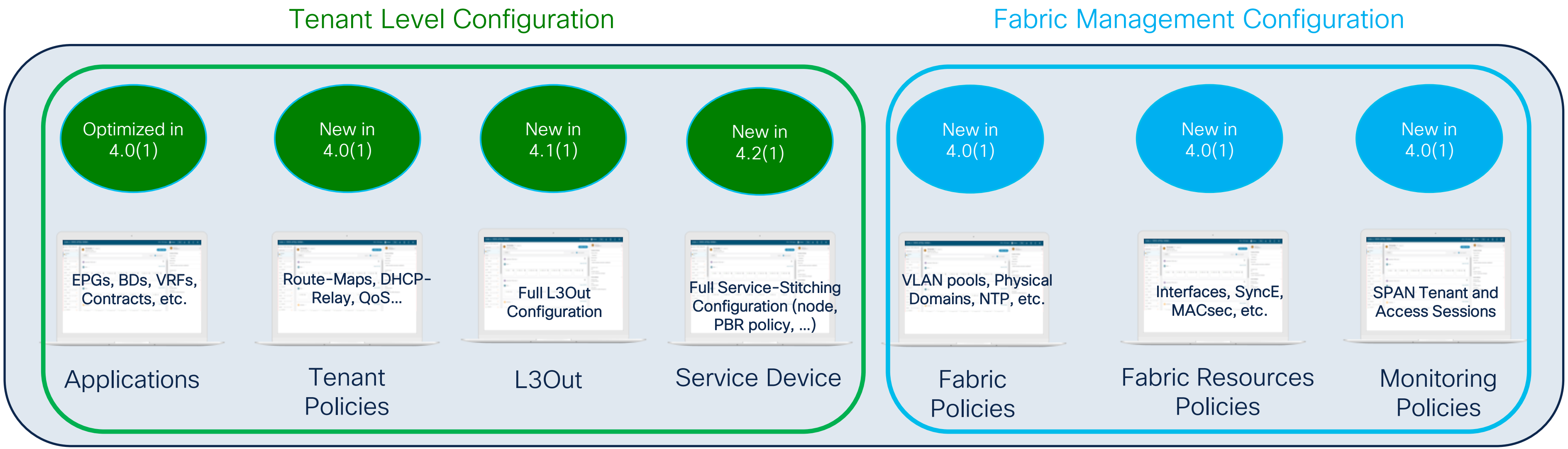
Fabric Management Configuration

Used to define fabric access policies, interface and monitoring policies

- 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

Optimized in 4.0(1)

EPGs, BDs, VRFs, Contracts, etc.

Applications

New in 4.0(1)

Route-Maps, DHCP-Relay, QoS…

Tenant Policies

New in 4.1(1)

Full L3Out Configuration

L3Out

New in 4.2(1)

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

Service Device

Fabric Management Configuration

New in 4.0(1)

VLAN pools, Physical Domains, NTP, etc.

Fabric Policies

New in 4.0(1)

Interfaces, SyncE, MACsec, etc.

Fabric Resources Policies

New in 4.0(1)

SPAN Tenant and Access Sessions

Monitoring Policies

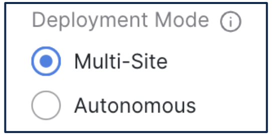
Benefits

Simplify	Single Pane of Glass

cisco Live!

Application Templates

Multi-Site Templates



Deployment Mode ⓘ

● Multi-Site

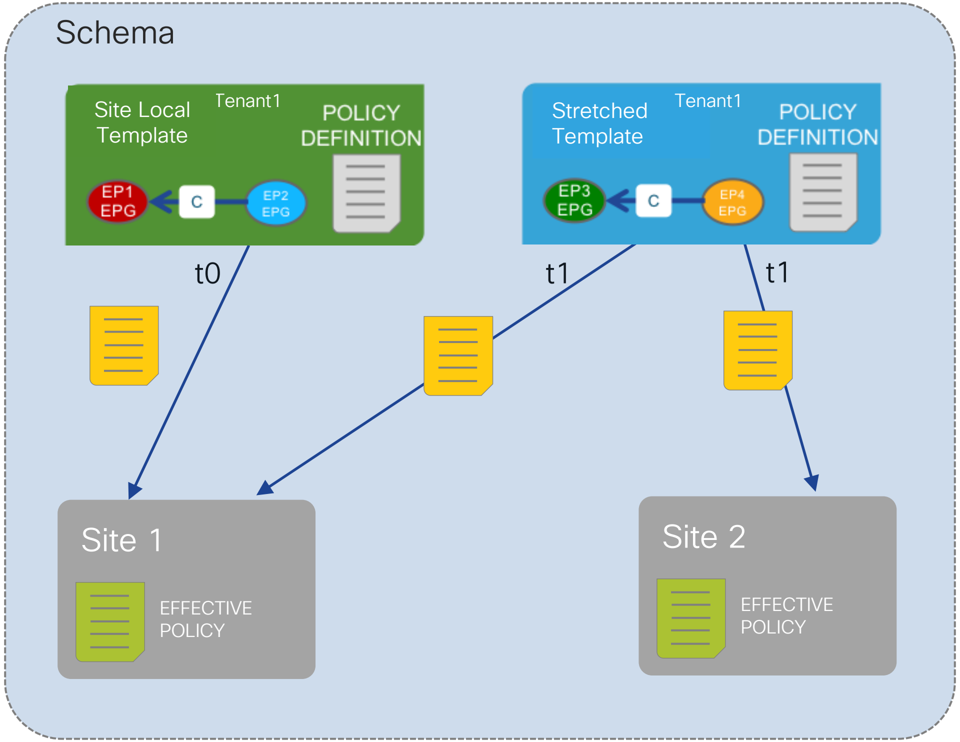
○ Autonomous

- 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

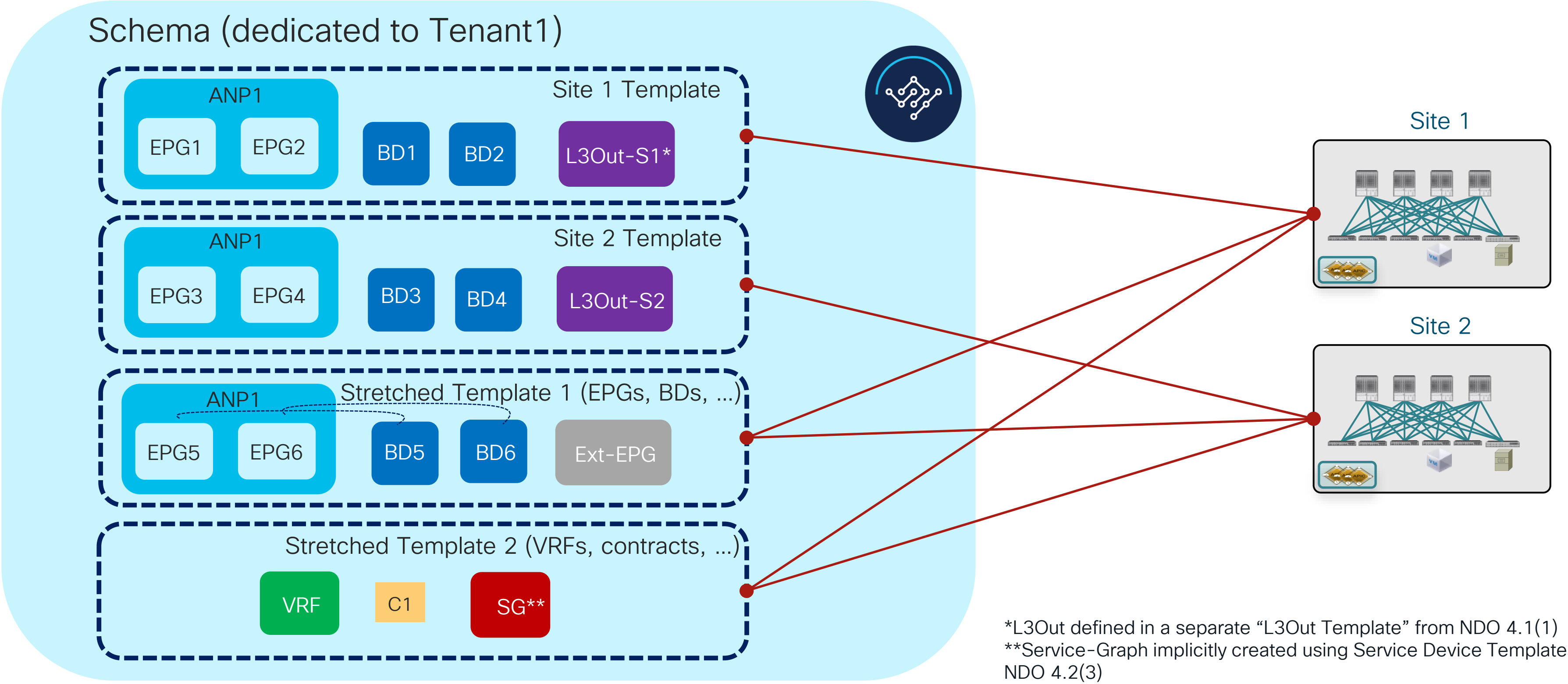
  A Multi-Site template associated to a single site can be pushed only to that site

  A Multi-Site template associated to multiple sites is concurrently pushed to all those sites



Best Practices for Multi-Site Templates

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

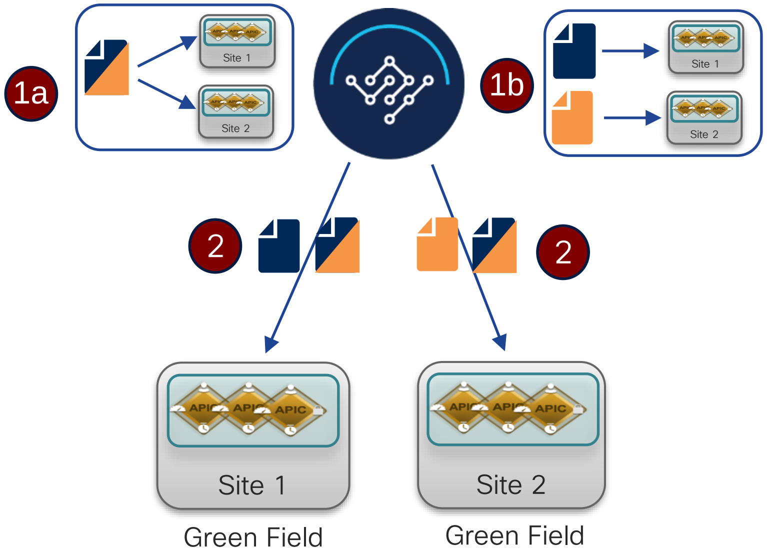


*L3Out defined in a separate “L3Out Template” from NDO 4.1(1)
**Service-Graph implicitly created using Service Device Templates from NDO 4.2(3)

