Cisco Application Centric Infrastructure Release 2.3
Design Guide
Contents

Introduction ........................................................................................................................................... 6
Components and Versions .................................................................................................................. 7
Cisco ACI Building Blocks .................................................................................................................. 7
Cisco Nexus 9000 Series Hardware ................................................................................................... 7
Leaf Switches ..................................................................................................................................... 8
Spine Switches ................................................................................................................................... 9
Cisco Application Policy Infrastructure Controller ............................................................................ 10
Fabric with Mixed Hardware or Software ........................................................................................ 11
Fabric with Different Spine Types ..................................................................................................... 11
Fabric with Different Leaf Types ....................................................................................................... 11
Fabric with Different Software Versions ............................................................................................ 12
Configuration Tools .......................................................................................................................... 12
Cisco NX-OS Style of CLI .................................................................................................................. 12
Virtual Machine Managers and Orchestration .................................................................................. 13
Evaluating the Impact of Configuration Changes ............................................................................. 13
Using Configuration Zones ............................................................................................................... 14
Formal Verification of Configurations ............................................................................................... 14
Software Upgrades and Downgrades ................................................................................................. 14
Physical Topology ............................................................................................................................. 15
Leaf-and-Spine Design ....................................................................................................................... 15
Leaf Uplinks ...................................................................................................................................... 16
Virtual Port Channel ........................................................................................................................ 16
Placement of Outside Connectivity .................................................................................................... 16
Border Leaf Design Considerations .................................................................................................... 18
VRF-lite or Layer 3 EVPN with MP-BGP ......................................................................................... 18
Dedicated Border Leaf or Computing and Border Leaf ..................................................................... 18
Using More Than Two Border Leaf Switches .................................................................................. 19
Service Leaf Considerations ............................................................................................................. 19
Overview of Cisco ACI Forwarding ..................................................................................................... 20
VXLAN Forwarding Infrastructure .................................................................................................... 20
VXLAN Tunnel Endpoints ................................................................................................................ 21
VXLAN Headers Used in the Cisco ACI Fabric .................................................................................. 23
Maximum Transmission Unit ............................................................................................................ 24
Underlay Multicast Trees .................................................................................................................. 24
Layer 2 Unicast Forwarding ............................................................................................................... 25
Mapping Database ........................................................................................................................... 27
Endpoint Learning Through ARP ..................................................................................................... 27
Mapping Database Timers ............................................................................................................... 30
Endpoint Aging with Multiple IP Addresses for the Same MAC Address ........................................ 31
Virtual Machine Mobility Considerations ......................................................................................... 32
Switching in the Overlay .................................................................................................................... 32
Layer 2 Traffic Versus Layer 3 Traffic ............................................................................................... 33
Bridge Domains and Unknown Unicast Handling .......................................................................... 33
ARP Handling ................................................................................................................................... 34
Multicast Traffic in the Overlay ......................................................................................................... 37
Mapping of Access Port VLANs to VNIDs ....................................................................................... 37
Global VNID and Locally Significant VLAN Numbers .................................................................... 38
BPDUs Handling ............................................................................................................................... 39
IP Routing in the Overlay .................................................................................................................. 40
Bridge Domain Subnets ..................................................................................................................... 40
Inside and Outside Routing .............................................................................................................. 41
Leaf Forwarding Tables: Local and Global Entries .......................................................................... 42
### Controller Design Considerations

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cisco APIC Basic Connectivity Requirements</td>
<td>47</td>
</tr>
<tr>
<td>Infrastructure VLAN</td>
<td>47</td>
</tr>
<tr>
<td>DHCP Function, Leaf Auto provisioning, and TEP Address Pool</td>
<td>48</td>
</tr>
<tr>
<td>Cisco APIC Teaming</td>
<td>49</td>
</tr>
<tr>
<td>In-Band and Out-of-Band Management of Cisco APIC</td>
<td>50</td>
</tr>
<tr>
<td>Cluster Sizing and Redundancy</td>
<td>50</td>
</tr>
<tr>
<td>Standby Controller</td>
<td>51</td>
</tr>
<tr>
<td>Fabric Recovery</td>
<td>51</td>
</tr>
<tr>
<td>Summary of Cisco APIC Design Considerations</td>
<td>52</td>
</tr>
</tbody>
</table>

### Preparing the Fabric Infrastructure

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defining a Naming Convention</td>
<td>53</td>
</tr>
<tr>
<td>Objects with Overlapping Names in Different Tenants</td>
<td>54</td>
</tr>
<tr>
<td>Interface Policy-Groups: Defining Port Channels and vPCs</td>
<td>55</td>
</tr>
<tr>
<td>Infrastructure VLAN</td>
<td>56</td>
</tr>
<tr>
<td>VTEP IP Address Pool</td>
<td>56</td>
</tr>
<tr>
<td>Fabric Access Policy Configuration</td>
<td>56</td>
</tr>
<tr>
<td>VLAN Pools and Domains</td>
<td>57</td>
</tr>
<tr>
<td>Attachable Access Entity Profiles</td>
<td>57</td>
</tr>
<tr>
<td>Interface Policies</td>
<td>57</td>
</tr>
<tr>
<td>Cisco Discovery Protocol, LLDP, and Policy Resolution</td>
<td>58</td>
</tr>
<tr>
<td>Port Channels and Virtual Port Channels</td>
<td>58</td>
</tr>
<tr>
<td>Interface Overrides</td>
<td>59</td>
</tr>
<tr>
<td>Miscabling Protocol</td>
<td>59</td>
</tr>
<tr>
<td>Storm Control</td>
<td>60</td>
</tr>
<tr>
<td>Port Tracking</td>
<td>60</td>
</tr>
<tr>
<td>Endpoint Loop Protection</td>
<td>60</td>
</tr>
<tr>
<td>Spanning-Tree Considerations</td>
<td>61</td>
</tr>
<tr>
<td>Fabric Policy Configuration</td>
<td>62</td>
</tr>
<tr>
<td>BGP Route-Reflector Policy</td>
<td>62</td>
</tr>
<tr>
<td>COOP Group Policy</td>
<td>62</td>
</tr>
</tbody>
</table>

### Designing the Tenant Network

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tenant Network Configurations</td>
<td>63</td>
</tr>
<tr>
<td>Implementing a Simple Layer 2 Topology</td>
<td>64</td>
</tr>
<tr>
<td>VRF Design Considerations</td>
<td>64</td>
</tr>
<tr>
<td>VRFs and Bridge Domains in the Common Tenant</td>
<td>64</td>
</tr>
<tr>
<td>VRFs in the Common Tenant and Bridge Domains in User Tenants</td>
<td>65</td>
</tr>
<tr>
<td>Ingress Versus Egress Filtering Design Recommendations</td>
<td>66</td>
</tr>
<tr>
<td>Bridge Domain Design Considerations</td>
<td>66</td>
</tr>
<tr>
<td>Connecting to Existing Networks and Migration Designs</td>
<td>67</td>
</tr>
<tr>
<td>Optimizing a Bridge Domain for Flooding Reduction of Layer 2 Unknown Unicast</td>
<td>67</td>
</tr>
<tr>
<td>ARP Flooding</td>
<td>68</td>
</tr>
<tr>
<td>Endpoint Learning Considerations</td>
<td>68</td>
</tr>
<tr>
<td>Endpoint Aging</td>
<td>68</td>
</tr>
<tr>
<td>Floating IP Address Considerations</td>
<td>68</td>
</tr>
<tr>
<td>Data-Plane Learning</td>
<td>68</td>
</tr>
<tr>
<td>NIC Teaming Design Considerations</td>
<td>68</td>
</tr>
<tr>
<td>Changing Bridge Domain Settings in a Production Network</td>
<td>69</td>
</tr>
<tr>
<td>Summary of Bridge Domain Recommendations</td>
<td>69</td>
</tr>
<tr>
<td>Default Gateway (Subnet) Design Considerations</td>
<td>69</td>
</tr>
<tr>
<td>Pervasive Gateway</td>
<td>69</td>
</tr>
<tr>
<td>Virtual Versus Custom MAC Address for the SVI with Fabrics Connected Using Layer 2 Extension</td>
<td>69</td>
</tr>
<tr>
<td>Virtual Versus Primary IP address for the SVI with Fabrics Connected Using Layer 2 Extension</td>
<td>69</td>
</tr>
</tbody>
</table>

Traffic Filtering in Cisco ACI................................................................. 44
EPG Lookup on the Leaf for Fabric-Attached Endpoints.......................... 44
EPG Lookup on the Leaf for Endpoints Outside the Fabric ....................... 45
Filtering Rules Programmed in the Policy CAM ...................................... 45
Virtual Port Channel.................................................................................. 47

© 2018 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.
Introduction

Cisco® Application Centric Infrastructure (Cisco ACI™) technology enables you to integrate virtual and physical workloads in a programmable, multihypervisor fabric to build a multiservice or cloud data center. The Cisco ACI fabric consists of discrete components that operate as routers and switches, but it is provisioned and monitored as a single entity.

This document describes how to implement a fabric such as the one depicted in Figure 1.

The design described in this document is based on this reference topology:

- Two spine switches interconnected to several leaf switches
- Top-of-Rack (ToR) leaf switches for server connectivity, with a mix of front-panel port speeds: 1, 10, 25, 40
- Physical and virtualized servers dual-connected to the leaf switches
- A pair of border leaf switches connected to the rest of the network with a configuration that Cisco ACI calls a Layer 3 outside (L3Out) connection
- A cluster of three Cisco Application Policy Infrastructure Controllers (APICs) dual-attached to a pair of leaf switches in the fabric

Figure 1. Cisco ACI Fabric

The network fabric in this design provides the following main services:

- Connectivity for physical and virtual workloads
- Partitioning of the fabric into multiple tenants, which may represent departments or hosted customers.
- A shared-services partition (tenant) to host servers or virtual machines whose computing workloads provide infrastructure services such as Network File System (NFS) and Microsoft Active Directory to the other tenants
- Capability to provide dedicated or shared Layer 3 routed connections to the tenants present in the fabric
Components and Versions

- Cisco ACI with APIC Release 2.3 or later
- Cisco Nexus® 9000 Series Switches with Cisco NX-OS Software Release 12.3 or later in Cisco ACI mode
- VMware ESXi hosts with VMware vSphere 6.0 integrated with Cisco ACI through either VMware vSphere Distributed Switch (vDS) or Cisco Application Virtual Switch (AVS); vSphere 6.5 U1 is supported with Cisco ACI 2.2 or later integrated with Cisco ACI through either vDS or AVS

**Note:** For more information about the support matrix for VMware vSphere with Cisco ACI through vDS, please refer to the online documentation: [https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/2-x/virtualization/b_ACI_Virtualization_Guide_2_3_1/b_ACI_Virtualization_Guide_2_3_1_chapter_011.html](https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/2-x/virtualization/b_ACI_Virtualization_Guide_2_3_1/b_ACI_Virtualization_Guide_2_3_1_chapter_011.html).


Cisco ACI Building Blocks

A Cisco ACI fabric can be built using a variety of hardware platforms. The choice depends on the following criteria:

- Type of physical layer connectivity required
- Amount of Ternary Content-Addressable Memory (TCAM) space required
- Requirements for IP-based or MAC-based classification of workloads into Endpoint Groups (EPGs)
- Analytics support
- Multicast routing in the overlay
- Support for link-layer encryption
- Fibre Channel over Ethernet (FCoE) support

Cisco Nexus 9000 Series Hardware

This section clarifies the naming conventions for the leaf nodes (N9K-C93xx) and spine line cards (N9K-X97xx) that are available for the Cisco ACI fabric.

