Cisco ACI Multi-Pod Design and Deployment

John Weston – Technical Marketing Engineer
Questions?
Use Cisco Spark to communicate with the speaker after the session

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

cs.co/ciscolivebot#BRKACI-2003
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 vs. Multi-Site)
- Understand the functionalities and specific design considerations associated to the ACI Multi-Pod Fabric option

Initial assumption:

- The audience already has a good knowledge of ACI main concepts (Tenant, BD, EPG, L2Out, L3Out, etc.)
Agenda

- ACI Network and Policy Domain Evolution
- ACI Multi-Pod Deep Dive
  - Overview, Use Cases and Supported Topologies
  - APIC Cluster Deployment Considerations
  - Inter-Pod Connectivity Deployment Considerations
  - Control and Data Planes
  - Connecting to the External Layer 3 Domain
  - Network Services Integration
  - Migration Scenarios
ACI Network and Policy Domain Evolution
Cisco ACI
Fabric and Policy Domain Evolution

ACI Single Pod Fabric

ACI Stretched Fabric

ACI Multi-Pod Fabric

ACI 1.0 - Leaf/Spine Single Pod Fabric
ACI 1.1 - Geographically Stretch a single Pod
ISE 2.1 & ACI 1.2 Federation of Identity and Interconnect TrustSec and ACI using IP based EPG/SGT
ACI 2.0 - Multiple Networks (Pods) in a single Availability Zone (Fabric)
ACI 3.0 – Multiple Availability Zones (Fabrics) in a Single Region ‘and’ Multi-Region Policy Management
ACI 3.1/3.2 - Remote Leaf and vPod extends an Availability Zone (Fabric) to remote locations

ACI Multi-Site
Fabric and Policy Domain Evolution
Deployment Options

**Single APIC Cluster/Single Fabric**
- Stretched Fabric
- DC1
- DC2
- ACI Fabric
- APIC Cluster

**Multiple APIC Clusters/Multiple Fabrics**
- Multi-Fabric (with L2 and L3 DCI)
- Fabric ‘A’
- Fabric ‘n’
- Inter-Site App
- L2/L3 DCI

**Deployment Options**
- Multi-Pod (from 2.0 Release)
- Pod ‘A’
- Pod ‘n’
- MP-BGP - EVPN
- IPN

- Multi-Site (3.0 Release, Q3CY17)
- Fabric ‘A’
- Fabric ‘n’
- MP-BGP - EVPN
- IP
- ACI Multi-Site
Terminology

- **Pod** – A Leaf/Spine network sharing a common control plane (ISIS, BGP, COOP, …)
  Pod == Network Fault Domain

- **Fabric** – Scope of an APIC Cluster, it can be one or more Pods
  Fabric == Availability Zone (AZ) or Tenant Change Domain

- **Multi-Pod** – Single APIC Cluster with multiple leaf spine networks
  Multi-Pod == Multiple Networks within a Single Availability Zone (Fabric)

- **Multi-Fabric** – Multiple APIC Clusters + associated Pods (you can have Multi-Pod with Multi-Fabric)*
  Multi-Fabric == Multi-Site == a DC infrastructure Region with multiple AZs

* Available from ACI release 3.1
ACI Multi-Site Overview

- Separate ACI Fabrics with independent APIC clusters
- 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
Typical Requirement
Creation of Two Independent Fabrics/AZs

Fabric ‘A’ (AZ 1)

Fabric ‘B’ (AZ 2)

Application workloads deployed across availability zones
Typical Requirement
Creation of Two Independent Fabrics/AZs

Application workloads deployed across availability zones
ACI Multi-Pod Deep Dive
Overview, Use Cases and Supported Topologies
ACI Multi-Pod

Overview

- 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
Configuration Zones can span any required set of switches, simplest approach may be to map a configuration zone to an availability zone, applies to infrastructure configuration and policy only.
Reducing the Impact of Configuration Errors
Introducing Configuration Zones