Nexus Dashboard Orchestrator

Migration Scenarios

Green Field Deployment

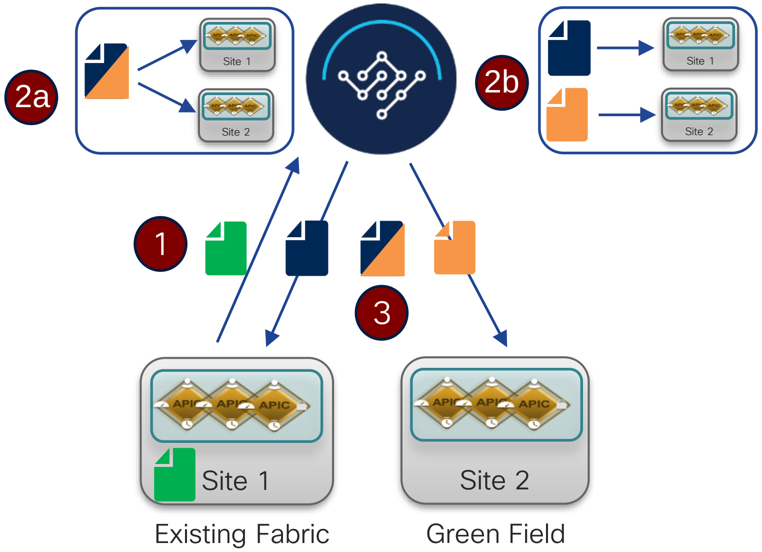


1a. Model new tenant and policies to a common template on NDO and associate the template to both sites (for stretched objects)

1b. Model new tenant and policies to site-specific templates and associate them to each site

2. Push policies to the ACI sites

Import Policies from an Existing Fabric



1. Import existing tenant policies from site 1 to new common and site-specific templates on NDO

2a. Associate the common template to both sites (for stretched objects)

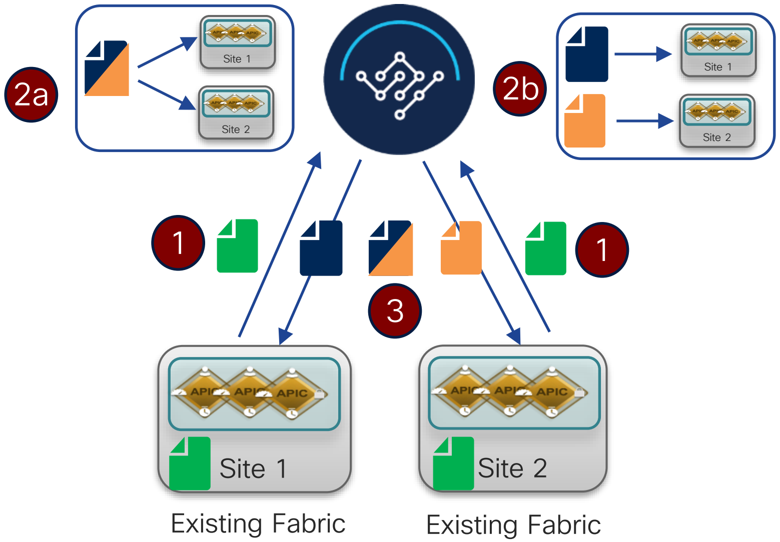
2b. Associate site-specific templates to each site

3. Push the policies back to the ACI sites

Nexus Dashboard Orchestrator

Migration Scenarios

Import Policies from Multiple Existing Fabrics



1. Import existing tenant policies from site 1 and site 2 to new common and site-specific templates on ACI MSO

2a. Associate the common template to both sites (for stretched objects)

2b. Associate site-specific templates to each site

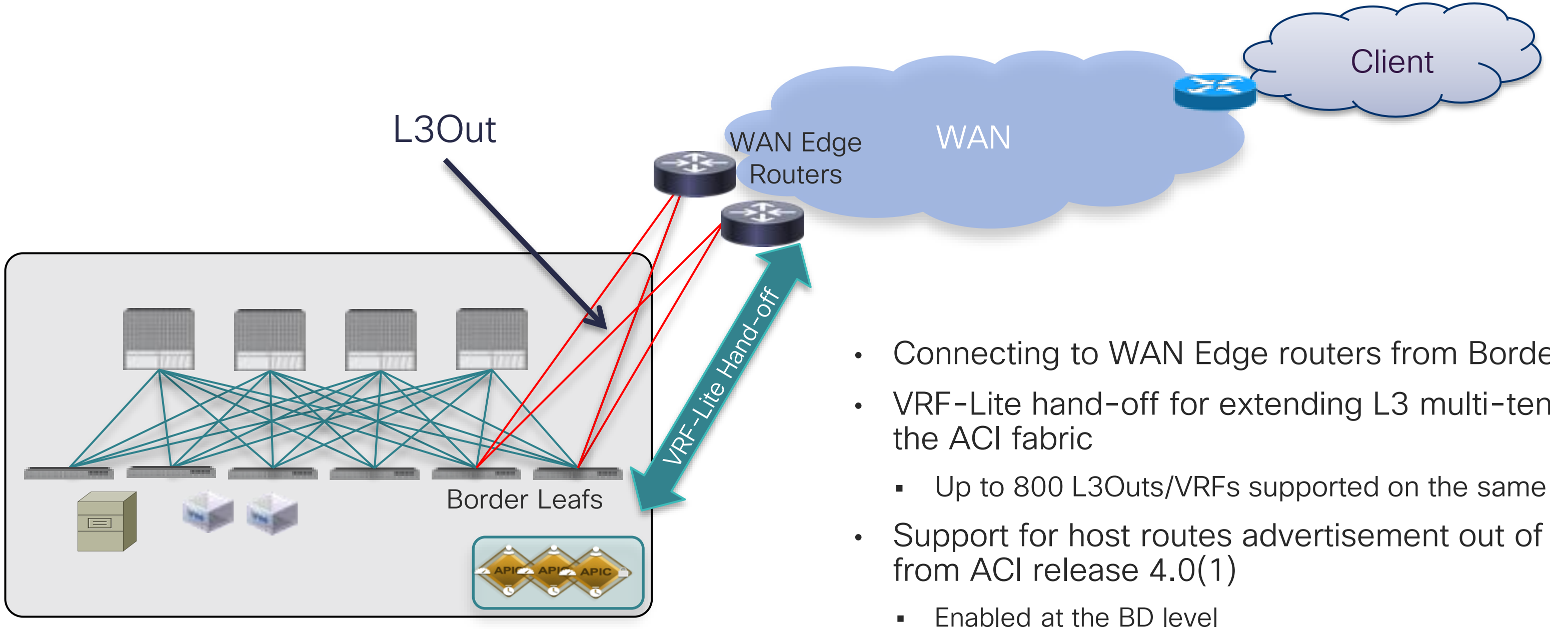
3. Push the policies back to the ACI sites

- NDO does not allow diff/merge operations on policies from different APIC domains
- It is still possible to import policies for the same tenant from different APIC domains, under the assumption those are no conflicting
  - Tenant defined with the same name
  - Name and policies for existing stretched objects are also common

Connecting to the External L3 Domain

Connecting to the External Layer 3 Domain

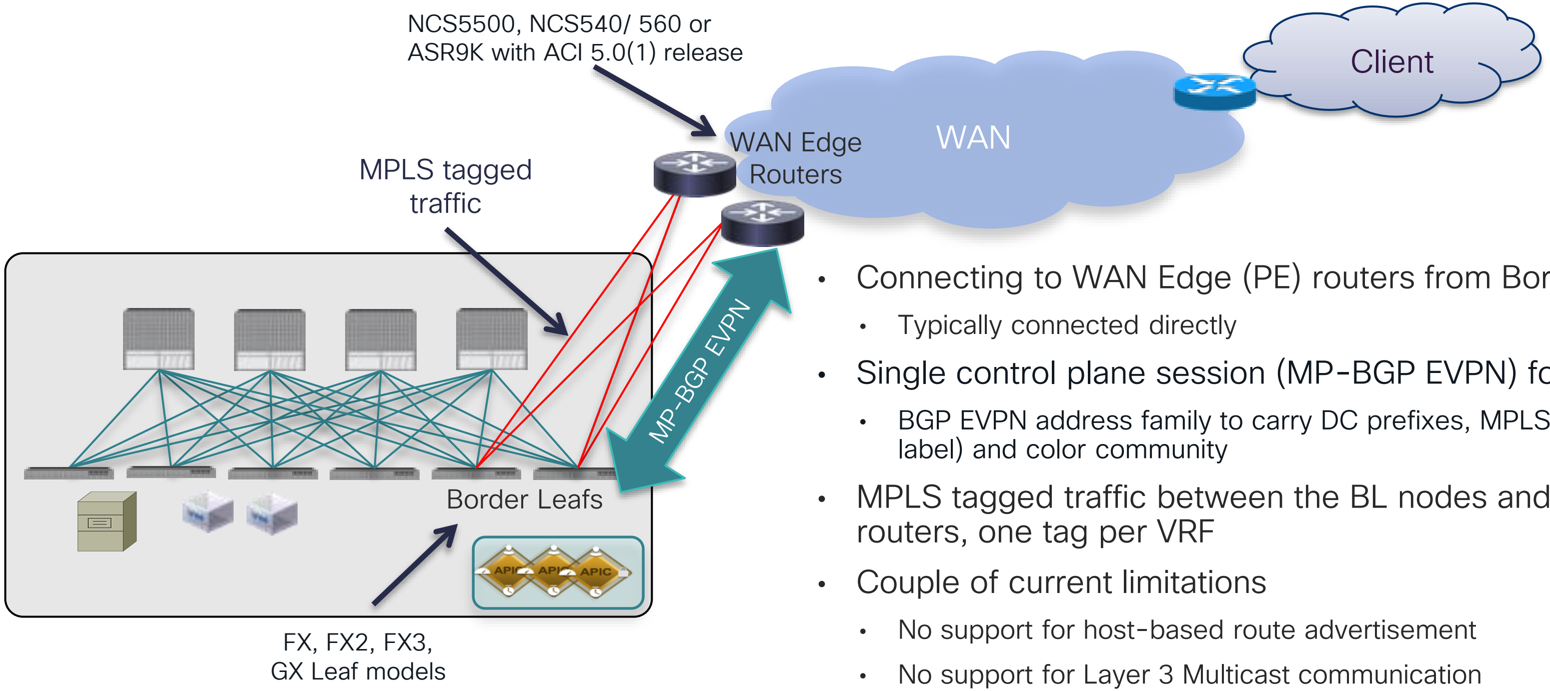
‘Traditional’ IP-Based L3Outs (Recommended Option)



- Connecting to WAN Edge routers from Border Leaf nodes
- VRF-Lite hand-off for extending L3 multi-tenancy outside the ACI fabric
    - Up to 800 L3Outs/VRFs supported on the same BL nodes pair
- 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