The trailing -E and -X signify the following:

- -E: Enhanced. This label refers to the ability of the switch to classify traffic into EPGs based on the source IP address of the incoming traffic.
- -X: Analytics. This label refers to the ability of the hardware to support analytics functions. The hardware that supports analytics includes other enhancements in the policy CAM, in the buffering capabilities, and in the ability to classify traffic into EPGs.

For port speeds, the naming conventions are as follows:

- PX: 1/10-Gbps Enhanced Small Form-Factor Pluggable (SFP+)
- TX: 100-Mbps, 1-Gbps, and 10GBASE-T copper
- Y: 10/25-Gbps SFP+
- Q: 40-Gbps Quad SFP+ (QSFP+)
- C: 100-Gbps QSFP28
The switch number following the 93xx or 97xx reflects either the port count if the ports are all of the same speed or the sum of all the ports if they are of different speeds, expressed in 10-Gbps speed multiples.

For example, N9K-C93180YC-EX has 48 ports of 25 Gbps (Y) + 6 ports of 100 Gbps (C), for a total of \((48 \times 25 + 6 \times 100) / 10 = 180\) Gbps.

**Leaf Switches**

In Cisco ACI, all workloads connect to leaf switches. The leaf switches used in a Cisco ACI fabric are Top-of-the-Rack (ToR) switches. There are a number of leaf switch choices that differ based on function:

- **Port speed and medium type:** The latest Cisco ACI leaf nodes allow connectivity up to 25 and 40 Gbps to the server and uplinks of 100 Gbps to the spine.
- **Buffering and queue management:** All leaf nodes in Cisco ACI provide several advanced capabilities for flowlet load balancing to load-balance traffic more precisely, including dynamic load balancing to distribute traffic based on congestion, and dynamic packet prioritization to prioritize short-lived, latency-sensitive flows (sometimes referred to as mouse flows) over long-lived, bandwidth-intensive flows (also called elephant flows). The newest hardware also introduces more sophisticated ways to keep track and measure elephant and mouse flows and prioritize them, as well as more efficient ways to handle buffers.
- **Policy CAM size and handling:** The policy CAM is the hardware resource that allows filtering of traffic between EPGs. It is a TCAM resource in which Access Control Lists (ACLs) are expressed in terms of which EPG (security zone) can talk to which EPG (security zone). The policy CAM size varies depending on the hardware. The way in which the policy CAM handles Layer 4 operations and bidirectional contracts also varies depending on the hardware.
- **Multicast routing support in the overlay:** A Cisco ACI fabric can perform multicast routing for tenant traffic (multicast routing in the overlay), depending on the leaf model.
- **Support for analytics:** The newest leaf switches and spine line cards provide flow measurement capabilities for the purposes of analytics and application dependency mappings. These capabilities may not be enabled yet in the current software release.
- **Support for link-level encryption:** The newest leaf switches and spine line cards provide line-rate MAC Security (MACsec) encryption. This functionality is not yet enabled with the Cisco ACI Release 2.3.
- **Scale for endpoints:** One of the major features of Cisco ACI is the mapping database, which maintains the information about which endpoint is mapped to which Virtual Extensible LAN (VXLAN) tunnel endpoint (VTEP), in which bridge domain, and so on. The newest hardware has bigger TCAM tables. This means that the potential storage capacity for this mapping database is higher, even if the software may not take advantage of this additional capacity yet.
- **FCoE:** Depending on the leaf model, you can attach FCoE-capable endpoints and use the leaf node as an FCoE NPV device.
- **Support for Layer 4 through Layer 7 (L4-L7) service redirect:** The L4-L7 service graph is a feature that has been available since the first release of Cisco ACI and it works on all leaf nodes. The L4-L7 service graph redirect option allows redirection of traffic to L4-L7 devices based on protocols. It works on all hardware versions, but it has some restrictions depending on the leaf chosen.
• **Microsegmentation, or EPG classification capabilities:** Microsegmentation refers to the capability to isolate traffic within an EPG (a function similar or equivalent to the private VLAN function) and to segment traffic based on virtual machine properties, IP address, MAC address, and so on. The capability for the fabric to provide the second capability depends on both software and hardware. Traffic entering a leaf from a virtualized server running a software switch that supports Cisco OpFlex™ protocol can be classified in different EPGs based on several parameters: IP address, MAC address, virtual machine properties, and so on. Traffic entering a leaf switch from a physical server can be categorized into an EPG based on IP address or MAC address; specific hardware is required to provide this capability. Traffic entering a leaf from a virtualized server not running a software switch that supports OpFlex protocol also requires the leaf to provide hardware support for microsegmentation.

Certain design choices depend on the type of hardware that you are using in the Cisco ACI fabric: first-generation Cisco ACI leaf switches or second-generation Cisco ACI leaf switches or both.

First-generation Cisco ACI leaf switches are the Cisco Nexus 9332PQ, 9372PX-E, 9372TX-E, 9372PX, 9372TX, 9396PX, 9396TX, 93120TX, and 93128TX Switches


**Note:** For more information about the differences between the Cisco Nexus 9000 Series switches, please refer to the following:


**Spine Switches**

The spine switches are available in several form factors.

At the time of this writing you can use these fixed-form-factor spine switches:

- Cisco Nexus 9336PQ Switch
- Cisco Nexus 9364C Switch (which requires Cisco ACI software Release 3.0 or newer)


At the time of this writing the connectivity to the leaf nodes is provided by these line cards:

- N9K-X9736PQ line card
- N9K-X9732C-EX line card
- N9K-X9736C-FX line card (which requires Cisco ACI software Release 3.0 or newer)
The differences between these spine switches and line cards are as follows:

- **Port speeds:** The Cisco Nexus 9364C Switch and 9732C-EX and 9736C-FX line cards make it possible to connect uplinks at both 40- and 100-Gbps speeds.
- **Line-card mode:** newer linecards have hardware that can be used in either Cisco NX-OS mode or Cisco ACI mode.
- **Support for analytics:** Although this capability is primarily a leaf function and it may not be necessary in the spine, in the future there may be features that use this capability in the spine. The Cisco Nexus 9732C-EX and 9376C-FX line cards offer this hardware feature.
- **Support for link-level encryption:** The Cisco Nexus 9364C Switch and the N9K-X9736C-FX line card can support MACsec encryption.
- **Support for Cisco ACI Multi-Pod and Multi-Site:** Cisco ACI Multi-Pod works with all spines in terms of hardware, but at the time of this writing software support for the Cisco Nexus 9364C Switch is not yet available. Cisco ACI Multi-Site (which requires Cisco ACI Release 3.0 or newer) at the time of this writing requires the Cisco Nexus 9700-EX or 9700-FX spine linecards. Please refer to the specific documentation on Multi-Pod, Multi-Site and release notes for more details.


The Cisco ACI fabric forwards traffic based on host lookups, with a mapping database used to store the information about the leaf switch on which each IP address resides. All known endpoints in the fabric are programmed in the spine switches. The endpoints saved in the leaf forwarding table are only those that are used by the leaf in question, thus preserving hardware resources at the leaf. As a consequence, the overall scale of the fabric can be much higher than the individual scale of a single leaf.

The spine models also differ in the number of endpoints supported in the mapping database, which depends on the type and number of fabric modules installed. Modular switches equipped with six fabric modules can support the following numbers of endpoints:

- **Modular 4-slot switch:** Up to 300,000 endpoints
- **Modular 8-slot switch:** Up to 600,000 endpoints
- **Modular 16-slot switch:** Up to 1.2 million endpoints

**Note:** The number of supported endpoints is a combination of the capacity of the hardware tables, what the software allows you to configure, and what has been tested. Please refer to the Verified Scalability Guide for a given release and to the Capacity Dashboard in the APIC GUI for this information.

**Cisco Application Policy Infrastructure Controller**

The APIC is the point of configuration for policies and the place where statistics are archived and processed to provide visibility, telemetry, and application health information and enable overall management of the fabric. The controller is a physical appliance based on a Cisco UCS rack server with two 10 Gigabit Ethernet interfaces for connectivity to the leaf switches. The APIC is also equipped with 1 Gigabit Ethernet interfaces for out-of-band management. Controllers can be configured with 10GBASE-T or SFP+ Network Interface Cards (NICs), and this configuration must match the physical format supported by the leaf. In other words, if controllers are configured with 10GBASE-T, they have to be connected to a Cisco ACI leaf with 10GBASE-T ports.
Two controller models are available: Cisco APIC-M (for medium-size configurations) and APIC-L (for large configurations). At the time of this writing, the recommendation is to use APIC-M for fabrics with fewer than 1000 edge ports, and APIC-L for fabrics with more than 1000 edge ports.

**Note:** A cluster may contain a mix of different APIC sizes (M and L); however, the scalability will be that of the less powerful cluster member.


**Fabric with Mixed Hardware or Software**

**Fabric with Different Spine Types**
In Cisco ACI you can mix new and old generations of hardware in the spines and in the leaf nodes. For instance, you could have first-generation hardware leaf nodes and new-generation hardware spines, or vice versa. The type of spines will not have an influence on the features supported in the fabric. The main considerations with spine hardware are as follows:

- Uplink bandwidth between leaf and spine nodes.
- Scalability of the mapping database (which depends primarily on the type of fabric line card that is used in the spine)
- Multi-Site requires spine nodes based on the Cisco Nexus 9500 platform Cloud Scale linecards to connect to the inter-site network

You can mix spine switches of different types, but the total number of endpoints that the fabric supports is the minimum common denominator.

**Fabric with Different Leaf Types**

When mixing leaf nodes of different hardware types in the same fabric, you may have varying support of features and different levels of scalability.

In Cisco ACI, the processing intelligence resides primarily on the leaf nodes, so the choice of leaf hardware determines which features may be used (for example, multicast routing in the overlay, or FCoE).

We can distinguish between two types of features:

- Features that are local to the leaf: classification features such as IP-based EPG, copy service, service-based redirect, FCoE and potentially microsegmentation (depending on whether or not you use a software switch that supports OpFlex protocol)
- Features that are not local to a leaf: for example, Layer 3 multicast in the overlay

For the first category of features, the following behavior applies: APIC pushes the managed object to the leaf nodes regardless of the ASIC that is present. If a leaf doesn't support a given feature, it raises a fault.

For the second category of features, the ones that are not local to a leaf (currently only multicast), you should ensure that the bridge domains and Virtual Routing and Forwarding (VRF) instances configured with the feature are deployed **only** on the leaf nodes that support the feature.
Fabric with Different Software Versions
The Cisco ACI fabric is designed to operate with the same software version on all the APICs and switches. During upgrades, there may be different versions of the OS running in the same fabric.

If the leaf nodes are running different software versions, the following behavior applies: APIC pushes features based on what it is supported in its software version. If the leaf is running an older version of software and it doesn’t understand a feature, it will reject it; however, it will not raise a fault.


Configuration Tools
The networking team has a number of tools available to configure a Cisco ACI fabric, including:

- The GUI
- Representational State Transfer (REST) calls with XML or JavaScript Object Notation (JSON), which can be sent with various tools, such as the POSTMAN client
- Python scripts using libraries provided by Cisco such as the Cisco ACI Software Development Kit (SDK), the Cisco ACI toolkit, or scripts that originate REST calls
- Other SDKs such as the Ruby or Microsoft Windows PowerShell SDK
- Command-Line Interface (CLI) for monitoring each device in the fabric (which is accessible through Secure Shell [SSH] or the console and uses traditional NX-OS show commands)
- CLI similar to the NX-OS CLI for managing the entire fabric from the APIC (which is accessible through SSH to the APIC IP address)

Cisco NX-OS Style of CLI
APIC provides a CLI that is similar to the NX-OS CLI; commands can be issued directly from the APIC console. This CLI uses the REST API to program the fabric. The CLI commands are similar to NX-OS commands used to configure networking constructs such as bridge domains, subnets, VRFs, and ACLs. New commands have been introduced to support tenants, application profiles, and EPGs.