- Three different zone deployment modes:
  - **Enabled (default):** updates are immediately sent to all nodes part of the zone
    
    Note: a node not part of any zone is equivalent to a node part of a zone set to enabled.
  - **Disabled:** updates are postponed until the zone deployment mode is changed (or a node is removed from the zone)
  - **Triggered:** send postponed updates to the nodes part of the zone

- The deployment mode can be configured for an entire Pod or for a specified set of leaf switches
Single Availability Zone with Tenant Isolation
Isolation for ‘Virtual Network Zone and Application’ Changes

- The ACI ‘Tenant’ construct provide a **domain of application** and associated virtual network policy change
- Domain of operational change for an application (e.g. production vs. test)
ACI Multi-Pod
Supported Topologies

Intra-DC

** 50 msec support added in SW release 2.3(1)

* 10G only with QSA adapters on EX/FX spines

Two DC sites directly connected

3 (or more) DC Sites directly connected

Multiple sites interconnected by a generic L3 network
ACI Multi-Pod
SW/HW Support and Scalability Values

- All existing Nexus 9000 HW supported as leaf and spine nodes
- Maximum number of supported ACI leaf nodes (across all Pods)
  - Up to 80 leaf nodes supported with a 3 nodes APIC cluster
  - 300 leaf nodes (across Pods) with a 5 nodes APIC Cluster
  - 400 leaf nodes (across Pods) with a 7 nodes APIC Cluster (from ACI release 2.2(2e))
  - Maximum 200 leaf nodes per Pod
  - Up to 6 spines per Pod
- Maximum number of supported Pods
  - 4 in 2.0(1)/2.0(2) releases
  - 6 in 2.1(1) release
  - 10 in 2.2(2e) release
  - 12 in 3.0(1) release
APIC Cluster Deployment Considerations
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
APIC Cluster Deployment Considerations
Multi-Pod – 2 Pods Scenario

- Pod isolation scenario: changes still possible on APIC nodes in Pod1 but not in Pod2
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 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
**APIC Cluster Deployment Considerations**

**Multi-Pod – 2 Pods Scenario**

- **Pod isolation scenario**: changes still possible on APIC nodes in Pod1 but not in Pod2

- **Pod hard failure scenario**: recommendation is to activate a standby node to make the cluster fully functional again
Pod isolation scenario: 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
APIC Cluster Deployment Considerations
Multi-Pod – 2 Pods Scenario

- **Pod isolation scenario**: changes still possible on APIC nodes in Pod1 but not in Pod2

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

Pod hard failure scenario: recommendation is to activate a standby node to make the cluster fully functional again

Pod isolation scenario: same considerations as with single Pod (different behaviour across shards)
APIC Cluster Deployment Considerations

Multi-Pod – 2 Pods Scenario

- **Pod isolation scenario**: changes still possible on APIC nodes in Pod1 but not in Pod2

- **Pod hard failure scenario**: recommendation is to activate a standby node to make the cluster fully functional again

- **Pod isolation scenario**: same considerations as with single Pod (different behaviour across shards)
APIC Cluster Deployment Considerations
Multi-Pod – 2 Pods Scenario

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

- Pod isolation scenario: same considerations as with single Pod (different behaviour across shards)
- P
  Possible to restore the whole fabric state to the latest taken configuration snapshot (‘ID Recovery’ procedure – needs BU and TAC involvement)
APIC Cluster Deployment Considerations

What about a 4 Nodes APIC Cluster?

- Intermediate scalability values compared to a 3 or 5 nodes cluster scenario (up to 170 leaf nodes supported)
- Pod isolation scenario: same considerations as with 5 nodes (different behaviour across shards)
- Pod hard failure scenario
  - No chance of total loss of information for any shard
  - Can bring up a standby node in the second site to regain full majority for all the shards

Pending internal validation, scoped for Q2CY18
APIC Cluster Deployment Considerations
What about a 4 Nodes APIC Cluster?