Connecting to the External Layer 3 Domain

SR-MPLS/MPLS Hand-Off on the BL Nodes



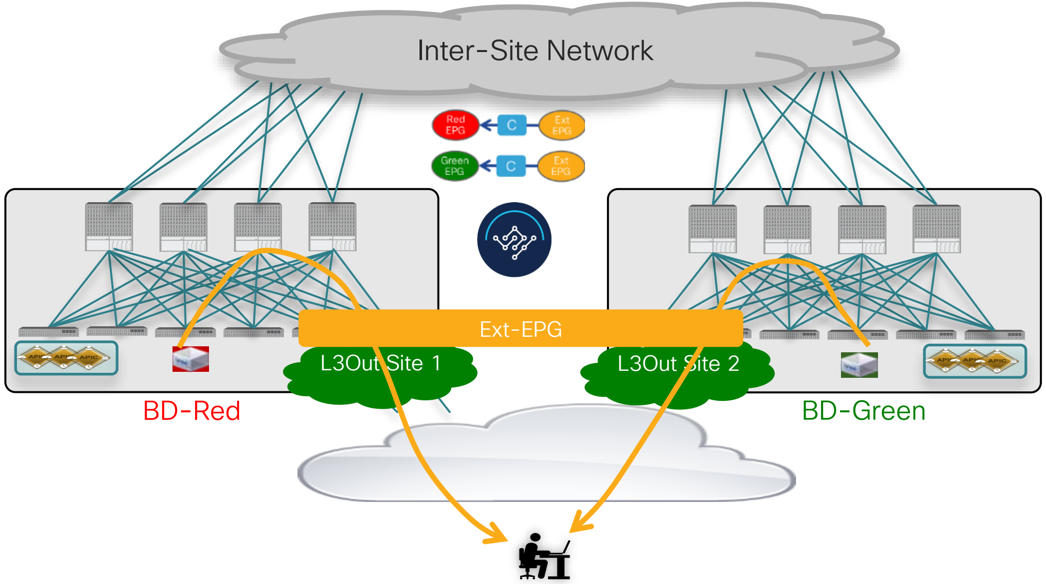
- Connecting to WAN Edge (PE) routers from Border Leaf nodes
  - Typically connected directly
- Single control plane session (MP-BGP EVPN) for all tenant VRFs
  - BGP EVPN address family to carry DC prefixes, MPLS label for VRF (VPN label) and color community
- MPLS tagged traffic between the BL nodes and the WAN Edge routers, one tag per VRF
- Couple of current limitations
  - No support for host-based route advertisement
  - No support for Layer 3 Multicast communication

https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-744107.html

Deploying External EPG(s) Associated to the L3Out

ACI Multi-Site and L3Out

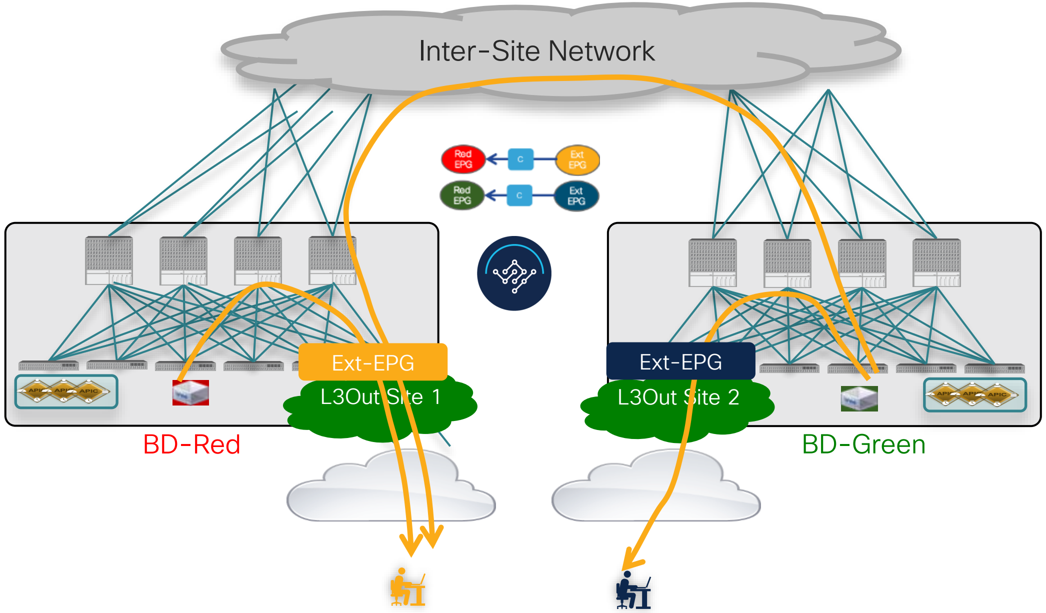
Stretching or Not Stretching the Ext-EPG?



- 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

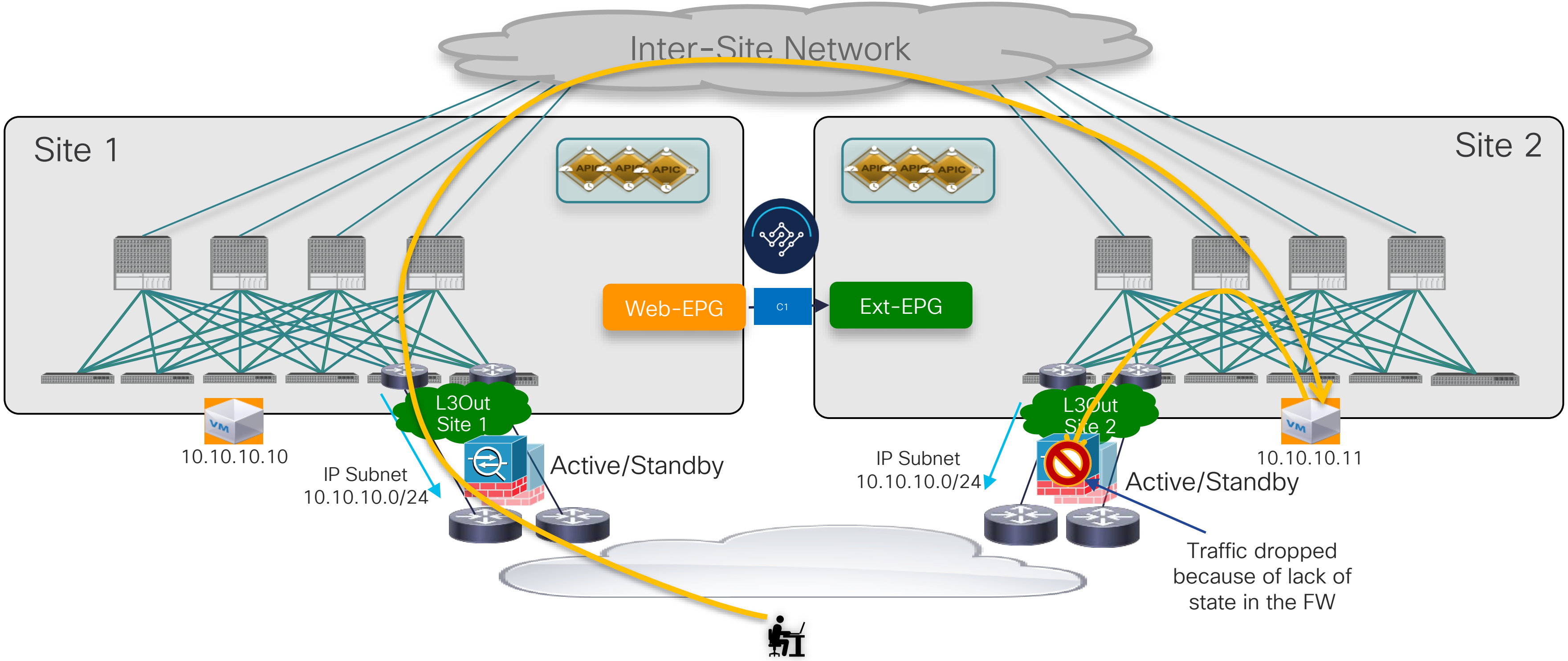
- Can still use the same 0.0.0.0/0 network configuration for classification on both sites