For example, a tenant is configured as follows:

```
apic1# configure
apic1(config)# tenant new
apic1(config-new)# exit
```

Figure 2 provides an example of a configuration performed using the NX-OS CLI.
Figure 2. Configuring the Cisco ACI Fabric from the Cisco NX-OS CLI

Note that the NX-OS CLI precreates certain objects to simplify the configuration steps.

The configuration scalability supported in the CLI is documented in the APIC CLI configuration guide and differs from the regular verified scalability limits, which refer to configuration performed either using the Advanced GUI or using REST calls.

These numbers are published in the document “Verified Scalability Using the CLI.”

Virtual Machine Managers and Orchestration
Cisco ACI provides support for multiple-hypervisor deployments. It is possible to integrate Cisco ACI with multiple Virtual Machine Managers (VMMs) such as VMware vSphere, Microsoft SCVMM, and OpenStack.

You can also configure networking from tools such as VMware vCenter using the vCenter plug-in. There are also plug-ins for additional orchestration and automation tools.

In a Microsoft environment, you can configure networking directly from Windows Azure Pack.

In this case, bridge domains, VRFs, tenants, and other Cisco ACI constructs can be configured directly either by the VMM administrator or by the predefined blueprints and plans that are offered by the orchestration and automation tools.

The APIC administrator controls what can be configured in the fabric by these tools by defining different users for each tool, user-to-role assignments, and the tasks that a role can perform.

Evaluating the Impact of Configuration Changes
In Cisco ACI you can make changes that apply to a very large number of leaf nodes and ports. This is a benefit of using Cisco ACI because you can manage a very large infrastructure with little configuration burden. However, if a change is made that involves many ports and you make a mistake, this change could have a significant impact. To help avoid this scenario, Cisco ACI has a Show Usage button next to each policy configuration to provide you with information to help you identify the elements that are affected by a configuration change. For example, if you decide to delete a policy that you believe is not being used by any other object in the fabric, you can verify your assumption using the Show Usage button.
Using Configuration Zones

In addition, you can scope the fabric nodes to which a configuration change is applied immediately by using a concept called configuration zones. Currently, zones can be used for policies contained in the Fabric Policies section.

Uses of configuration zones allows you to test configuration changes on a subset of leaf nodes before applying them to the entire fabric.

Each zone is a collection of leaf nodes.

Each zone can be in one of the following deployment modes:

- Disabled: Any update to a node that is a member of a disabled zone will be postponed until the zone deployment mode is changed or the node is removed from the zone.
- Enabled: Any update to a node that is a member of an enabled zone will be immediately sent. This is the default behavior. By default, switch nodes are not a member of any zone; therefore, updates will be sent immediately to those nodes.

Changes to infrastructure policies are immediately applied to nodes that are members of a zone with a deployment mode of Enabled. The same configuration changes are queued for those nodes that are members of a zone with a deployment mode of Disabled.

In this way, you can verify that configurations are working correctly on the nodes in the Enabled zone, and then subsequently change the deployment mode to Triggered in the Disabled zone. This will cause the changes to be applied on the leaf nodes within the zone in question.

Formal Verification of Configurations

In addition to using the built-in tools that Cisco ACI offer to troubleshoot the fabric, such as the embedded fault-management tools, you can formally verify configurations.

This can be done using ACILint, a Python tool built with the Cisco ACI toolkit, which works in a way similar to lint for Python scripts.

You can use the ACILint tool to verify that the configuration is formally correct in terms of object tree relationships. For example, if an EPG is configured that misses a relationship to a bridge domain, the tool will alert the user, or if a bridge domain is missing a relationship to a VRF, the tool will highlight this.

**Note:** It is outside of the scope of this document to explain the concept of the object tree in Cisco ACI and how classes and objects are organized. It is assumed that the reader is familiar with these concepts.

Software Upgrades and Downgrades

As with any other technology, prior to upgrading to another release of software you should verify any bugs that are listed in the release notes. The tool to perform this verification is at the following link:

https://bst.cloudapps.cisco.com/bugsearch
Additionally, if you are upgrading from very old software to the most recent version, chances are that Cisco has not tested that exact upgrade path. Cisco publishes guidelines about the starting and destination releases that have been validated. This is documented at this link:

Starting from Release 2.3, Cisco ACI performs a synchronous verification of the configuration for non-hardware-specific features. Certain configurations that in previous releases were accepted and would raise faults afterward are now instead verified at configuration time and not accepted.

If you need to reinstall a configuration after upgrading to Release 2.3, you can import this configuration and choose a Best Effort import model to replicate the same behavior as in versions predating Release 2.3.

Alternatively, before upgrading to Cisco ACI Release 2.3, you can try an atomic import of an existing configuration on a test setup running Release 2.3 to see if the configuration is accepted. If it is not accepted, you can fix the configuration until it is accepted and then upgrade the production environment to Release 2.3.

Before upgrading, you should also verify your Configuration Zones configuration to make sure that every configuration zone is in the Enabled state.

**Note:** It is a best practice to verify that there are no faults in the fabric before you upgrade.

**Physical Topology**

Cisco ACI uses a leaf-and-spine topology, in which each leaf switch is connected to every spine switch in the network, with no interconnection between leaf switches or spine switches:

- Each leaf and spine is connected with one or more 40 Gigabit Ethernet links or with 100 Gigabit links.
- Each APIC appliance should connect to two leaf switches for resiliency purposes.

**Leaf-and-Spine Design**

The fabric is based on a leaf-and-spine architecture in which leaf and spine nodes provide the following functions:

- **Leaf nodes:** These devices have ports connected to Classic Ethernet devices (servers, firewalls, router ports, etc.) and 40 or 100 Gigabit Ethernet uplink ports connected to the fabric spines. Leaf switches are at the edge of the fabric and provide the VXLAN Tunnel Endpoint (VTEP) function. In Cisco ACI terminology, the IP address that represents the leaf VTEP is called the Physical Tunnel Endpoint (PTEP). The leaf nodes are responsible for routing or bridging tenant packets and for applying network policies.
- **Spine nodes:** These devices interconnect leaf devices. They can also be used to build a multipod fabric by connecting a Cisco ACI pod to an IP network, or they can connect to a supported WAN device (see more details in the section “Designing External Layer 3 Connectivity”). Spine devices also provide the mapping database function—hardware used for spine devices is specifically designed for this function.

All leaf nodes connect to all spine nodes, and all spine nodes connect to all leaf nodes, but no direct connectivity is allowed between spine nodes or between leaf nodes. If you incorrectly cable spine switches to each other or leaf switches to each other, the interfaces will be disabled. You may have topologies in which certain leaf devices are not connected to all spine devices (such as in stretched fabric designs), but traffic forwarding may be suboptimal in this scenario.
Leaf Uplinks
Uplink ports on leaf switches are hard-coded as fabric (iVXLAN) ports and can connect only to spine switches.

Virtual Port Channel
Cisco ACI provides a routed fabric infrastructure with the capability to perform equal cost multipathing for Layer 2 and Layer 3 traffic, using sophisticated algorithms that optimize mouse and elephant flows and can distribute traffic based on flowlets.

In addition, Cisco ACI supports virtual-Port-Channel (vPC) technology on leaf ports to optimize server connectivity to the fabric.

It is very common for servers connected to Cisco ACI leaf nodes to be connected through vPC (that is, a port channel on the server side) to increase throughput and resilience. This is true for both physical and virtualized servers.

vPC can also be used to connect to existing Layer 2 infrastructure. vPC can also be used for L3Out connections (vPC plus Layer 3 Switch Virtual Interface [SVI]).

It is therefore important to decide which pairs of leaf nodes in the fabric should be configured as part of the same vPC domain.

When creating a vPC domain between two leaf switches, both switches must be of the same switch generation. Switches not of the same generation are not compatible vPC peers.

Placement of Outside Connectivity
The external routed connection, also known as L3Out, is the Cisco ACI building block that defines the way that the fabric connects to the external world. This can be the point of connectivity of the fabric to a campus core, to the WAN, to the MPLS-VPN cloud, and so on.

Layer 3 connectivity to the outside can be implemented in one of two ways: by attaching routers to leaf nodes (normally designated as border leaf nodes), or directly to spine switches:

- Connectivity through border leaf nodes using VRF-lite: This connectivity can be achieved with any routing-capable device that supports static routing, OSPF, Enhanced Interior Gateway Routing Protocol (EIGRP), and Border Gateway Protocol (BGP) as shown in Figure 3. Leaf node interfaces connecting to the external router are configured as Layer 3 routed interfaces, subinterfaces, or SVIs.

- Connectivity through spine ports with multiprotocol BGP (MP-BGP) EVPN and VXLAN: This connectivity option requires that the WAN device that communicates with the spines is MP-BGP EVPN capable. This feature uses VXLAN to send traffic to the spine ports as illustrated in Figure 4. Optionally, it supports OpFlex protocol. At the time of this writing, this topology is possible only with Cisco Nexus 7000 Series and 7700 platform (F3) switches, Cisco ASR 9000 Series Aggregation Services Routers, or Cisco ASR 1000 Series Aggregation Services Routers. In this topology, there is no need for direct connectivity between the WAN router and the spine. For example, there could be an OSPF-based network in between. The spine ports should be connected at 40 or 100 Gbps.
Figure 3. Connectivity to the Outside with VRF-lite (Standard L3Out in Cisco ACI)

![Figure 3](image)

Figure 4. Connectivity to the Outside with Layer 3 EVPN Services

![Figure 4](image)

The topology in Figure 3 works with any router connected to a leaf node. This topology is based on VRF-lite.

The topology in Figure 4 requires that the WAN routers support MP-BGP EVPN, OpFlex protocol, and VXLAN. The advantage of the topology in Figure 4 is that there is only one MP-BGP session for all VRFs and tenants. The other advantage is that OpFlex protocol helps ensure that the VRFs created within Cisco ACI are also created on the WAN router. In addition, the BGP route targets defined on the Cisco ACI side are matched with the route targets on the WAN router.

With the topology in Figure 4, the fabric infrastructure is extended to the WAN router, which effectively becomes the equivalent of a border leaf in the fabric.

The three main advantages of the MP-BGP EVPN solutions are:

- **VRF scalability:** At the time of this writing, this solution supports 1000 VRFs.
- **Operational simplicity:** Multiple tenants can use the same spines to connect to the outside without the need to define multiple logical nodes, interfaces, and dynamic routing on each VRF instance.
- **Automation of configurations on the WAN router device with OpFlex protocol:** This feature is optional.

It is possible to start from the topology in Figure 3 and migrate later if desired to the topology in Figure 4.
Note: This design guide does not cover the details of extending the fabric using MP-BGP EVPN, OpFlex protocol, and VXLAN.

Border Leaf Design Considerations
The following choices are available for outside connectivity:

- Use a pair of leaf nodes as both the computing and VRF-lite L3Out border leafs (or border leaf for short). These leaf nodes are used to connect endpoints and to connect to WAN or campus routers.
- Use a dedicated pair of border leaf nodes. In this case, no servers connect to the leaf. Connections are only to WAN or campus routers.
- Use Layer 3 EVPN services through the spine (instead of using a border leaf).

VRF-lite or Layer 3 EVPN with MP-BGP
At the time of this writing, the following considerations apply. For a design with fewer than 400 VRF-lite instances, or if you need multicast routing, you can use VRF-lite-based L3Out connectivity. For designs in which you foresee the need for more than 400 VRF-lite instances (L3Out) and in which there is no need for multicast routing, you should consider Layer 3 EVPN with MP-BGP on the spine because of the improved control-plane scalability and ease of configuration. L3Out connectivity based on Layer 3 EVPN with MP-BGP supports 1000 VRF instances at the time of this writing.