- Intermediate scalability values compared to a 3 or 5 nodes cluster scenario (up to 170 leaf nodes supported)
- **Pod isolation scenario**: same considerations as with 5 nodes (different behaviour across shards)
- **Pod hard failure scenario**
  - No chance of total loss of information for any shard
  - Can bring up a standby node in the second site to regain full majority for all the shards

Pending internal validation, scoped for Q2CY18
APIC Cluster Deployment Considerations
What about a 4 Nodes APIC Cluster?

- Intermediate scalability values compared to a 3 or 5 nodes cluster scenario (up to 170 leaf nodes supported)
- **Pod isolation scenario**: same considerations as with 5 nodes (different behaviour across shards)
- **Pod hard failure scenario**
  - No chance of total loss of information for any shard
  - Can bring up a standby node in the second site to regain full majority for all the shards

Pending internal validation, scoped for Q2CY18
APIC Cluster Deployment Considerations
What about a 4 Nodes APIC Cluster?

- Intermediate scalability values compared to a 3 or 5 nodes cluster scenario (up to 170 leaf nodes supported)
- **Pod isolation scenario**: same considerations as with 5 nodes (different behaviour across shards)
- **Pod hard failure scenario**
  - No chance of total loss of information for any shard
  - Can bring up a standby node in the second site to regain full majority for all the shards

Pending internal validation, scoped for Q2CY18

© 2018 Cisco and/or its affiliates. All rights reserved. Cisco Public
APIC Cluster Deployment Considerations
What about a 4 Nodes APIC Cluster?

- Intermediate scalability values compared to a 3 or 5 nodes cluster scenario (up to 170 leaf nodes supported)
- **Pod isolation scenario**: same considerations as with 5 nodes (different behaviour across shards)
- **Pod hard failure scenario**
  - No chance of total loss of information for any shard
  - Can bring up a standby node in the second site to regain full majority for all the shards

Pending internal validation, scoped for Q2CY18
APIC Cluster Deployment Considerations

What about a 4 Nodes APIC Cluster?

- Intermediate scalability values compared to a 3 or 5 nodes cluster scenario (up to 170 leaf nodes supported)
- **Pod isolation scenario**: same considerations as with 5 nodes (different behaviour across shards)
- **Pod hard failure scenario**
  - No chance of total loss of information for any shard
  - Can bring up a standby node in the second site to regain full majority for all the shards

Pending internal validation, scoped for Q2CY18
APIC Cluster Deployment Considerations

Deployment Recommendations

- **Main recommendation**: deploy a 3 nodes APIC cluster when less than 80 leaf nodes are deployed across Pods
- From Q2CY18 can deploy 4 nodes if the scalability requirements are met
- When 5 (or 7) nodes are really needed for scalability reasons, follow the rule of thumb of never placing more than two APIC nodes in the same Pod (when possible):

<table>
<thead>
<tr>
<th>Pod1</th>
<th>Pod2</th>
<th>Pod3</th>
<th>Pod4</th>
<th>Pod5</th>
<th>Pod6</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Pods*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Pods</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Pods</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Pods</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6+ Pods</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*’ID Recovery’ procedure possible for recovering of lost information
Inter-Pod Connectivity Deployment Considerations
ACI Multi-Pod

Inter-Pod Network (IPN) Requirements

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

* Broadcast, Unknown unicast, Multicast
Inter-Pod Connectivity
Frequently Asked Questions

What platforms can or should I deploy in the IPN?

- Nexus 9200s, 9300-EX, but also any other switch or router supporting all the IPN requirements
- First generation Nexus 9300s/9500s not supported as IPN nodes

Can I use a 10G connection between the spines and the IPN network?