- May require enablement of Intersite L3Out

Solving Asymmetric Routing Issues with the External Network

ACI Multi-Site and L3Out

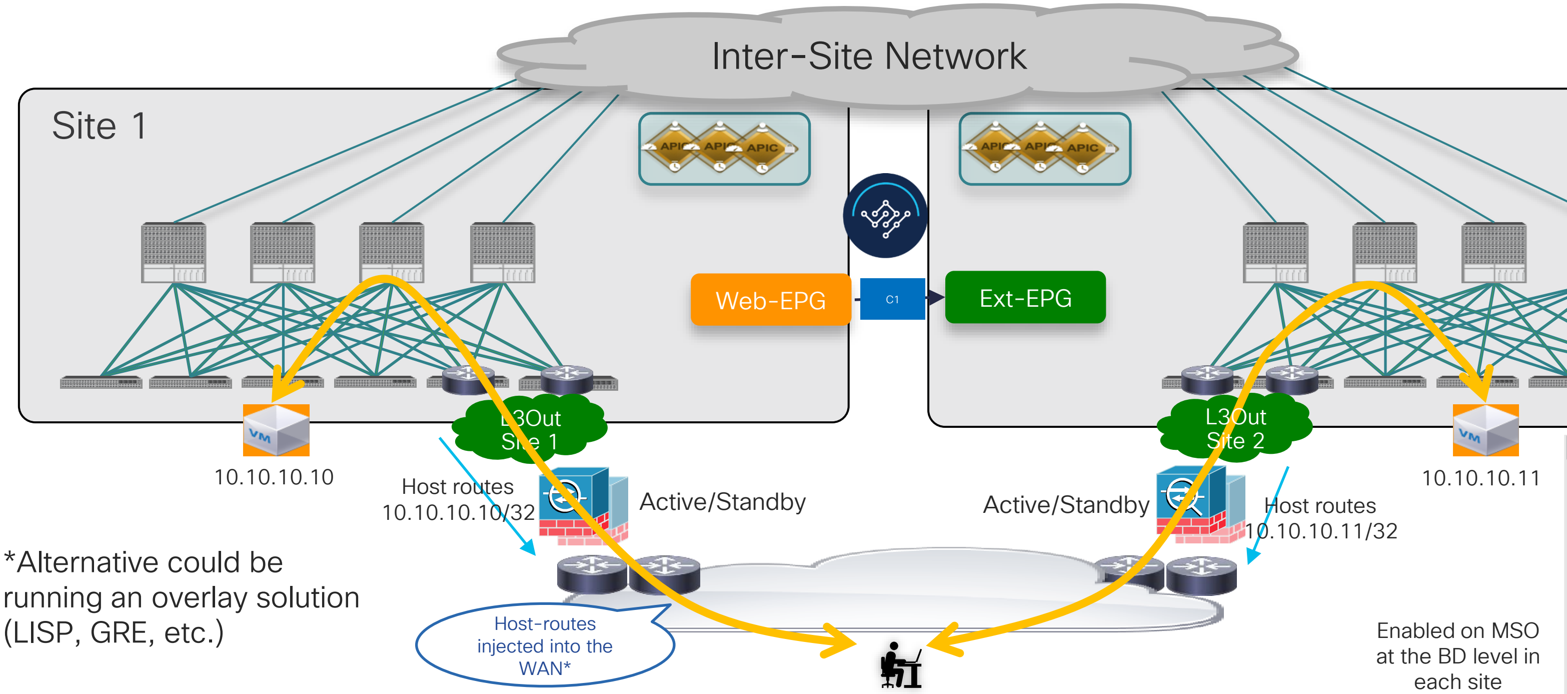
Typical Deployment of Perimeter FWs



Traffic dropped because of lack of state in the FW

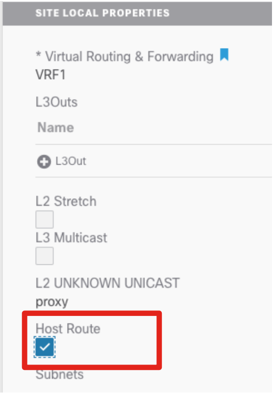
Solving Asymmetric Routing Issues

Use of Host-Routes Advertisement



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

Enabled on MSO at the BD level in each site


SITE LOCAL PROPERTIES

* Virtual Routing & Forwarding
VRF1

L3Outs
Name
L3Out

L2 Stretch
□

L3 Multicast
□

L2 UNKNOWN UNICAST
proxy

Host Route
☑

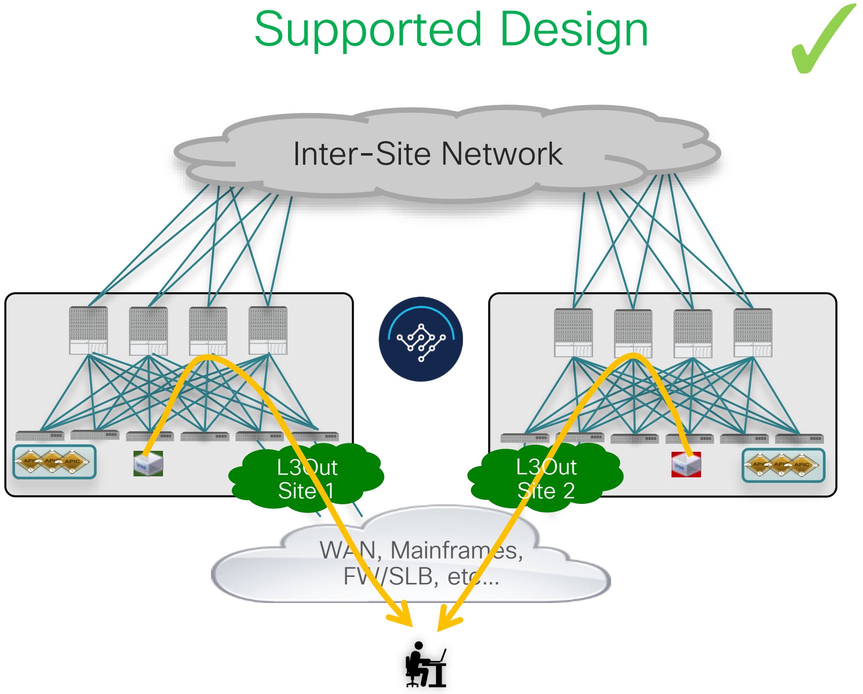
Subnets

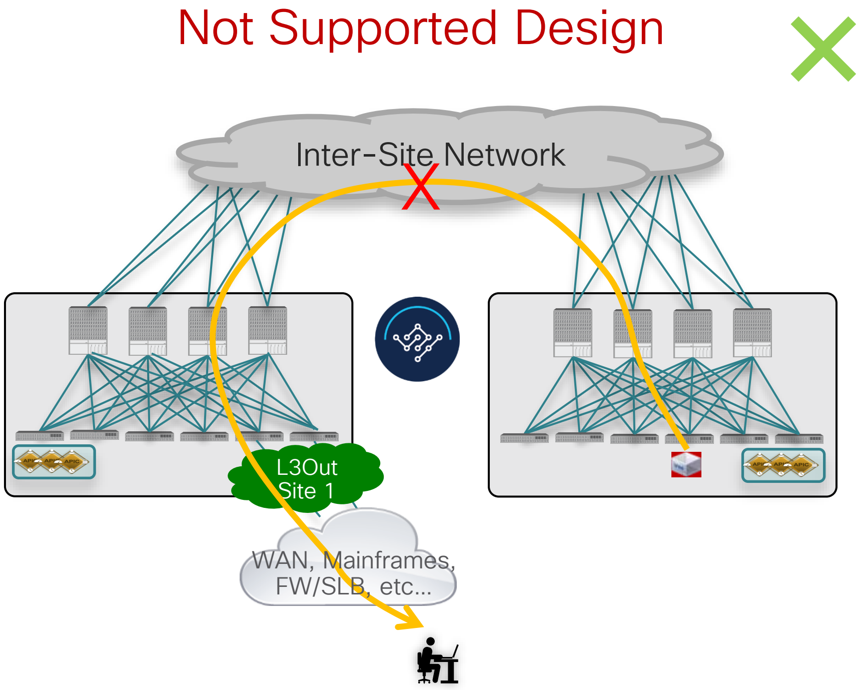
- Ingress optimization requires host-routes advertisement on the L3Out
  - Native support on ACI Border Leaf nodes available from ACI release 4.0(1)
  - Not currently supported for SR-MPLS L3Outs

Intersite L3Out Support

Problem Statement

Behavior before ACI Release 4.2(1)

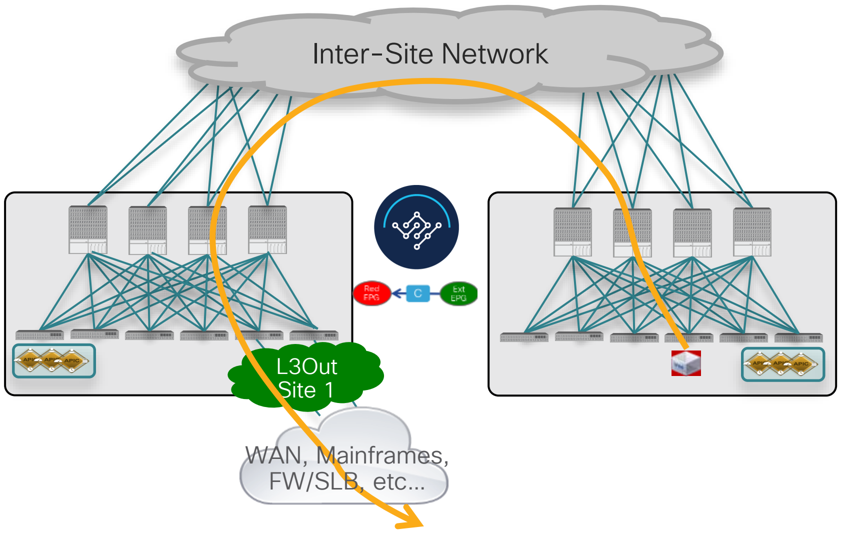


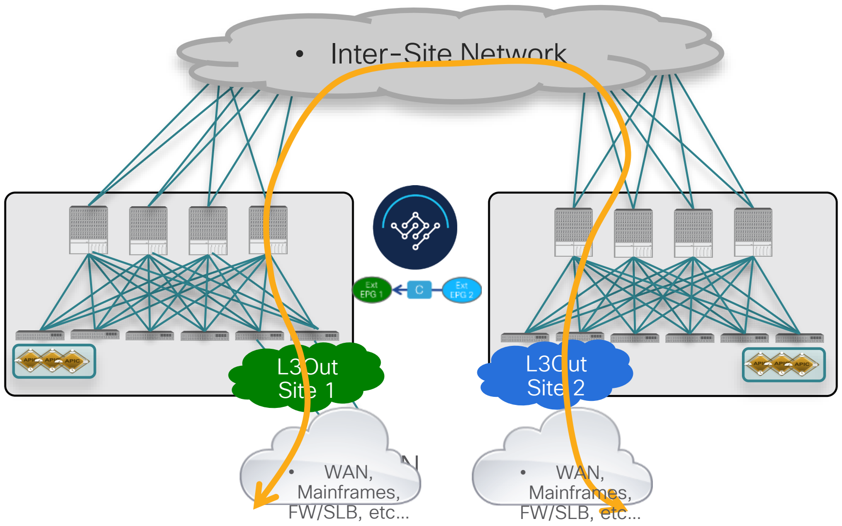


Note: the same consideration applies to both IP-Based L3Outs and SR-MPLS L3Outs

ACI Multi-Site and Intersite L3Out

Supported Scenarios





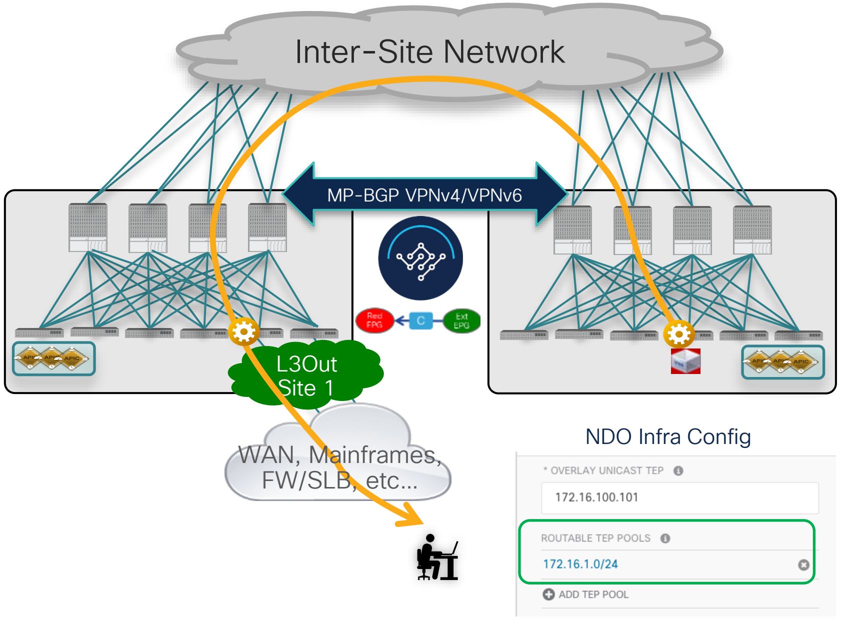
Inter-Site Network
Ext EPG 1
C
Ext EPG 2
L3Out Site 1
L3Out Site 2
WAN, Mainframes, FW/SLB, etc...
WAN, Mainframes, FW/SLB, etc...