Dedicated Border Leaf or Computing and Border Leaf
For VRF-lite L3Out designs, you can either dedicate leaf nodes to border leaf functions, or use a leaf as both a border node and a computing node.

Using a dedicated border leaf is usually considered beneficial compared to using a leaf for both computing and VRF-lite for scalability reasons. Tables that are used to store endpoint /32 and /128 addresses can be dedicated to the classification of traffic coming from the outside into the correct EPG, and they don’t need to be shared with locally connected endpoints.

Another design consideration is the way that traffic filtering is implemented for the computing-to-outside traffic. This is controlled using the Policy Control Enforcement parameter at the VRF level. This parameter can be set to ingress or egress. For information about policy control enforcement configuration, please refer to the section “Ingress or Egress Policy Filtering for External EPGs.”

The following considerations apply to the use of VRF-lite L3Out designs:

- Attachment of endpoints to border leaf switches is fully supported when the leaf switches are all Cisco Nexus 9300-EX and Cisco 9300-FX platform switches. In presence of L3Out with SVI and external bridge domain stretched between border leafs you should disable remote IP address endpoint learning on the border leaf from Fabric > Access Policies > Global Policies > Fabric Wide Setting Policy by selecting Disable Remote EP Learn.
- If the computing leaf switches—that is, the leaf switches to which the servers are connected—are first-generation leaf switches, you need to consider one of the following options:
  - If VRF ingress policy is enabled (which is the default and recommended setting), you need to make sure that the software is Cisco ACI Release 2.2(2e) or later, and you should configure the option to disable endpoint learning on the border leaf switches. You can disable remote IP address endpoint learning on the border leaf switches from Fabric>Access Policies>Global Policies>Fabric Wide Setting Policy by selecting Disable Remote EP Learn.
You can configure the VRF instance for egress policy by selecting the Policy Control Enforcement Direction option Egress under Tenants>Networking>VRFs.

**Using More Than Two Border Leaf Switches**

Border leaf switches support three types of interfaces to connect to an external router:

- Layer 3 (routed) interface
- Sub interface with IEEE 802.1Q tagging
- Switched Virtual Interface (SVI)

When configuring an SVI on a L3Out interface, you specify a VLAN encapsulation. Specifying the same VLAN encapsulation on multiple border leaf nodes on the same L3Out interface results in the configuration of an external bridge domain.

You can configure static or dynamic routing protocol peering over a vPC for a L3Out connection by specifying the same SVI encapsulation on both vPC peers.

Depending on the hardware used for the leaf switches and on the software release, you need to keep in mind that using more than two border leaf switches as part of the same L3Out connection in Cisco ACI may have restrictions if:

- The L3Out connection consists of more than two leaf switches with SVI in the same encapsulation (VLAN)
- The border leaf switches are configured with static routing to the external device
- The connectivity from the outside device to the fabric is vPC based

These restrictions occur because traffic may be routed from one data center to the local L3Out interface and then bridged on the external bridge domain to the L3Out interface in the other data center.

With topologies consisting of more than two border leaf switches, the preferred approach is to use dynamic routing and to use a different VLAN encapsulation for each vPC pair on the L3Out SVI. This approach is preferred because the fabric can route the traffic to the L3Out interface that has reachability to the external prefix without the need to perform bridging on an outside bridge domain.

With Cisco Nexus 9300EX and newer used as border leaf nodes and starting from Cisco ACI Release 2.3, you can use static routing to an external device (for example, a firewall) with a single L3Out interface consisting of more than two border leaf switches with the same VLAN encapsulation.

**Service Leaf Considerations**

When attaching firewalls, load balancers, or other L4-L7 devices to the Cisco ACI fabric, you have the choice of whether to dedicate a leaf or leaf pair to aggregate all service devices, or to connect firewalls and load balancers to the same leaf nodes that are used to connect servers.

This is a consideration of scale. For large datacenters, it makes sense to have leaf nodes dedicated to connection of L4-L7 services.

For deployment of service graphs with the service redirect feature, dedicated service leaf nodes must be used if you are using first-generation Cisco ACI leaf switches. With Cisco Nexus 9300EX and newer switches, you can use the service graph redirect feature, and you don't have to use dedicated leaf switches for the L4-L7 service devices.
Overview of Cisco ACI Forwarding

This section provides an explanation of some of the main Cisco ACI concepts relevant to the rest of this design guide, including how forwarding works.

Note: Cisco ACI configures automatically all protocols described in this section, and you don’t need a deep understanding of them to bring up or operate a Cisco ACI fabric. However, the information in this section helps you make the right design choices, and it may be needed for advanced troubleshooting.

Cisco ACI forwarding is based on a VXLAN overlay. Leaf nodes are the VTEPs, which, in Cisco ACI terminology, are known as PTEPs.

Cisco ACI maintains a mapping database containing information about where (that is, on which TEPs) endpoints (MAC and IP addresses) reside.

Cisco ACI can perform Layer 2 or Layer 3 forwarding on the overlay. Layer 2 switched traffic carries a VXLAN Network Identifier (VNID) to identify bridge domains, whereas Layer 3 (routed) traffic carries a VNID with a number to identify the VRF.

Depending on the configuration, learning and forwarding on the leaf nodes can be based on flood and learn over the multicast tree(similar to other VXLAN implementations), or it can use the mapping database.

In a Layer 2 domain, there are two main sources of flooding: unknown unicast destination MAC traffic and ARP traffic. To reduce the amount of flooding in the fabric, Cisco ACI can do the following:

- Cisco ACI discovers the MAC or IP address—or both—of the endpoints.
- Cisco ACI forwards ARP requests to the intended recipient of the ARP request, without flooding the ARP request through the entire fabric.
- The leaf maintains a table with the MAC or IP address—to–TEP mapping based on active conversations.
- The leaf switch maintains a table with the locally discovered endpoints for local forwarding.
- Cisco ACI keeps entries updated in the tables by actively probing the endpoints whose entry is about to expire.

Furthermore, Cisco ACI classifies traffic into security zones called endpoint groups, or EPGs, and the ability for two endpoints to communicate depends on access control lists, or ACLs, that filter based on EPGs instead of IP addresses. These ACLs are modeled in Cisco ACI using a concept called a contract.

VXLAN Forwarding Infrastructure

VXLAN overlays are designed to address the shortcomings associated with regular VLANs:

- VXLAN provides greater scalability in the number of Layer 2 segments supported. Whereas VLANs are limited to just over 4000 segments, VXLAN can scale (through the use of a 24-bit ID) to up to 16 million individual segments.
- VXLAN allows extension of Layer 2 across Layer 3 boundaries through the use of MAC address in User Datagram Protocol (MAC-in-UDP) encapsulation.
VXLAN Tunnel Endpoints

The VTEP is the network device that terminates a VXLAN tunnel. A VTEP is a virtual or physical device that maps end devices to VXLAN segments and performs encapsulation and de-encapsulation. A VTEP has two interfaces: one on the local LAN segment, used to connect directly to end devices, and the other on the IP transport network, used to encapsulate Layer 2 frames into UDP packets and send them over the transport network.

Figure 5 illustrates the concept of a VTEP.

Figure 5. VTEP

In a Cisco ACI environment, VXLAN is used to encapsulate traffic inside the fabric: in other words, each leaf switch acts as a hardware VTEP, as shown in Figure 6.

Figure 6. VTEPs in a Cisco ACI Fabric

In addition to its scalability, VXLAN allows the separation of location from identity. In a traditional IP-based environment, the IP address is used to provide information about an endpoint’s identity, as well as information about where that endpoint resides in the network. An overlay technology such as VXLAN separates these functions and creates two name spaces: one for the identity, and another to signify where that endpoint resides.

In the case of Cisco ACI, the endpoint’s IP address is the identifier, and a VTEP address designates the location (leaf) of an endpoint in the network.

Cisco ACI uses a dedicated VRF and a subinterface of the uplinks as the infrastructure to carry VXLAN traffic. In Cisco ACI terminology, the transport infrastructure for VXLAN traffic is known as Overlay-1, which exists as part of tenant Infra.
Note: You may wonder why Cisco ACI uses subinterfaces to carry the VXLAN encapsulated traffic instead of using direct Layer 3 links. It would have been possible to simply run VXLAN directly on a Layer 3 interface without the need for subinterfaces, but the approach used by Cisco ACI provides flexibility to support future features and technologies.

The Overlay-1 VRF contains /32 routes to each VTEP, vPC virtual IP address, APIC and spine proxy IP address.

The VTEPs representing the leaf and spine nodes in Cisco ACI are called physical tunnel endpoints, or PTEPs. In addition to their individual PTEP addresses, spines can be addressed by a proxy TEP. This is an anycast IP address that exists across all spines and is used for forwarding lookups into the mapping database. Each VTEP address exists as a loopback on the Overlay-1 VRF. The fabric is also represented by a fabric loopback TEP (FTEP), used to encapsulate traffic in VXLAN to a vSwitch VTEP if present. Cisco ACI defines a unique FTEP address that is identical on all leaf nodes to allow mobility of downstream VTEP devices.

The control-plane protocols running inside the fabric are as follows:

- Intermediate Switch–to–Intermediate Switch (IS-IS) protocol is run on the subinterfaces to maintain infrastructure reachability.
- Council of Oracles Protocol (COOP) runs on the PTEP loopback to synchronize and to help ensure consistency of the endpoint database.
- MP-BGP is run on the PTEP loopback to advertise external WAN routes throughout the fabric.
- VXLAN tunnels to PTEPs of other leaf and spine proxy TEPs.

Each leaf maintains VXLAN tunnels with all other leaf nodes on Overlay-1. For instance, consider the inventory of the fabric in Figure 7.

Figure 7. Example of Fabric Inventory with Overlay-1 IP Addresses for Each Leaf and Spine

Looking at the configuration of the first leaf (whose IP address on Overlay-1 is 10.0.0.95), you will see the tunnels to leaf 10.0.0.90 as shown in Figure 8.
The important loopback addresses that are present on the leaf nodes are as follows:

- **Infrastructure loopback VTEP IP address:** This address is used for communication with the APIC, for MP-BGP peering, and for trace route or ping in the infrastructure.
- **vPC loopback VTEP address:** This IP address is used when the two leaf nodes forward traffic that enters through a vPC port. Traffic is forwarded by the leaf using the VXLAN encapsulation. This address is shared with the vPC peer.
- **Fabric loopback TEP address:** All leaf nodes in the fabric have an IP address to communicate with the VTEP southbound (for instance, with AVS).

**VXLAN Headers Used in the Cisco ACI Fabric**

In the Cisco ACI fabric, some extensions have been added to the VXLAN header to support the segmentation of security zones and the management of filtering rules, as well as to support the enhanced load-balancing techniques used in the fabric.

The VXLAN header used in the Cisco ACI fabric is shown in **Figure 9**.
Maximum Transmission Unit

From Figure 9 you can also calculate the Maximum Transmission Unit (MTU) that is necessary in the fabric to transport an original Ethernet frame. An Ethernet frame may arrive at a fabric access port encapsulated with a VLAN header, but the VLAN header is removed so the Ethernet frame size that is encapsulated in the VXLAN payload is typically 1500 for the original MTU size + 14 bytes of headers (the FCS is recalculated, and appended, and the IEEE 802.1q header is removed). In addition to this, the Ethernet frame transported on the fabric wire carries IP headers (20 bytes), UDP headers (8 bytes), and iVXLAN headers (8 bytes).

Therefore, the minimum MTU size that the fabric ports need to support is the original MTU + 50 bytes. The Cisco ACI fabric uplinks are configured for 9150 bytes, which is large enough to accommodate the traffic of servers sending jumbo frames.