- Yes, with QSA adapters supported on the ACI spine devices
  - Available from 2.1(1h) release on EX/FX based HW
  - No plans to introduce support for first generation spines (including 9336-PQ ‘baby spine’)
Inter-Pod Connectivity
Frequently Asked Questions (2)

I have two sites connected with dark fiber/DWDM circuits, can I connect the spines back-to-back?

- No, because of multicast requirement for L2 multi-destination inter-Pod traffic

Do I need a dedicated pair of IPN devices in each Pod?

- Can use a single pair of IPN devices, but before 2.1(1h) release mandates the use of 40G/100G inter-Pod links
Control and Data Planes
ACI Multi-Pod
Auto-Provisioning of Pods

1. APIC Node 1 connected to a Leaf node in ‘Seed’ Pod 1
2. Discovery and provisioning of all the devices in the local Pod
3. Provisoning interfaces on the spines facing the IPN and EVPN control plane configuration
4. Spine 1 in Pod 2 connects to the IPN and generates DHCP requests
5. DHCP requests are relayed by the IPN devices back to the APIC in Pod 1
6. DHCP response reaches Spine 1 allowing its full provisioning
7. Discovery and provisioning of all the devices in the local Pod
8. APIC Node 2 connected to a Leaf node in Pod 2
9. APIC Node 2 joins the Cluster
10. Discover other Pods following the same procedure

For more information on how to setup an ACI Fabric from scratch: BRKACI-2004, BRKACI-2820
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 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

<table>
<thead>
<tr>
<th>IP Prefix</th>
<th>Next-Hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.1.0.0/16</td>
<td>Pod1-S1, Pod1-S2, Pod1-S3, Pod1-S4</td>
</tr>
</tbody>
</table>
ACI Multi-Pod
Inter-Pod MP-BGP EVPN Control Plane

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

1. VM1 sends traffic destined to remote EP2
2. EP2 unknown, traffic is encapsulated to the local Proxy A Spine VTEP (adding S_Class information)
3. Spine encapsulates traffic to remote Proxy B Spine VTEP
4. If policy allows it, EP2 receives the packet
5. Leaf learns remote EP1 location and enforces policy
6. Leaf 4

Policy and network information carried across Pods

= VXLAN Encap/Decap

EP1 | Leaf 4 | Proxy B
---|---|---
EP2 | Proxy B

Spine encapsulates traffic to local leaf

APIC Cluster
ACI Multi-Pod
Inter-Pod Data Plane (2)

* Leaf learns remote VM2 location (no need to enforce policy)

9
Leaf
enforces policy in ingress and, if allowed, encapsulates traffic to remote Leaf node L4

VM1 receives the packet

10
VM2 sends traffic back to remote VM1

© 2018 Cisco and/or its affiliates. All rights reserved. Cisco Public
ACI Multi-Pod
Inter-Pod Data Plane (3)

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

BD1 GIPo1: 225.1.1.128

BUM: Broadcast, Unknown Unicast, Multicast
ACI Multi-Pod
Use of Multicast for Inter-Pod BUM Traffic

Spine 2 is designated to send MG1 traffic toward the IPN

1. VM1 in BD1 generates a BUM* frame
2. BD1 has associated MG1, traffic is flooded intra-Pod via one multi-destination tree
3. IPN replicates traffic to all the PODs that joined MG1 (optimized delivery to Pods)
4. BUM frame is flooded along one of the trees associated to MG1
5. VM2 receives the BUM frame

BUM: Layer 2 Broadcast, Unknown Unicast, Multicast
ACI Multi-Pod
PIM Bidir for BUM – Supported Topologies

- 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

---

Full Mesh between remote 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

---

Directly connect local IPN devices

- Recommendation is to increase the OSPF cost of the interfaces between IPN and spines

---

```
interface Ethernet1/49.4
  description L3 Link to Pod1-SpIne1
  mtu 9150
  encapsulation dot1q 4
  ip address 192.168.1.1/31
  ip ospf cost 100
  ip ospf network point-to-point
  ip router ospf IPN area 0.0.0.0
  ip pim sparse-mode
  ip dhcp relay address 10.1.0.2
  ip dhcp relay address 10.1.0.3
```
Connecting to the External Layer 3 Domain
Connecting ACI to Layer 3 Domain