- 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



NDO Infra Config

* OVERLAY UNICAST TEP

172.16.100.101

ROUTABLE TEP POOLS

172.16.1.0/24

ADD TEP POOL

- 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)

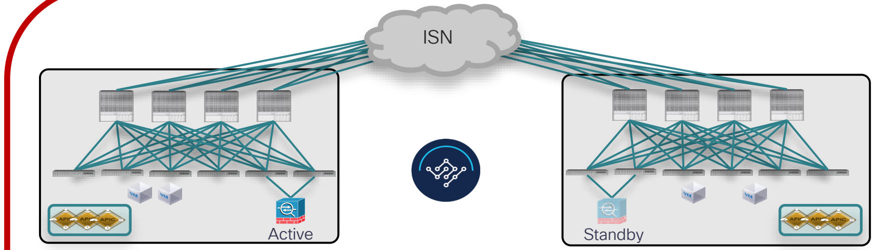
Network Services Integration

Integration Models

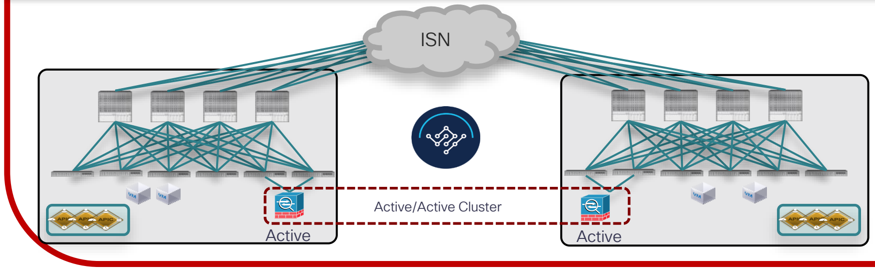
ACI Multi-Site and Network Services

Integration Models

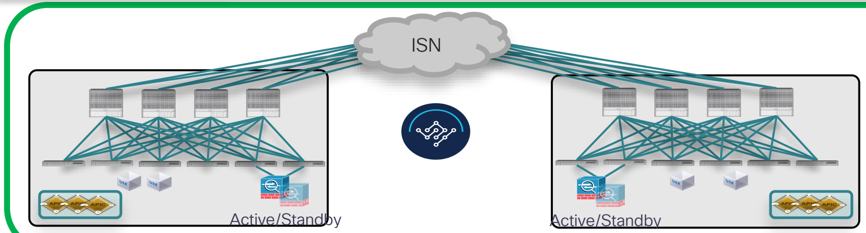
Deployment options fully supported with ACI Multi-Pod



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



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



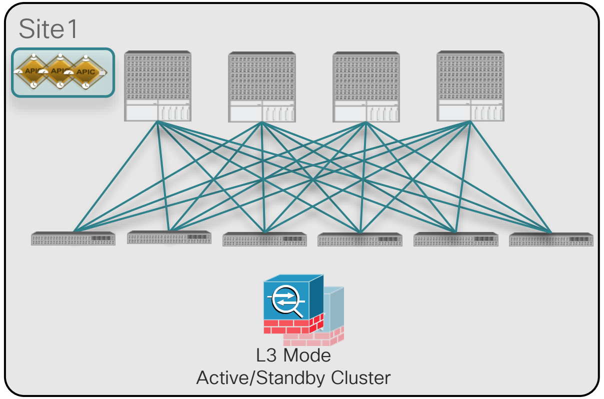
- Typical deployment model for ACI Multi-Site, each fabric leverages a dedicated service node function
- Use of PBR to avoid creating asymmetric paths through stateful devices (FWs, LBs, etc.) for both North-South and East-West communication

Use of Service Graph and PBR

Resilient Service Node Deployment in Each Site

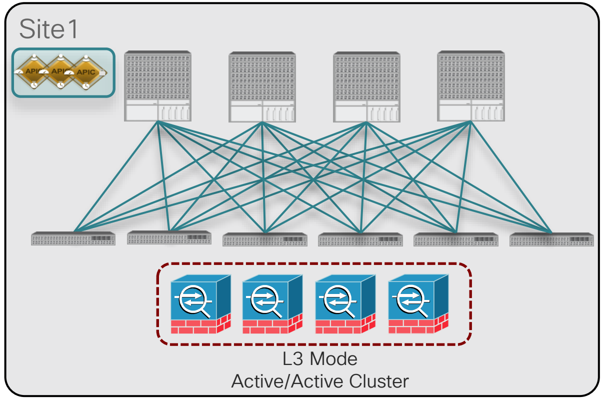
PBR redirection only supported to a local service function, hence it is important to deploy such function in a resilient way

Active/Standby Cluster



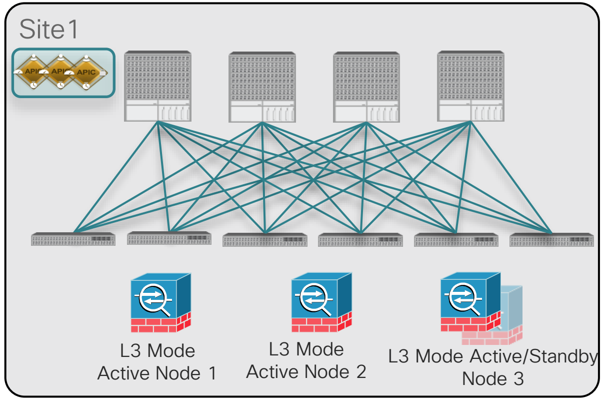
- The Active/Standby pair represents a single MAC/IP entry in the PBR policy

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

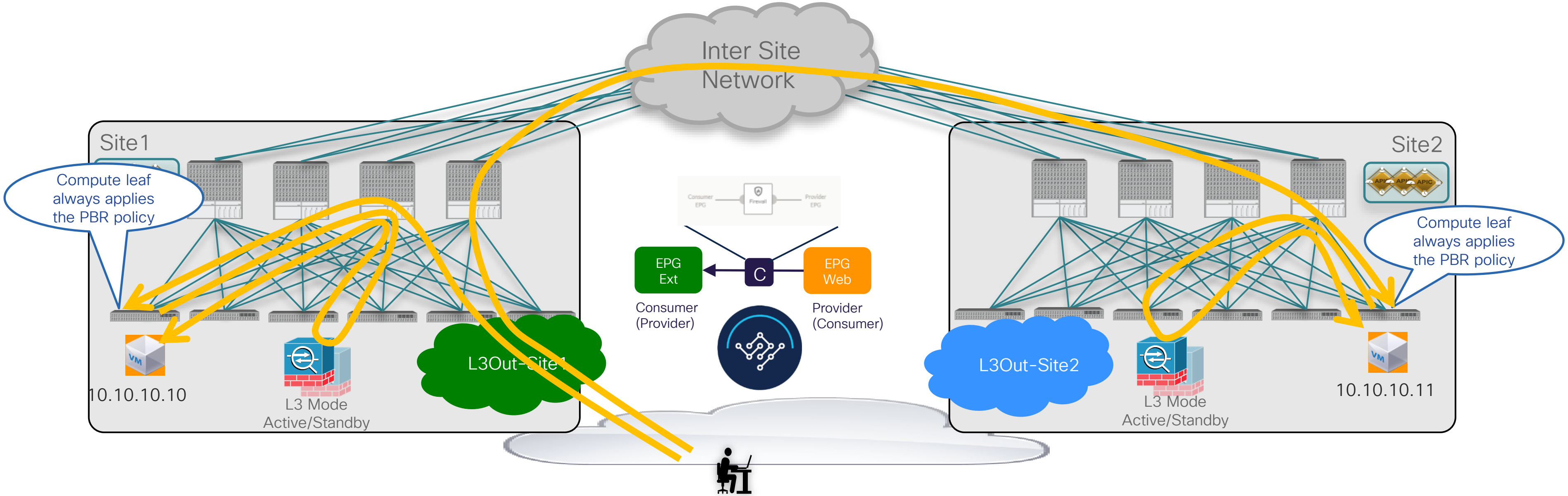


- Each Active node represent a unique MAC/IP entry in the PBR policy
- Use of Symmetric PBR to ensure each flow is handled by the same Active node in both directions

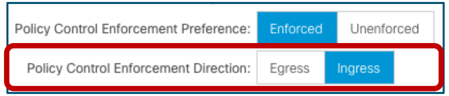
Use of Service Graph and PBR North-South and East-West

North-South Communication

Inbound Traffic



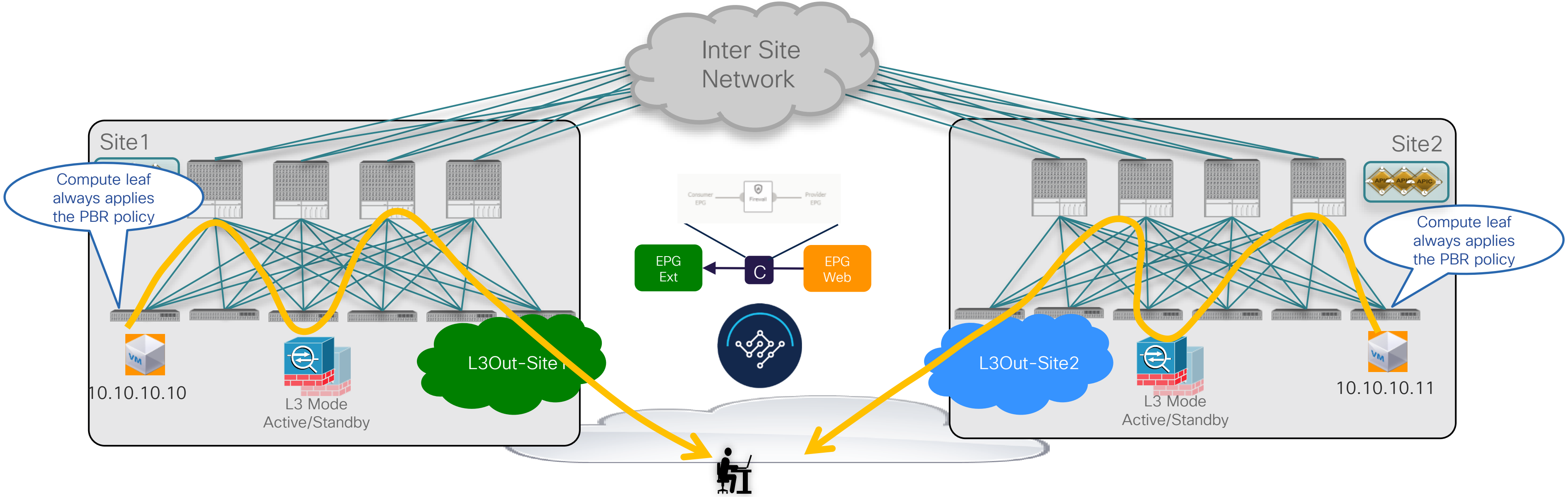
- Inbound traffic can enter any site when destined to a stretched subnet (if ingress optimization is not deployed or possible)
- PBR policy is always applied on the compute leaf node where the destination endpoint is connected
  - Requires the VRF to have the default policies for enforcement preference 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