The MTU of the fabric access ports is 9000 bytes, to accommodate servers sending jumbo frames.

**Note:** In Cisco ACI, in contrast to traditional fabrics, which have a default MTU of 1500 bytes, there is no need to configure jumbo frames manually because the MTU is already set to 9000 bytes.

Underlay Multicast Trees

Most VXLAN forwarding implementations are based on multicast trees to support broadcast, unknown unicast, and multicast traffic. In this case, VTEPs learn the location of the remote endpoints by flooding unknown unicast packets or ARP packets.

Cisco ACI implements routed multicast trees in the underlay to support multidestination traffic. If the bridge domain is not configured to use the mapping-database lookup, the location of the MAC of remote endpoints is learned as a result of the flooding over the multicast tree.

Each bridge domain is assigned a group IP outer (GIPo) address (as opposed to group IP inner [GIPi] or the multicast address in the overlay). This is also referred to as the flood GIPo for the bridge domain and is used for all multidestination traffic on the bridge domain inside the fabric.

As you can see in Figure 10, each bridge domain has a multicast IP address, which in this example is 225.0.103.240.
The configuration on the APIC defines which leaf nodes have which bridge domain enabled on them since bridge domains are deployed only to leaf nodes that actually need them (that is, leaf nodes that have endpoints that are located in those bridge domains). Learning of GI Po membership information for a given leaf occurs through the IS-IS protocol.

The multicast tree in the underlay is set up automatically without any user configuration. The roots of the trees are always the spine switches, and traffic can be distributed along multiple trees according to a tag, known as the Forwarding Tag ID (FTAG). Forwarding on the spine is independent of the bridge domain and uses only the GI Po address for the forwarding decisions.

It is possible to view the multicast tree information by choosing Fabric Policies > Global Policies as shown in Figure 11. As you can see in Figure 11, there are 12 FTAGs. With 12 FTAGs, the multidestination traffic of a given leaf can be spread across up to 12 uplinks. The FTAG trees are shared across all bridge domains.

Layer 2 Unicast Forwarding
When a traditional Ethernet switch needs to send a packet to an unknown destination MAC address, it floods it on all its ports, so that the packet will reach its destination. This mechanism is known as unknown unicast flooding. The same happens on all switches in a network, until the destination system receives the packet and eventually sends a packet itself as an answer, which is used by the switches in the path to learn where this MAC address is located.
MAC address entries learned in the switches expire after a certain time if there is no traffic using these addresses. As a consequence, unknown unicast flooding is a significant source for bandwidth waste, and is one of the reasons why network administrators tend to over provision network bandwidth.

Cisco ACI implements forwarding of Layer 2 unicast frames over VXLAN, which is optimized to remove the need for unknown unicast flooding.

Several methods exist in the industry to reduce the need for flooding of Layer 2 traffic on the multicast tree in VXLAN. These methods can be based on the pushing of endpoint-to-VTEP mapping information to the VTEPs (leaf nodes) in the fabric through a control-plane mechanism, or by having a mapping database that a VTEP consults before forwarding a packet.

Cisco ACI uses a method that combines the benefits of both:

- Cisco ACI uses a mapping database built into the spine switches that is used to perform a lookup to learn to which VTEP to forward a packet. The spine hardware contains the entire mapping database for the fabric. Looking up packets in this database does not require any additional forwarding steps because traffic from one leaf to another must always traverse a spine switch.
- Cisco ACI uses a local cache on the leaf nodes (VTEPs) to learn the MAC-to-VTEP mapping for switched traffic and IP-to-VTEP mapping for routed traffic based on active conversations, thus reducing the need to store the entire fabric mapping database on a single leaf (and therefore enabling a Cisco ACI fabric to scale way beyond the scale of a single leaf).

When Cisco ACI forwards Layer 2 frames, the forwarding behavior is as follows, depending on whether or not the destination MAC address was already learned by the originating leaf node:

- If the leaf node where the source endpoint is located already knows about the destination MAC address, the leaf node encapsulates the packet with VXLAN and sends it to the leaf node on which the destination endpoint is located.
- Assuming that the appropriate configuration is in place, when a leaf needs to forward a packet whose MAC-to-VTEP mapping is not known to that leaf (that is an unknown unicast packet), it sends the packet to one of the spines nodes for a lookup in the mapping database. The spine will know behind which egress leaf node the destination address is located, and will forward the frame accordingly. If the spine doesn’t have an entry, the packet is dropped.

**Note:** In common with other traffic types, ARP traffic populates the MAC-to-VTEP entry in the leaf. Therefore, when a node sends traffic to a specific destination MAC address, it is likely that this MAC entry is already present in the leaf cache table. This unknown unicast flooding optimization is relevant for servers in which the ARP cache timers hold an IP-to-MAC entry longer than the leaf caches the remote MAC information, in which case these servers may send traffic to a MAC address that is unknown to the leaf.

When the packet is received at the egress leaf, the local cache of the mapping database is checked and updated according to the information contained in the packet. For subsequent packets, the leaf node looks up information in the cache and forwards to the destination VTEP without the need for lookups on the spine.

In summary, Cisco ACI uses the mapping database information maintained in hardware tables in the spine switches (referred to as the spine-proxy function) to remove flooding of unknown unicast. Cisco ACI also uses cached information on the leaf nodes to maintain a local copy of the mapping.
The mapping information consists of MAC-to-VTEP, IPv4-to-VTEP, and IPv6-to-VTEP entries.

**Mapping Database**

Cisco ACI implements a mapping database, which holds the information about the MAC, IPv4 (/32), and IPv6 (/128) addresses of all endpoints and the VTEP on which they are located. This mapping information exists in hardware in the spine switches (referred to as the spine-proxy function).

The mapping database can be useful for the following:

- Routing traffic
- Maintaining an updated view of where each endpoint resides and tracking endpoint moves between leaf nodes
- Troubleshooting (iTraceroute, for instance)
- Using the endpoint attach function with service graphs

The mapping database is always populated with MAC-to-VTEP mappings, regardless of configuration. IP-to-VTEP information is populated in the mapping database only when the ip routing option is enabled in the bridge domain Layer 3 configuration.

**Note:** It is possible to have ip routing enabled without having a default gateway (subnet) configured.

MAC-to-VTEP mapping information in the spine is used only for:

- Handling unknown DMAC unicast if hardware-proxy is enabled (as described in the section “Optimizing a Bridge Domain for Flooding Reduction”)

IP-to-VTEP mapping information in the spine is used for:

- Handling ARP if ARP flooding is set to disabled
- Handling routing when the leaf node is not aware yet of the destination IP host address but the destination IP belongs to a subnet defined in the ACI fabric or when the destination does not match the Longest-Prefix-Match (LPM) table for external prefixes. The leaf is configured to send unknown destination IP traffic to the spine-proxy node by installing a subnet route for the bridge domain on the leaf and pointing to the spine-proxy TEP for this bridge domain subnet.

You can explore the content of the mapping database by opening the GUI to Fabric>Inventory>Spine>Protocols, COOP> End Point Database as in Figure 12.
Figure 12. You Can Explore the Mapping Database from the GUI

In the database you can see the MAC address, the IP address, and the VRF of the endpoint.

If the bridge domain is configured for unicast routing, the fabric learns the MAC address, IP address, VRF, and location of the endpoint in the following ways:

- The administrator can statically program the identity-to-location mapping.
- Upon creation of a new virtual machine, the VMM can update the APIC with the identity and location information. The location—that is, the port to which the virtual machine is connected—is learned through a combination of VMM-to-APIC communication (for example, the VMware ESXi host on which the virtual machine is located) and the information that the APIC retrieves from the leaf through either the OpFlex protocol, or LLDP or Cisco Discovery Protocol. This information is not programmed into the mapping database.
- Dynamic Host Configuration Protocol (DHCP) packets can be used to learn the identity-to-location mapping.
- Learning of the endpoint IPv4 or IPv6 address can occur through Address Resolution Protocol (ARP), Gratuitous ARP (GARP), and Reverse ARP (RARP) traffic; Neighbor Discovery; or data-plane routing. The leaf switch, upon discovering a new endpoint, updates its local mapping cache with an entry for this host and informs the centralized mapping database of the update through COOP.
The learning of the MAC address, bridge domain, and VTEP of the endpoint occurs on the leaf on which the endpoint generates traffic. This MAC address is then installed on the spine switches through COOP.

COOP is responsible for updating the spine-proxy database and helping ensure consistency between spine nodes. Each leaf node has COOP adjacencies with the spine nodes as shown in Figure 13.

Figure 13. COOP Adjacencies as Seen from the Leaf

The leaf switch selects one of the spine switches at random to which to send the COOP update. The spine switch in question then updates all other spine switches to help ensure consistency of the database across the nodes. The spines do not notify leaf nodes about the endpoint-to-VTEP mapping except when an endpoint moves from one leaf to another.

When an ingress leaf switch forwards a packet, it checks the local cache of the mapping database. If it does not find the endpoint address it is looking for, it encapsulates the packet with the destination set to the spine-proxy anycast address and forwards it as a unicast packet.

The any cast address is a VTEP that is configured on all spine switches in the fabric. Figure 14 shows the any cast VTEPs that are present in a fabric.

There are three any cast VTEPs in any fabric:

- Anycast VTEP for MAC address lookup
- Anycast VTEP for IPv4 address lookup
- Anycast VTEP for IPv6 address lookup

Figure 14. Anycast VTEP Addresses
The spine switch, upon receiving the packet, looks up the destination identifier address in its forwarding tables, which contain the entire mapping database. The spine then re-encapsulates the packet using the destination locator while retaining the original ingress source locator address in the VXLAN encapsulation. The packet is then forwarded as a unicast packet to the intended destination. The receiving leaf node uses information in the VXLAN packet to update its local cache with the endpoint IP and MAC information and information about which VTEP the packet is sourced from.

To be more precise, leaf nodes learn the remote endpoints and VTEP where they are located as follows:

- With ARP traffic, the leaf node learns the MAC address of the remote endpoint and the tunnel interface that the traffic is coming from.
- With bridged traffic, the leaf node learns the MAC address of the remote endpoint and the tunnel interface that the traffic is coming from.
- With flooded GARP traffic (if ARP flooding is enabled), the leaf node learns the MAC and IP addresses of the remote endpoint and the tunnel interface that the traffic is coming from.
- With routed traffic, the leaf node learns the IP address of the remote endpoint and the tunnel interface that the traffic is coming from.

In summary, the mapping database consists of:

- Spine entries learned through COOP
- Leaf switch entries (in the LST) that are learned through active conversations as depicted in Figure 15.

**Figure 15.** Mapping Tables in the Cisco ACI Fabric on Spines and on Leaf Nodes

Endpoint Learning Through ARP
Cisco ACI discovers endpoint IP addresses when traffic is routed (that is, destined for the bridge domain SVI MAC address) or through ARP gleaning.

Parsing of the ARP packets is performed partially in hardware and partially in software, and ARP packets are handled differently depending on whether the Cisco ACI leaf is a first- or second-generation switch.
**Note:** A knob in the bridge domain called data-plane learning affects the way that ARP is handled. Discovery of endpoint IP addresses based on ARP is performed in hardware through data-plane learning. If the data-plane learning knob is deselected to help ensure that Cisco ACI correctly learns the endpoint IP addresses, you should also enable the GARP-based detection option. Do not change the data-plane learning configuration except for service graph redirect configurations. For more information, refer to the section titled "Data-plane learning" later in this document.

With first-generation Cisco ACI leaf switches, Cisco ACI uses ARP packet information as follows:

- Cisco ACI learns the source MAC address of the endpoint from the payload of the ARP packet.
- Cisco ACI learns the IP address of the endpoint from the payload of the ARP packet.