‘Traditional’ L3Out on the BL Nodes

- Connecting to WAN Edge devices at Border Leaf nodes
  - Definition of a L3Out logical construct
- VRF-lite hand-off for extending L3 multi-tenancy outside the ACI fabric
  - Each tenant defines one (or more) L3Out with a set of Logical Nodes, Logical Interfaces, peering protocol
Connecting Multi-Pod to Layer 3 Domain

‘Traditional’ L3Out on the BL Nodes

- A Pod does not need to have a dedicated WAN connection (i.e. can offer transit services to other Pods)
- Multiple WAN connections can be deployed across Pods
- Outbound traffic: by default VTEPs always select WAN connection in the local Pod based on preferred metric

By default traffic flows are hashed across L3Outs of remote Pods
Connecting Multi-Pod to Layer 3 Domain
‘Traditional’ L3Out on the BL Nodes (2)

- Asymmetric traffic paths creates issues when independent active perimeter FWs are deployed across Pods
- Tuning routing is possible to ensure ingress/egress traffic leverages always the same Pod’s L3Out
  - Reverting to an “Active/Standby” mode of operation for the deployed FWs
- Host routes advertisement is a best option to ensure all the deployed FWs are actively utilized
  - Support for host route advertisement on BL nodes planned for a future ACI release
  - Requires an L3Out connection in each Pod
Connecting ACI to Layer 3 Domain

‘GOLF’ Design

- Direct or indirect connection from spines to WAN Edge routers
- Better scalability, one protocol session for all VRFs, no longer constraint by border leaf HW table
- VXLAN handoff with MP-BGP EVPN
- Simplified tenant L3Out configuration
- Support for host routes advertisement out of the ACI Fabric
- VRF configuration automation on GOLF router through OpFlex exchange

For More Information on GOLF Deployment: LABACI-2101
GOLF and Multi-Pod Integration

Centralized and Distributed Models

Centralized WAN Edge Devices

- Common when Pods represent rooms/halls in the same physical DC
- MP-BGP EVPN peering required from spines in each Pod and the centralized WAN Edge devices

Distributed WAN Edge Devices

- Pods usually represent separate physical DCs
- Full mesh of EVPN peerings between Pods and WAN Edge routers

For more info on GOLF and Multi-Pod integration: https://www.ciscolive.com/online/connect/sessionDetail.ww?SESSION_ID=94038&backBtn=true
GOLF and Multi-Pod Integration

Inter-DC Scenario

WAN Edge devices inject host routes into the WAN or register them in the LISP database.

Host routes for endpoint belonging to public BD subnets in Pod ‘A’

Host routes for endpoint belonging to public BD subnets in Pod ‘B’

MP-BGP EVPN Control Plane

Pod ‘A’

Pod ‘B’

IPN

APIC Cluster
GOLF and Multi-Pod Integration
Inter-DC Scenario (2)

IPN WAN

Remote Router Table
- 10.10.10.0/24 B
- 10.10.10.11/32 B

G3,G4 Routing Table
- 10.10.10.0/24 B
- 10.10.10.11/32 B

Granular inbound path optimization (host route advertisement into the WAN or integration with LISP)

G1,G2 Routing Table
- 10.10.10.0/24 A
- 10.10.10.10/32 A

G1,G2

G3,G4

Granular inbound path optimization (host route advertisement into the WAN or integration with LISP)

Proxy A

Proxy B

APIC Cluster

© 2018 Cisco and/or its affiliates. All rights reserved. Cisco Public
Network Services Integration
ACI Multi-Pod
Network Services Integration Models