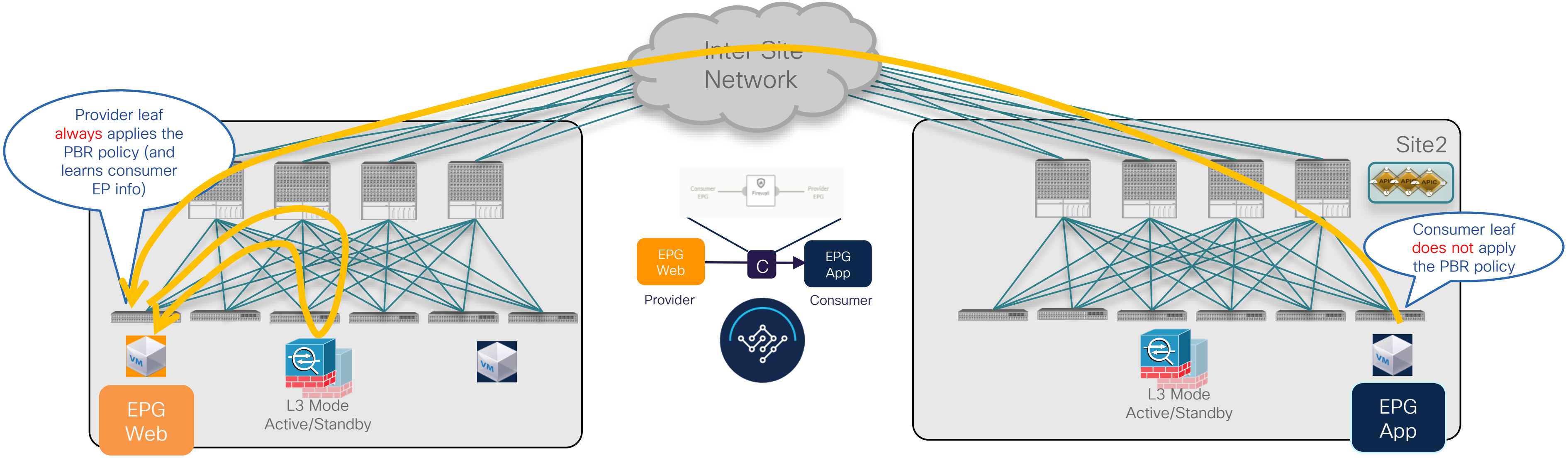
Outbound Traffic



- 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

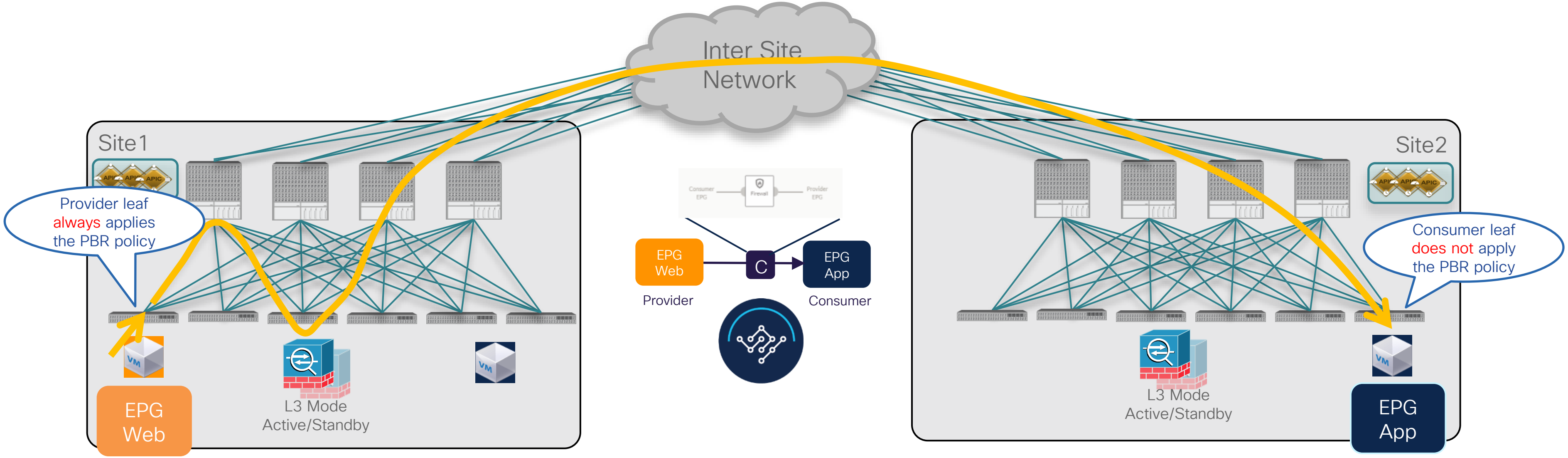


EP-App	O-UTEP S2

- 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

East-West Communication

Provider to Consumer Return Flow

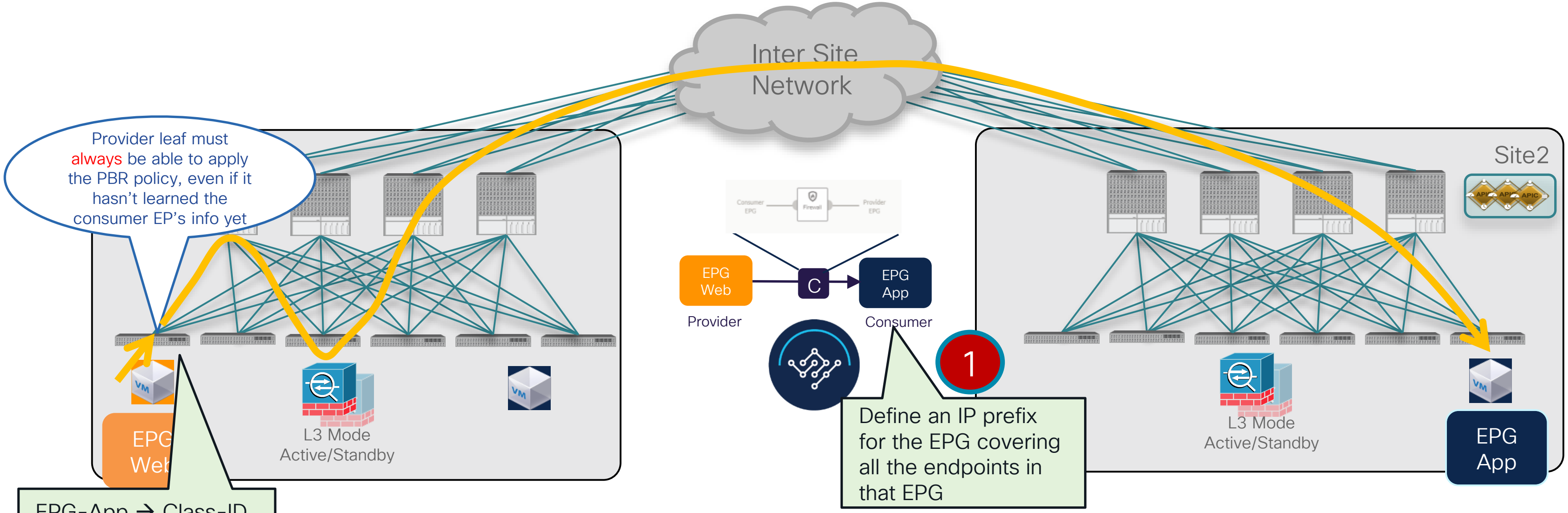


EP-App	O-UTEP S2

- 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 <u>always</u> 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?



Provider leaf must always be able to apply the PBR policy, even if it hasn't learned the consumer EP's info yet

1 Define an IP prefix for the EPG covering all the endpoints in that EPG

EPG-App → Class-ID information statically configured on the provider leaf node

2

- 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

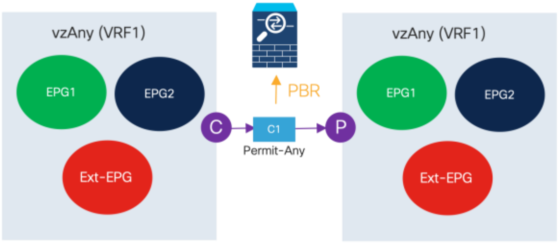
New PBR Supported Use Cases

ACI Multi-Site and PBR Enhancements

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

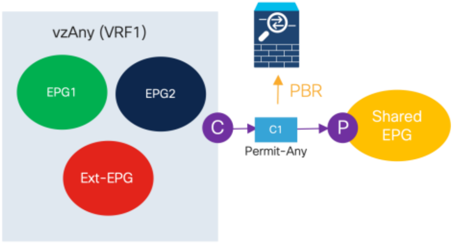
New Supported Use Cases

Any-to-Any



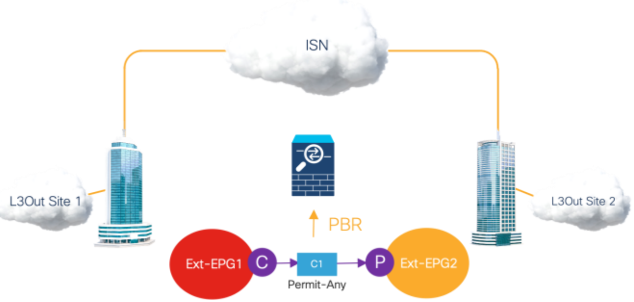
- Support only for single service node iertion (one-arm)
- Distributed deployment model (traffic is redirected via both local and remote service node)
- Intra-VRF only
- Works for both "network centric" and "app centric" designsns

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



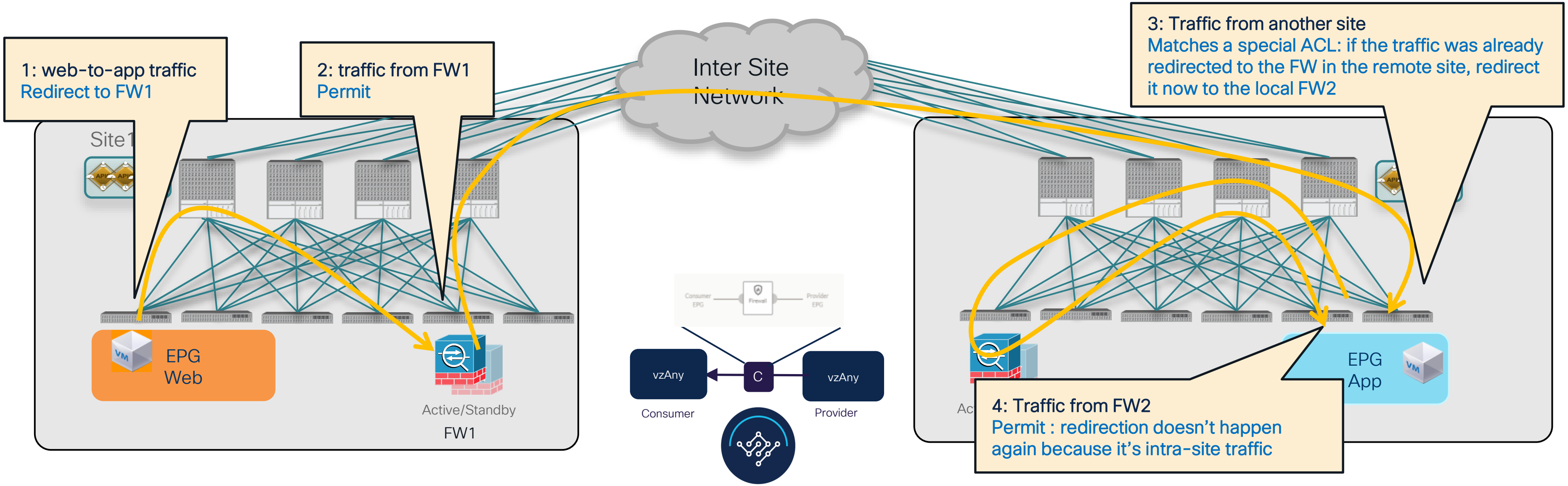
- Support only for single service node insertion (one-arm)
- Redirect intersite transit routing traffic flows
- Traffic is redirected via both local and remote service node
- Intra-VRF and inter-VRF