With second-generation Cisco ACI leaf switches, Cisco ACI uses ARP packets information as follows:

- If the ARP packet is destined for the bridge domain subnet IP address, Cisco ACI learns the endpoint MAC address from the payload of the ARP packet.
- If the ARP packet is not directed to the bridge domain subnet IP address, Cisco ACI learns the source MAC address of the endpoint from the source MAC address of the ARP packet.
- Cisco ACI learns the endpoint IP address from the payload of the ARP packet.

**Mapping Database Timers**

Entries in the mapping database can expire. The default timer for the table that holds the host information on the leaf switches is 900 seconds. When 75 percent of this value is reached, the leaf sends three ARP requests as unicast packets in a staggered fashion (with a time delta between the requests) to check for the endpoint’s existence. If there is no ARP response, then the endpoint is removed from the local table and from the mapping database in the spine.

Leaf nodes also have a cache for remote entries that have been programmed as a result of active conversations. The purpose of this cache is to store entries for active conversations with a given remote MAC or IP address, so if there are no active conversations with this MAC or IP address, the associated entries are removed after the expiration of the timer (which is 300 seconds by default).

**Note:** You can tune this behavior by changing the **Endpoint Retention Policy** setting for the bridge domain.

**Endpoint Aging with Multiple IP Addresses for the Same MAC Address**

Cisco ACI maintains a hit bit to verify whether an endpoint is in use or not. If neither the MAC address nor the IP address of the endpoint is refreshed by the traffic, the entry ages out.

If there are multiple IP addresses for the same MAC address as in the case of a device that performs Network Address Translation (NAT), these are considered to be the same endpoint. Therefore, only one of the IP addresses needs to be hit for all the other IP addresses to be retained.

First- and second-generation Cisco ACI leaf switches differ in the way that an entry is considered to be hit:

- With first-generation Cisco ACI leaf switches, an entry is considered still valid if the traffic matches the entry IP address even if the MAC address of the packet doesn't match.
- With first- and second-generation Cisco ACI leaf switches, an entry is considered still valid if the traffic matches the MAC address and the IP address.
If you want to age out the IP addresses individually, you need to enable the IP Aging option under Fabric>Access Policies>Global Policies>IP Aging Policy.

**Virtual Machine Mobility Considerations**

Cisco ACI handles virtual machine migration from one leaf (VTEP) to another as follows *(Figure 16)*:

- When the virtual machine migrates, the hypervisor sends out a RARP message on behalf of the virtual machine. The RARP packet does not include any IP address information. However, it carries the source MAC address of the virtual machine that moved.
- The destination leaf node forwards that RARP message to the leaf on which the virtual machine was originally located.
- The destination leaf also originates a COOP message to update the mapping database.
- The spine proxy clears the IP address information associated with the MAC address of the virtual machine that moved.
- As a result of receiving the RARP packet or the COOP update message, the original leaf marks the MAC address (and IP address) of the virtual machine as a bounce entry pointing to the new switch where the VM moved to. The bounce entry is maintained for a configurable amount of time, after which it is removed.
- All traffic received by the original leaf is sent to the spine proxy. The spine proxy performs gleaning for the destination IP address if it has not yet been learned.
- Meanwhile, all leaf switches in the fabric update their forwarding tables.

*Figure 16. Virtual Machine Migration in the Cisco ACI Fabric*

---

**Switching in the Overlay**

The abstract model of Cisco ACI organizes endpoint connectivity in security zones called endpoint groups, or EPGs. EPGs are always associated with a bridge domain, and traffic between endpoints in different EPGs must be explicitly allowed by defining an ACL (contract).

Forwarding of traffic between endpoints that are allowed to communicate (through contracts) is based on bridge domains and VRFs. Bridge domains and routing instances provide the transport infrastructure for the workloads defined in the EPGs.
The relationships among the objects are as follows: the EPG references a bridge domain, and the bridge domain references a Layer 3 network (Figure 17).

**Figure 17.** Relationships Among VRF Instances, Bridge Domains, and EPGs

---

**Layer 2 Traffic Versus Layer 3 Traffic**

Layer 2 traffic in Cisco ACI terminology can be either IP or non-IP traffic. Layer 2 traffic in this context refers to traffic for which the destination MAC address is not the router MAC address. In Cisco ACI, the router MAC address refers to the MAC address of the pervasive gateway configured inside the bridge domain.

In Cisco ACI, Layer 2 traffic is forwarded according to the destination MAC address only, whether or not this involves consulting the spine-proxy database (hardware proxy enabled). When IP routing is enabled for the bridge domain, Cisco ACI can also forward a packet based on the IP-to-VTEP information.

In Cisco ACI, Layer 3 traffic is IP traffic for which the destination MAC address is the Cisco ACI bridge domain subnet MAC address (router MAC address). When forwarding on the overlay, Cisco ACI looks up the IP address of the packet only for Layer 3 traffic. For Layer 2 traffic it uses the destination MAC address (even if routing is enabled in the bridge domain).

Routing in Cisco ACI refers to host-based routing on the leaf if either /32 or /128 exists, and LPM routing is used for external routes. Routing in the spine consists exclusively of /32 and /128 routing based on the mapping database. Spines do not perform LPM routing (that is, routing for external Layer 3 destinations).

**Bridge Domains and Unknown Unicast Handling**

The concept of the bridge domain is similar to the concept of the VLAN in a traditional network. The bridge domain can act as a broadcast or flooding domain if broadcast or flooding is enabled. A bridge domain should be thought of as a distributed Layer 2 broadcast domain, which on a leaf node can be translated as a VLAN with local significance.

A bridge domain can be set to operate in flood mode for unknown unicast frames or in an optimized mode that eliminates flooding for these frames.
When operating in flood mode, Layer 2 unknown unicast traffic is flooded over the multicast tree of the bridge domain (GIPo). The multidestination tree is built using IS-IS. Each leaf advertises membership for the bridge domains that are locally enabled.

When operating in an optimized mode, the Layer 2 unknown unicast configuration is set to hardware-proxy. In this case, Layer 2 unknown unicast frames are sent to the spine-proxy any cast VTEP address. Note that enabling the hardware-proxy setting does not enable mapping database learning of MAC addresses. Learning of MAC addresses in the mapping database is always on. In addition, enabling hardware-proxy does not enable the mapping database learning of IP addresses. This happens only if IP routing is enabled.

Figure 18 illustrates forwarding with hardware-proxy enabled.

In this example, the leaf is not aware of the destination MAC address of H3, so it sends the frame to the spine. If the spine doesn't know about the destination MAC address of H3 because H3 never sent a single frame, the packet is dropped. However, in Figure 18, the spine knows that H3 is on Leaf3, so it forwards the packet to this leaf.

**Figure 18.** Forwarding of Unknown Unicast in Cisco ACI Doesn't Cause Flooding

**ARP Handling**

A source of bandwidth consumption in traditional Layer 2 networks is ARP flooding.

If an endpoint is trying to communicate with another endpoint located in the same subnet, it needs to learn the MAC address corresponding to the IP address it is trying to reach. To resolve the destination MAC address, the host sends an ARP broadcast packet, which is flooded by the Layer 2 network.

Other network implementations ameliorate this problem by using a function called proxy ARP, in which the network answers to the ARP request at the place of the destination endpoint.

Cisco ACI uses a different approach to reduce the flooding due to ARP requests. Cisco ACI forwards the ARP request to the egress endpoint as if this were a unicast packet. This is possible because Cisco ACI learns the location of each endpoint in the fabric. The egress endpoint then answers the ARP request. The network stacks of the end hosts do not notice any difference between this request and a regular ARP request, and a significant amount of bandwidth is preserved.
This option can be used only when IP-to-VTEP mapping information learning is enabled in the mapping database: in other words, when IP routing is configured for a given bridge domain.

If ARP flooding is enabled, Cisco ACI floods all ARP requests along all paths in the bridge domain to all leaf nodes that have the bridge domain configured: that is, the ARP packet is sent along the multicast tree for the GIPo of the bridge domain, as can be seen in **Figure 19**.

The multidestination tree is built using IS-IS. Each leaf advertises membership for the bridge domains that are locally enabled.

**Figure 19.** When ARP Flooding Is Enabled, ARP Requests Are Flooded on the Bridge Domain Multidestination Tree

If IP routing is enabled (which is required for the fabric to learn the endpoint IP addresses) and ARP flooding is disabled on the bridge domain, Cisco ACI will forward ARP traffic to the endpoint location based on the target IP address in the ARP packet. ARP optimizations (with the exception of ARP gleaning) do not require a subnet address to be configured for the bridge domain.

The forwarding of ARP traffic works as follows:

- If the ingress leaf knows the local endpoint IP address, it sends the ARP packet to the port where the endpoint resides, as depicted in **Figure 20**. If the endpoint is attached to that port through a vSwitch, the packet is flooded on the vSwitch within the host. The ARP packet is unaltered, the source MAC address is the MAC address of the client, and the destination MAC address is the broadcast MAC address.

- If the ingress leaf knows the remote endpoint IP address, it unicasts the ARP packet encapsulated in VXLAN to the egress leaf, after which the egress leaf sends to the port where the endpoint is known. If that endpoint is attached to that port through a vSwitch, the packet is flooded on the vSwitch within the host, as illustrated in **Figure 21**. The ARP packet is unaltered, the source MAC address is the MAC address of the client, and the destination MAC address is the broadcast MAC address.

- If the ingress leaf does not know the remote endpoint, it sends the ARP packet encapsulated in VXLAN destined for the any cast TEP of the spine proxy. If the spine knows the endpoint, it sends the ARP packet encapsulated as unicast to the egress leaf, after which the egress leaf sends to the port where the endpoint resides without flooding. The ARP packet is unaltered, the source MAC address is the MAC address of the client, and the destination MAC address is the broadcast MAC address.
- If neither the leaf nor the spine knows the endpoint, the leaf sends the ARP request to the spine. The spine drops the ARP request, but generates a new ARP request with a source of the Cisco ACI bridge domain subnet address. This new ARP request is sent to all leaf nodes in the bridge domain. The SVI of each leaf node sends this ARP request in the bridge domain. The ARP packet in this case is generated by the Cisco ACI fabric. This feature is called ARP gleaning. It is not configurable, and it requires a subnet address to be configured for the bridge domain.

Figure 20. ARP Handling with No ARP Flooding: Local Endpoint

Figure 21. ARP Handling with No ARP Flooding: Remote Endpoint
Multicast Traffic in the Overlay

Cisco ACI provides support for flooding in the bridge domain through routed multicast trees in the underlay that transport VXLAN traffic. These multicast trees are called FTAGs.

Each bridge domain is assigned a GIPo address (as opposed to a GIPi or the multicast address in the overlay). This is also referred to as the flood GIPo for the bridge domain and is used for all multidestination traffic on the bridge domain inside the fabric. This is the IP address used by the VXLAN packet.

Link local traffic such as 224.0.0.x, broadcast traffic (FFFF.FFFF.FFFF), Layer 2 multicast traffic (0100.x.x), and unknown unicast traffic are flooded in the bridge domain GIPo along one of the FTAG trees. If a bridge domain is not active on a given leaf node, the FTAG path to that leaf is pruned for that specific GIPo.

Flooding occurs across all encapsulations associated with a bridge domain (each encapsulation is represented by a flood domain VLAN [FD_VLAN]): that is, flooding is not scoped to just the VLAN encapsulation from which the originating packet came.

Only one copy of the multidestination traffic traverses the FTAG tree. The receiving leaf node performs replication across all encapsulations and across all the ports.