- Active and Standby pair deployed across Pods
- No issues with asymmetric flows but may cause traffic hair-pinning across the IPN
- Works in all scenarios from release 2.3

- Independent Active/Standby pair deployed in each Pod
- Use PBR (managed or unmanaged mode)
- Only for perimeter FW use case assuming proper solution is adopted to keep symmetric ingress/egress flows

- FW cluster deployed across Pods
- Not currently supported (scoped for 1HCY18)
Active/Standby Pair across Pods
Option 1: FW in L2 Mode
Active/Standby Pair across Pods Option 2: FW in L3 Mode and PBR

PBR Policy Applied Here

L3 Mode Active

L3 Mode Standby

WAN

IPN

= East-West
= North-South
Active/Standby Pair across Pods
Option 2: FW in L3 Mode and PBR

IPN

APIC Cluster

PBR Policy Applied Here

L3 Mode Active

L3Out-1 WAN

WAN

L3 Mode Standby

L3Out-2 WAN

= East-West

= North-South
FW in L3 Mode and L3Outs
Single L3Out Defined across Pods

APIC Cluster

Web VM1
L3 Mode Active

Web VM2
L3 Mode Standby

IPA

BDs associated to L3Outs extended via Multi-Pod
FW in L3 Mode and L3Outs
Single L3Out Defined across Pods (Dynamic Routing)

Note: supported from ACI SW releases 2.1(3), 2.2(3), 2.3(1) and 3.0(1) and deploying EX/FX HW for ACI service leaf nodes
FW in L3 Mode and L3Outs
Single L3Out Defined across Pods (Static Routing)

Note: supported from ACI SW releases 2.1(3), 2.2(3), 2.3(1) and 3.0(1) and deploying EX/FX HW for ACI service leaf nodes
Migration Scenarios
Migration Scenarios
Adding Pods to an Existing ACI

1. Distribute the APIC nodes across Pods
   - Add connections to the IPN network
   - Connect and auto-provision the other Pod(s)

2. Distribute the APIC nodes across Pods
   - Add connections to the IPN network
   - Connect and auto-provision the other Pod(s)
Migration Scenarios
Converting Stretched Fabric to Multi-Pod

- Re-cabling of the physical interconnection (especially when using DWDM circuits that must be reused)
- Re-addressing the VTEP address space for the second Pod → disruptive procedure as it requires a clean install on the second Pod
- Not internally QA-validated or recommended
ACI Multi-Pod & Multi-Site
A Reason for Both

Multi-Pod Fabric ‘A’ (AZ 1)
‘Classic’ Active/Active

Multi-Pod Fabric ‘B’ (AZ 2)
‘Classic’ Active/Active

Application workloads deployed across availability zones
Conclusions

- Cisco ACI offers different multi-fabric options that can be deployed today
- There is a solid roadmap to evolve those options in the short and mid term
  - Multi-Pod represents the natural evolution of the existing Stretched Fabric design
  - Multi-Site will replace the Dual-Fabric approach
- Cisco will offer migration options to drive the adoption of those new solutions
Where to Go for More Information

✓ ACI Stretched Fabric White Paper

✓ ACI Multi-Pod White Paper

✓ ACI Dual Fabric Design Guide

✓ ACI and GOLF High Level Integration Paper
Questions?
Use Cisco Spark to communicate with the speaker after the session

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

cs.co/ciscolivebot#BRKACI-2003
• Please complete your Online Session Evaluations after each session

• Complete 4 Session Evaluations & the Overall Conference Evaluation (available from Thursday) to receive your Cisco Live T-shirt

• All surveys can be completed via the Cisco Live Mobile App or the Communication Stations

Don’t forget: Cisco Live sessions will be available for viewing on-demand after the event at www.ciscolive.com/global/on-demand-library/.
Continue Your Education

• Demos in the Cisco campus
• Walk-in Self-Paced Labs
• Tech Circle
• Meet the Engineer 1:1 meetings
• Related sessions
Thank you