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

How to Keep Traffic Symmetric

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

- Redirect "inter-site" traffic in both ingress and egress sites



1: web-to-app traffic
Redirect to FW1

2: traffic from FW1
Permit

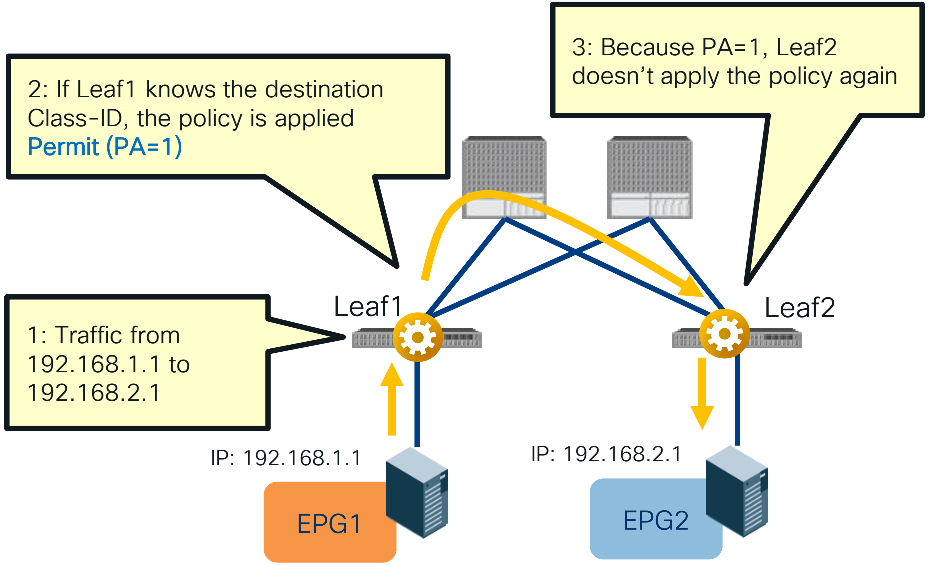
3: Traffic from another site
Matches a special ACL: if the traffic was already redirected to the FW in the remote site, redirect it now to the local FW2

4: Traffic from FW2
Permit : redirection doesn't happen again because it's intra-site traffic

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



1: Traffic from 192.168.1.1 to 192.168.2.1

2: If Leaf1 knows the destination Class-ID, the policy is applied
Permit (PA=1)

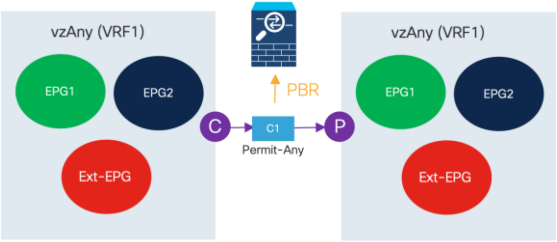
3: Because PA=1, Leaf2 doesn't apply the policy again

	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

NEW

"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



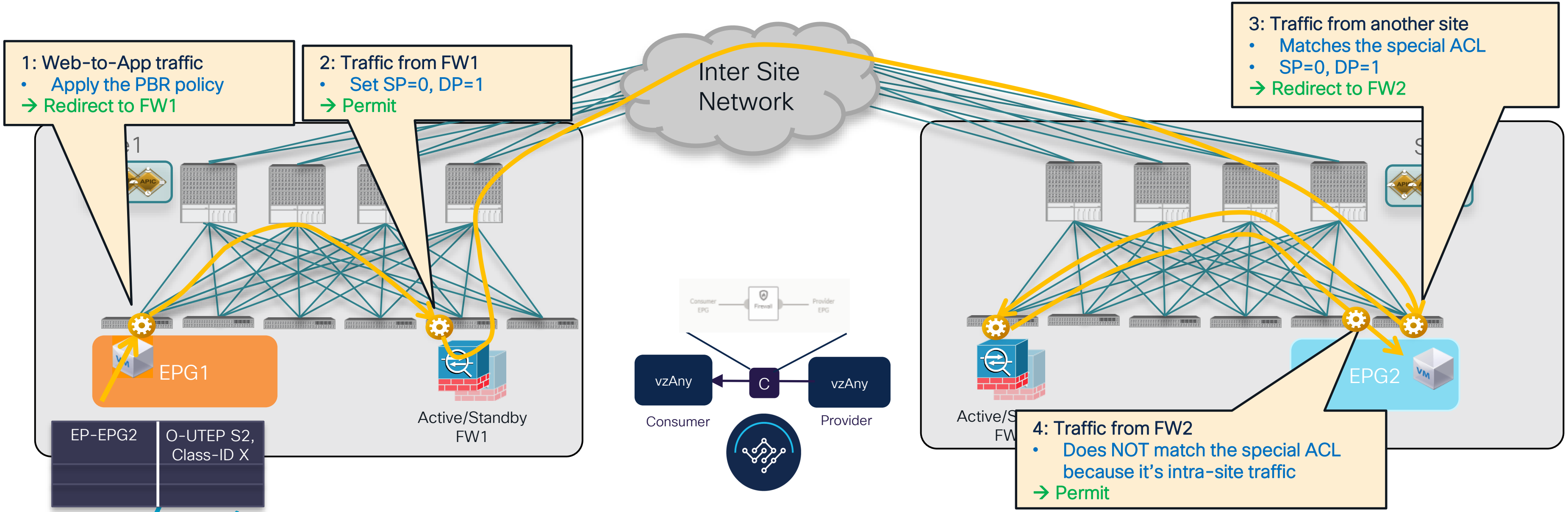
NDO 4.2(3)/ACI 6.0(4)

vzAny-to-vzAny PBR Use Case

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



1: Web-to-App traffic
- Apply the PBR policy

→ Redirect to FW1

2: Traffic from FW1
- Set SP=0, DP=1

→ Permit

3: Traffic from another site
- Matches the special ACL
- SP=0, DP=1

→ Redirect to FW2

4: Traffic from FW2
- Does NOT match the special ACL because it's intra-site traffic

→ Permit

EP-EPG2	O-UTEP S2, Class-ID X

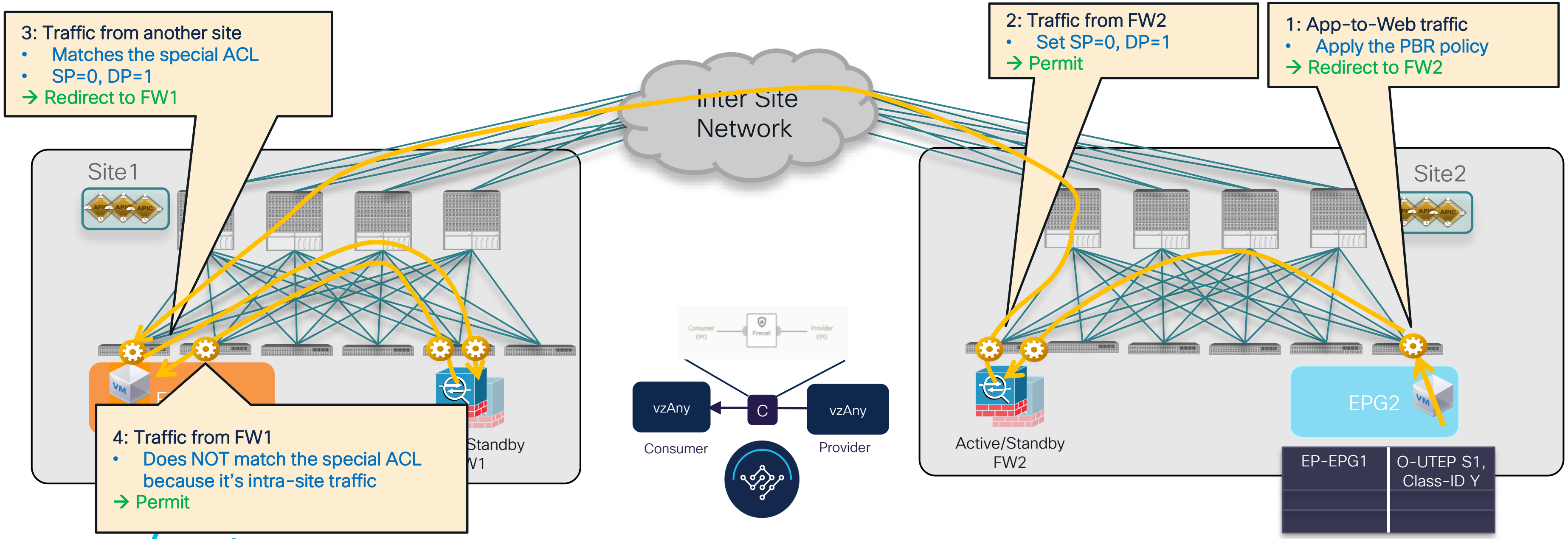
vzAny-to-vzAny PBR Use Case

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 provider endpoint **is known** on the consumer leaf



1: App-to-Web traffic
- Apply the PBR policy

→ Redirect to FW2

2: Traffic from FW2
- Set SP=0, DP=1

→ Permit

3: Traffic from another site
- Matches the special ACL
- SP=0, DP=1

→ Redirect to FW1

4: Traffic from FW1
- Does NOT match the special ACL because it's intra-site traffic