Within an bridge domain, you may have hosts (endpoints) that have joined a multicast group through IGMP. This multicast group is referred to by the GIPi address. Traffic destined for the GIPi is forwarded in the fabric encapsulated in the GIPo of the bridge domain. This means that all leaf nodes where the bridge domain is active receive a copy of the traffic. The receiving leaf sends the traffic to all ports that joined the GIPi. The spine does not prune traffic based on the inner GIPi address.

IGMP queries are flooded in the bridge domain, and a copy is sent to the CPU. The port from which the query came is added to the multicast router (mrouter) port list. If there is no querier outside the fabric, the bridge domain can be configured to generate IGMP queries to populate the IGMP snooping information.

The fabric can operate as a distributed IGMP querier, by originating queries in each leaf for the bridge domain and sending reports through COOP.

When the first host of a leaf sends an IGMP report, the leaf communicates the bridge domain and GIPi information to the mapping database through COOP, and when the last host leaves the group, the entry is withdrawn. This information is used by border leaf nodes to perform proxy reporting to connected mrouter.

The bridge domain offers these configuration options to control the forwarding of IP multicast traffic that is unknown (that is, for which there was no IGMP report):

- Flooding: In which case, the traffic is flooded in the bridge domain
- Optimized flooding: In which case, the traffic is sent only to mrouter ports

Mapping of Access Port VLANs to VNIDs

Cisco ACI uses VLANs to segment traffic from servers or other connected devices. This VLAN tag is locally significant on the access port and is known as the access encapsulation VLAN (encap VLAN). Note that the access encapsulation can also be a VXLAN.

Each EPG is mapped to VLANs or VXLANs on access ports. The access encapsulation VLAN has a 1:1 mapping with an FD_VLAN construct, which has a fabric encapsulation segment ID. This is to support Bridge Protocol Data Unit (BPDU) forwarding.
The same access encapsulation VLAN for the same EPG or different EPGs of the same bridge domain is associated with the same FD_VLAN fabric encapsulation (segment ID) for the purpose of BPDU flooding.

All multidestination traffic other than BPDUs is carried in the fabric as part of the bridge domain VLAN (BD_VLAN) fabric encapsulation (VXLAN VNID). The BD_VLAN construct is represented as a VLAN by CLI commands, even though it does not have any access encapsulation VLAN associated with it.

When you issue a `show vlan` command on the leaf, you see output that resembles the following:

```
leaf1-a1# show vlan
VLAN Name                     Status    Ports
---- --------------------------- ------- ----------------
1    infra:default              active   Eth1/1
3    Baekerei:BDforendpoints   active   Eth1/29, Eth1/31, Eth1/32
9    Baekerei:client-FW-server:client active Eth1/29, Eth1/31
```

VLAN 3 is a software construct that represents the bridge domain. This is the bridge domain VLAN, and you can see that it is mapped to ports 1/29, 1/31, and 1/32, which may be ports associated with either the same or different EPGs.

The second VLAN in the example shows one of the flood domain VLANs that are mapped to the EPG.

You can also see the SVI IP address of the pervasive gateway for the bridge domain `BDforeendpoints` by issuing the following command:

```
leaf1-a1# show int vlan 3
Vlan3 is up, line protocol is up
  Hardware EtherSVI, address is 0022.bdf8.19ff
  Internet Address is 20.20.20.2/24
```

**Global VNID and Locally Significant VLAN Numbers**

Consider the example shown in Figure 22. In this example, bridge domain 1 (BD1) has two EPGs, EPG1 and EPG2, and they are respectively configured with a binding to VLANs 5,6,7, and 8 and VLANs 9,10,11, and 12. The right side of the figure shows to which ports the EPGs have a binding. EPG1 has a binding to leaf 1, port 1, on VLAN 5; leaf 1, port 2, on VLAN 6; leaf 4, port 5, on VLAN 5; leaf 4, port 6, on VLAN 7; etc. These ports are all part of the same broadcast domain, regardless of which VLAN is used. For example, if you send a broadcast to leaf 1, port 1/1, on VLAN 5, it is sent out all ports that are in the bridge domain across all EPGs, regardless of the VLAN encapsulation.
**Figure 22.** Bridge Domain, EPGs and VLANs

**BPDU Handling**

When virtualized hosts are directly connected to the leaf nodes, VLANs are used to segment traffic from virtual machines. In this case, the topology is loop free because the Cisco ACI fabric is routed. It uses a multicast distribution tree for multidestination traffic and can also reduce the amount of multidestination traffic by using the spine-proxy mapping database.

When a switching device is attached to a leaf node, a mechanism is needed to help ensure interoperability between a routed VXLAN-based fabric and the loop-prevention features used by external networks to prevent loops inside Layer 2 broadcast domains.

Cisco ACI addresses this by flooding external BPDUs within a specific encapsulation, not through the entire bridge domain. Because Per-VLAN Spanning Tree carries the VLAN information embedded in the BPDU packet, the Cisco ACI fabric must also be configured to take into account the VLAN number itself.

For instance, if EPG1, port 1/1, is configured to match VLAN 5 from a switch, another port of that switch for that same Layer 2 domain can be connected only to EPG1 using the same encapsulation of VLAN 5; otherwise, the external switch would receive the BPDU for VLAN 5 tagged with a different VLAN number. Cisco ACI floods BPDUs only between the ports in the BD that have the same encapsulation.

As Figure 23 Illustrates, if you connect an external switch to leaf 1, port 1/1, the BPDU sent by the external switch would be flooded only to port 1/5 of leaf 4 because it is also part of EPG1 and tagged with VLAN 5.

BPDUs are flooded throughout the fabric with a different VNID than the one associated with the bridge domain that the EPG belongs to. This is to keep the scope of BPDU flooding separate from general multidestination traffic in the bridge domain.
Figure 23. BDPU Flooding in the Fabric

IP Routing in the Overlay

If IP routing is enabled in the bridge domain, the mapping database learns the IP address of the endpoints in addition to the MAC address. Each leaf node maintains a cache with the mapping of local endpoints (MAC and IP), and each leaf also caches the mapping of the remote endpoints (MAC and IP).

To route to networks outside the fabric, the leaf nodes are programmed with LPM tables through MP-BGP. These must be present because the spine-proxy function performs only exact-match lookups, not LPM matches.

When a packet reaches a leaf with a destination of the bridge domain MAC address, if the destination host IP address is known in the local cache, it is routed accordingly to the VTEP destination and encapsulated with the VNI of the VRF that it belongs to.

If the spine does not know either address, it tries to identify the host by sending an ARP request to all the leaf nodes that have the corresponding bridge domain and subnet, which then forward the ARP request locally.

Bridge Domain Subnets

In a Cisco ACI fabric, the bridge domain is configured as an abstracted entity that can be instantiated on any leaf, depending on whether there are endpoints attached. The SVI address for a given bridge domain in Cisco ACI is configured as a subnet. This SVI is a pervasive SVI in the sense that the same IP address and the same MAC address are instantiated on all the leaf nodes where the bridge domain exists. The MAC address is an anycast gateway MAC address. Should a virtual machine move from one leaf to another, that virtual machine does not need to change its default gateway information because the same gateway IP address and the same MAC address are present on the new leaf node.

The SVIs instantiated on the leaf for a bridge domain can be seen from the GUI in the Fabric Inventory view as shown in Figure 24.
Inside and Outside Routing

When a leaf switch receives a frame from the host, it must determine whether the destination IP address is inside or outside the fabric (that is, reachable through a L3Out connection).

The leaf forwarding table must include all the bridge domain subnets present in the fabric, and it points to the spine-proxy VTEP for all of them.

The forwarding space used to forward a packet is determined by the IP network in which it is located and to which it is going:

- Inside networks are those associated with tenants and their bridge domains.
- Outside networks are those associated with the outside routes for each of those tenants.

The lookup steps are as follows:

- If the destination IP address matches any /32 host route entry in the global station table, that means the destination is an endpoint inside the fabric and the leaf switch has already learned it, so the leaf sends the packet to the leaf node where the endpoint resides.
- If the destination IP address does not match any /32 host route entry, the leaf switch checks whether the destination IP address is within the IP address range of the bridge domain subnets of the tenant.
  - If the address is within a bridge domain subnet range, the destination IP address is inside the fabric. If the leaf switch hasn’t yet learned the destination IP address, the leaf switch encapsulates the frame in VXLAN and sets the destination address to the spine proxy VTEP.
  - The spine proxy checks the inner destination IP address against its proxy database. The packet is then sent to the leaf node where the endpoint resides.
If there is no match in the spine proxy, the spine can perform gleaning of the destination IP address by generating an ARP request with a source IP address set to the primary IP address (no virtual) of the bridge domain subnet. This ARP request is sent to all the leaf nodes on which the bridge domain is instantiated.

- If the destination address does not match either the /32 routes or a bridge domain subnet route, then the leaf looks up the routes within the external routing table. If the destination of the packet is outside the fabric, it matches it with one of the routes in the external routing table, and the leaf forwards to the VTEP of the border leaf (Figure 25).

If there is a conflict between subnets—for instance, if a bridge domain subnet 29.1.1.0/24 exists and an external route has a more specific match (for instance, 29.1.1.0/29)—the bridge domain subnet takes precedence over the external route.

Figure 25. Forwarding to Known Endpoints Inside the Fabric or to the Outside

Leaf Forwarding Tables: Local and Global Entries
The leaf contains information about local endpoints and a cache for remote endpoints. For cache misses, the leaf relies on the spine nodes to provide the VTEP information; hence, the leaf also has LPM tables programmed with the bridge domain subnets under each tenant. The APIC programs these subnets for the bridge domains that are in the tenants that are associated with the leaf, and they point to the spine-proxy VTEP.

The leaf node holds the forwarding information for local and remote endpoints:

- Information about the local endpoints is populated upon discovery of an endpoint. These entries are also synchronized with the spine-proxy forwarding tables. If a bridge domain has routing enabled, the leafs learn both IP addresses and MAC addresses. If the bridge domain is not configured for routing, the leafs learn only the MAC addresses. In first generation leafs these entries are populated in the LST table.

- Information about remote endpoints learned through active conversations. In first generation leafs these entries are populated in the GST table.

The hardware GST and LST tables are a characteristics of first generation leafs. In -EX leafs and newer there are local and remote entries but they are stored in a different way in the hardware. Just like first generation leafs, second generation leafs learn local endpoints and remote endpoints and maintain an LPM table as well. In second generation leafs a unique table is used both for /32s and LPM routes.

The remainder of this document uses the concepts of GST and LST to illustrate the forwarding decisions in ACI and even if in second generation leafs these tables are merged, for the purpose of the network administrator, the same forwarding behavior applies.
The information that is maintained in these tables is as follows:

- Local MAC and IP entries
- Remote MAC entries for which there is an active conversation: VRFs, bridge domains, and MAC addresses
- Remote IP entries for which there is an active conversation: VRFs and IP addresses

Figure 26 illustrates the forwarding tables that are present in each Cisco ACI leaf.

The traffic flows in first generation leafs use the tables as follows:

- Traffic arriving from the fabric and directed to a node attached to a leaf switch goes first through the GST and then through the LST. The source address is checked against the GST, and the destination address is checked against the LST.
- Traffic sourced by a locally attached endpoint and directed to the fabric goes first through the LST and then through the GST. The source address is checked against the LST, and the destination address is checked against the GST.
- Traffic that is locally switched goes first through the LST, then to the GST, and then back to the LST and to the destination endpoint. The GST also contains the locally connected endpoints because the destination IP address is looked up in the GST.

The lookup order is as follows:

- Match the destination IP address against the /32 routes for local or remote endpoints. If there is a match, send to the VTEP associated with the entry.
- If there is no match in the host routes table, try to match the LPM table for the subnets. If there is a match in this table, send to the spine-proxy VTEP for lookup.
- If there is no match either in the host table or in the subnet LPM table, try to match the LPM table for external routes.