→ Permit

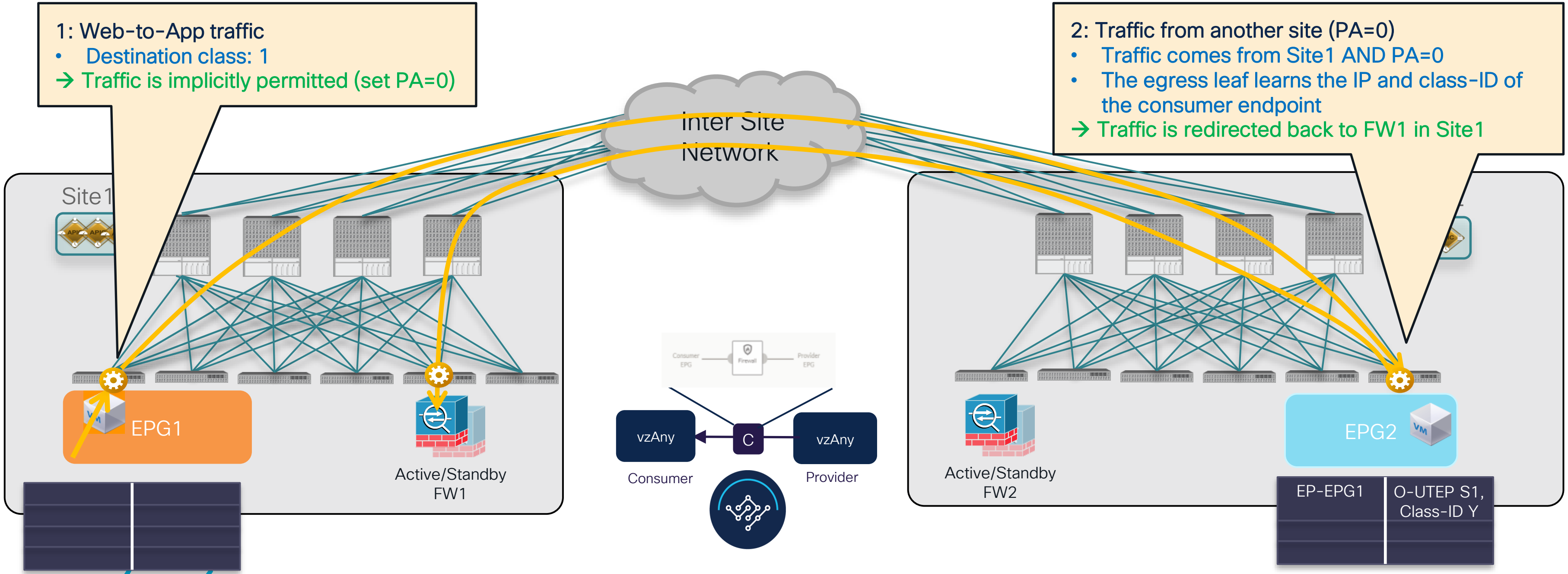
EP-EPG1	O-UTEP S1, Class-ID Y

NDO 4.2(3)/ACI 6.0(4)

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



1: Web-to-App traffic
- Destination class: 1

→ Traffic is implicitly permitted (set PA=0)

2: Traffic from another site (PA=0)
- Traffic comes from Site1 AND PA=0
- The egress leaf learns the IP and class-ID of the consumer endpoint

→ Traffic is redirected back to FW1 in Site1

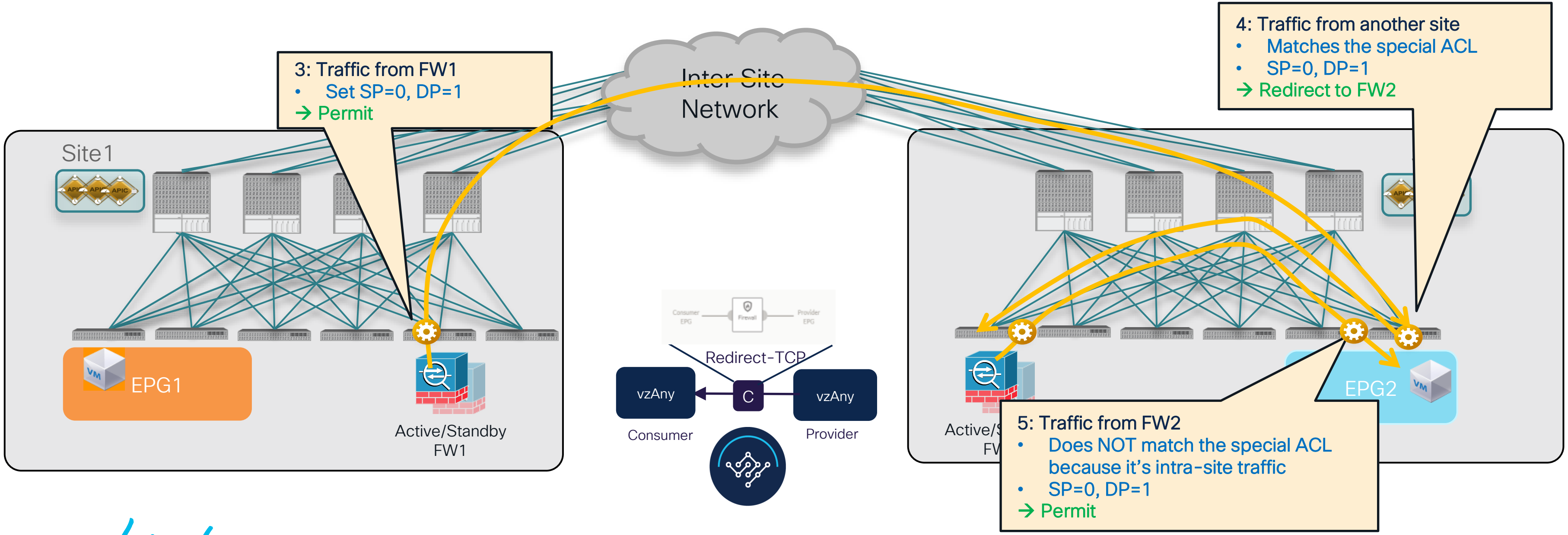
EP-EPG1	O-UTEP S1, Class-ID Y

NDO 4.2(3)/ACI 6.0(4)

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



3: Traffic from FW1
- Set SP=0, DP=1

→ Permit

4: Traffic from another site
- Matches the special ACL
- SP=0, DP=1

→ Redirect to FW2

5: Traffic from FW2
- Does NOT match the special ACL because it's intra-site traffic
- SP=0, DP=1

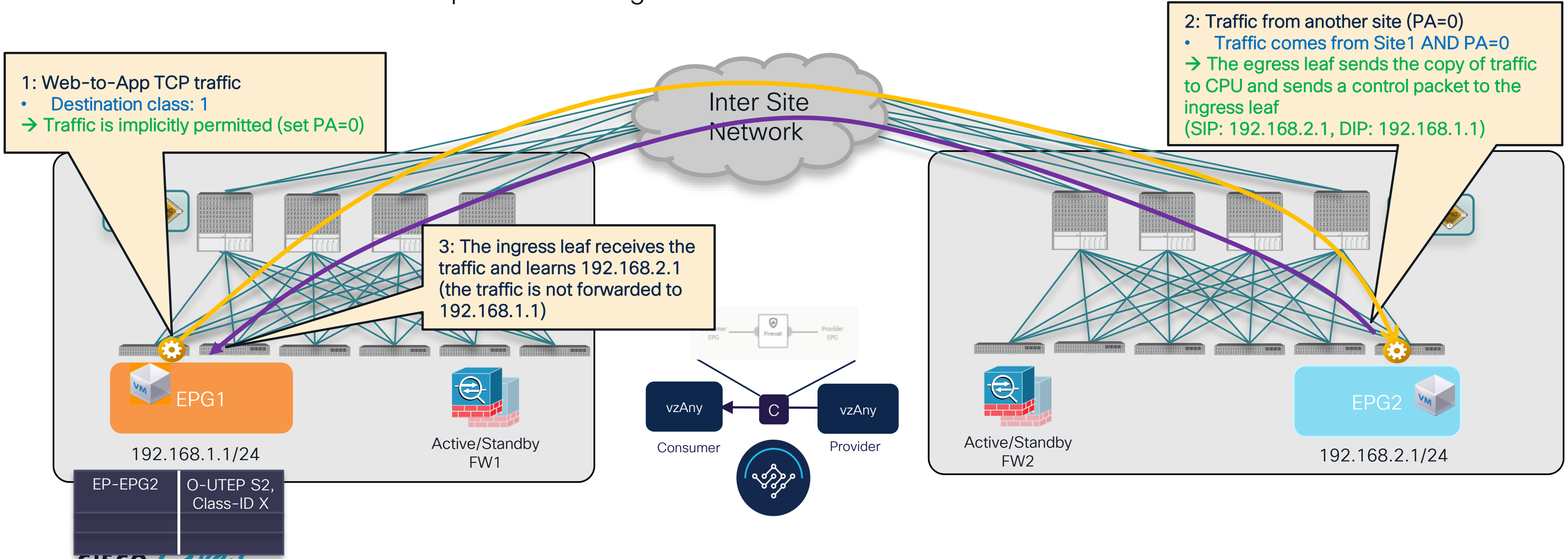
→ Permit

NDO 4.2(3)/ACI 6.0(4)

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



1: Web-to-App TCP traffic
- Destination class: 1

→ Traffic is implicitly permitted (set PA=0)

2: Traffic from another site (PA=0)
- Traffic comes from Site1 AND PA=0

→ The egress leaf sends the copy of traffic to CPU and sends a control packet to the ingress leaf
(SIP: 192.168.2.1, DIP: 192.168.1.1)

3: The ingress leaf receives the traffic and learns 192.168.2.1 (the traffic is not forwarded to 192.168.1.1)

EP-EPG2	O-UTEP S2, Class-ID X

ACI Multi-Site

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

Webex App

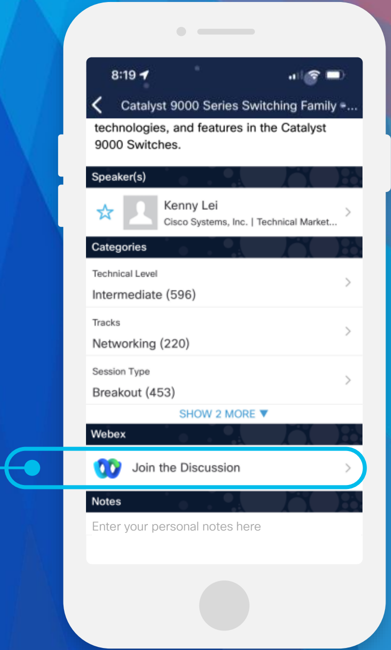
Questions?

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

How

1. Find this session in the Cisco Events 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 February 23, 2024.



8:19

Catalyst 9000 Series Switching Family ...

technologies, and features in the Catalyst 9000 Switches.

Speaker(s)

☆ Kenny Lei
Cisco Systems, Inc. | Technical Market...

Categories

Technical Level
Intermediate (596)

Tracks
Networking (220)

Session Type
Breakout (453)

SHOW 2 MORE ▼

Webex

Join the Discussion

Notes

Enter your personal notes here

https://ciscolive.ciscoevents.com/ciscolivebot/# BRKDCN-2980

Fill out your session surveys!

Participants who fill out a minimum of **four session surveys and the overall event survey** will get a Cisco Live t-shirt (from 11:30 on Thursday, while supplies last)!

All surveys can be taken in the Cisco Events Mobile App or by logging into the Session Catalog and clicking the ‘Participant Resource Center’ link at https://www.ciscolive.com/emea/learn/session-catalog.html.

Continue your education

- Visit the Cisco Showcase for related demos
- Book your one-on-one Meet the Engineer meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at ciscolive.com/on-demand. Sessions from this event will be available from February 23.

The bridge to possible

Thank you

cisco Live!

Let’s go