Figure 26. Forwarding Tables in First Generation Leafs
Traffic Filtering in Cisco ACI

Cisco ACI introduces an important security design concept: it classifies endpoints into EPGs. Endpoints within an EPG have the same access control requirements. Cisco ACI then filters traffic based on which EPG is allowed to talk to which class-ID.

**Note:** When the EPG information is included in the VXLAN packet, it is referred to as the source class \((S\_Class)\) or destination class \((D\_Class)\), and the EPG number that identifies the EPG is known as the class identifier \((\text{class-ID})\).

**EPG Lookup on the Leaf for Fabric-Attached Endpoints**

Leaf nodes forward traffic based on the destination MAC or IP address, but traffic is filtered or redirected based on the source and destination EPGs.

The relevant tables entries for the EPG classifications are the following:

- The host route entries: these entries are the /32 or /128 IP addresses and the MAC addresses of local and remote endpoints, the VRF or bridge domain, and the EPG with which they are associated.
- Longest Prefix Match entries: These have the information about external IP addresses and masks, the VRF, and the EPG to which they belong.

For fabric-attached endpoints, the local station table is populated the first time a packet arrives on a leaf from the server. The incoming packet carries a VLAN or VXLAN ID that is used to derive the EPG information. The IP and MAC address, VRF and bridge domain, and EPG information is then programmed into the LST. With IP-based EPGs (a feature used for automatically classifying hosts into an EPG based on IP address), the LST is programmed directly by the APIC with the IP-to-EPG association.

The IP and MAC address, VRF and bridge domain, and EPG information is then programmed together with the local endpoints information. With IP-based EPGs (a feature used for automatically classifying hosts into an EPG based on IP address) instead APIC programs directly the IP-to-EPG association.

**Note:** IP-based EPG classification for physical endpoints works only when traffic is routed. IP-based EPG classification for virtual workloads using AVS works for both routed and bridged traffic.

The information about the EPG of the fabric-attached remote endpoint is implemented through active conversations. The first packet that reaches a leaf destined for a local endpoint on the leaf is cached in the global station table together with the EPG information that it is in the VXLAN header.

When forwarding a packet from a locally connected endpoint, the Cisco ACI leaf has two forwarding options for applying policy filtering:

- If the leaf has cached the information about the destination endpoint, it also knows the remote endpoint EPG and can apply the policy locally.
- If the leaf has not yet cached the information about the destination endpoint, policy filtering is not applied at ingress, and it will be the egress leaf node’s job to apply it. In the VXLAN header, the \(\text{policy\_applied} \) bits are turned off to indicate that the destination leaf (after the packet goes through the spine proxy) should apply the policy.

When receiving a packet from a remote leaf, if the policy has not yet been applied, the receiving leaf looks up the destination MAC and IP addresses of the packet in the LST. It then derives the destination EPG and applies
filtering using the source EPG included in the VXLAN header as well as the information about the destination EPG contained in the LST.

**EPG Lookup on the Leaf for Endpoints Outside the Fabric**

In contrast to locally attached endpoints, traffic entering the Cisco ACI fabric through border leaf nodes is classified based on rules that have been programmed by the APIC in advance.

The classification for endpoints outside the fabric is based on IP subnet and mask matching.

The relevant tables on the border leaf for external EPG classification (that is, for endpoints outside the fabric) depends on whether **Policy Control Enforcement Direction** for the VRF is set to **ingress** or **egress**:

- With egress policy control: Endpoint-to-outside traffic is always filtered on the border leaf. Outside-to-endpoint traffic is filtered on the border leaf after the EPG has been learned. The border leaf learns the remote endpoint information from active conversations.
- With ingress policy control: Outside-to-endpoint traffic and endpoint-to-outside traffic is always filtered at the computing leaf (that is, the leaf to which the host is attached). In this case, the border leaf does not learn remote endpoint addresses.

You can find more details in the section "Ingress or Egress Policy Filtering for External EPGs".

**Filtering Rules Programmed in the Policy CAM**

Filtering rules (known as contracts) are programmed in the policy content-addressable memory (CAM) on the leaf nodes within a Cisco ACI fabric.

**Figure 27** shows the main elements that are programmed in the policy CAM, which provides the information to match the EPGs and the filters between the EPGs.

**Figure 27.** EPGs, Contracts, and Filters

You can see the rules programmed in the policy CAM by using the command `leaf1# show zoning-rule`, which shows entries programmed as follows:

- Rule ID
- SrcEPG
- DstEPG
- FilterID
For example, with two EPGs you can see the following:

- client-EPG: Consumes the contract called `client-consumes-server`
- server-EPG: Provides the contract called `client-consumes-server`

These contracts allow the client EPG to originate a connection to the server, and the server to answer that request. This configuration translates to two policy CAM entries: one for the traffic from the client EPG to the server EPG, and one for the return traffic from the server EPG to the client EPG.

The table output for `show zoning-rule` uses the class-ID information of the EPG, which you can derive from the GUI by looking at the EPG.

**Figure 28** shows the class-ID of EPG client-EPG, and **Figure 29** shows the class-ID of the EPG server-EPG.

**Figure 28.** The Class-ID for client-EPG is 32772

**Figure 29.** The Class-ID for server-EPG is 49156
The rules programmed in the policy CAM for this contract are equivalent to the following output:

<table>
<thead>
<tr>
<th>Rule ID</th>
<th>SrcEPG</th>
<th>DstEPG</th>
<th>FilterID</th>
<th>Scope</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32772</td>
<td>49156</td>
<td>reference to the filter</td>
<td>VRF-number</td>
<td>permit</td>
</tr>
<tr>
<td>2</td>
<td>49156</td>
<td>32772</td>
<td>reference to the filter</td>
<td>VRF-number</td>
<td>permit</td>
</tr>
<tr>
<td>3</td>
<td>any</td>
<td>any</td>
<td>implicit</td>
<td>VRF-number</td>
<td>deny</td>
</tr>
</tbody>
</table>

There are two TCAM types in the leaf nodes: policy TCAM, used for the SrcEPG and DstEPG information and the filter for the Layer 4 ports, and application TCAM, which is used for Layer 4 port ranges operations.

**Virtual Port Channel**

In Cisco ACI, you can group leaf nodes to form a vPC. In contrast to the traditional Cisco Nexus vPC feature, in Cisco ACI there is no requirement for a direct vPC peer link between leaf nodes.

The peer communication happens in the fabric, and there is no out-of-band verification mechanism required to verify whether the peer is alive.

IP and MAC addresses associated with a vPC are learned as belonging to the vPC anycast VTEP of the leaf pair. ACI forwards traffic destined to a vPC connected endpoint based on ECMP of the Anycast VTEP address: only one of the vpc peer leaf nodes receives and forwards the frame.

The configuration of vPC requires two main steps: the creation of an explicit vPC protection group, which you find in Fabric> Access Policies> Switch Policies> Policies> Virtual Port-Channel Default.

In traditional NX-OS, you define which interfaces are in the same port channel by giving them the same vPC ID. To define which interfaces are part of the same port channel, Cisco ACI uses a vPC policy group. This means that you should define a different policy group for each server that is dual-connected in vPC mode.

In an NX-OS vPC, if the peer link fails, the vPC secondary device brings down its vPC ports. In Cisco ACI, as there are multiple paths between the leaf nodes, it is assumed that if the peer is unreachable, it has failed. Role election still occurs, but it is used only in the event of inconsistencies, in which case only the vPC slave brings down the vPC ports.

The configurations are maintained consistently by the APIC. As a result, inconsistencies between vPC peers are unlikely. Hence, there is rarely a need to bring down the vPC ports. Despite this, there may still be inconsistencies as a result of operational mistakes, which can cause a LAGID mismatch or the negotiation of a different port speed, or as a result of loss of APIC connectivity by a leaf.

vPC also keeps tracks of fabric uplinks, and it brings down the vPC ports if the fabric uplinks go down. This occurs to prevent dual-active scenarios.

**Controller Design Considerations**

The Cisco Application Policy Infrastructure Controller, or APIC, is a clustered network control and policy system that provides image management, bootstrapping, and policy configuration for the Cisco ACI fabric.

**Cisco APIC Basic Connectivity Requirements**

The APIC provides the following control functions:

- Policy manager: Manages the distributed policy repository responsible for the definition and deployment of the policy-based configuration of Cisco ACI
- Topology manager: Maintains up-to-date Cisco ACI topology and inventory information
- Observer: The monitoring subsystem of the APIC; serves as a data repository for Cisco ACI operational state, health, and performance information
- Boot director: Controls the booting and firmware updates of the spine and leaf switches as well as the APIC elements
- Appliance director: Manages the formation and control of the APIC appliance cluster
- Virtual Machine Manager (or VMM): Acts as an agent between the policy repository and a hypervisor and is responsible for interacting with hypervisor management systems such as VMware vCenter
- Event manager: Manages the repository for all the events and faults initiated from the APIC and the fabric nodes
- Appliance element: Manages the inventory and state of the local APIC appliance

APICs are connected and configured as follows:

- APICs should be dual-connected through the 10 Gigabit Ethernet adapters to Cisco ACI leaf switches.
- The infrastructure VLAN should not overlap with existing VLANs that may be in the existing network infrastructure to which you are connecting.
- The TEP IP address pool should not overlap with existing IP address pools that may be in use by the servers (in particular, by virtualized servers).

Figure 30 shows a typical example of the connection of the APIC to the Cisco ACI fabric.

Figure 30. Cisco APIC Connection to the Cisco ACI Fabric

APICs discover the IP addresses of other APICs in the cluster using an LLDP-based discovery process. This process maintains an appliance vector, which provides a mapping from an APIC ID to an APIC IP address and a Universally Unique Identifier (UUID) of the APIC. Initially, each APIC has an appliance vector filled with its local IP address, and all other APIC slots are marked as unknown.

Upon switch reboot, the policy element on the leaf switch gets its appliance vector from the APIC. The switch then advertises this appliance vector to all its neighbors and reports any discrepancies between its local appliance vector and the neighbors’ appliance vectors to all the APICs in the local appliance vector.

Using this process, APICs learn about the other APICs connected to the Cisco ACI fabric through leaf switches. After the APIC validates these newly discovered APICs in the cluster, the APICs update their local appliance vector and program the switches with the new appliance vector. Switches then start advertising this new appliance vector.
This process continues until all the switches have the identical appliance vector, and all APICs know the IP addresses of all the other APICs.

**Infrastructure VLAN**

The APIC communicates with the Cisco ACI fabric through a VLAN that is associated with the tenant Infra. The infrastructure VLAN number is chosen at the time of fabric provisioning. This VLAN is used for internal connectivity between the APIC and the leaf switches. Cisco ACI uses Layer 3 links in the fabric, and if the fabric needs to be extended to an external device fully managed by Cisco ACI, the traffic is carried tagged with the infrastructure VLAN. An example of this is when AVS is used as a software switch. In this case, the infrastructure VLAN must be extended between the leaf and the AVS software switch.

**Note:** The infrastructure VLAN simplifies the extension of the Cisco ACI fabric to the virtual hosts. The AVS component can then send a DHCP request on this VLAN and get an address dynamically from the Cisco ACI fabric.

See the section “Preparing the Fabric Infrastructure” for more details.

**DHCP Function, Leaf Auto provisioning, and TEP Address Pool**

The Cisco ACI fabric is brought up in a cascading manner, starting with the leaf nodes that are directly attached to the APIC. LLDP and control-plane IS-IS protocol convergence occurs in parallel to this boot process. The Cisco ACI fabric uses LLDP-based and DHCP-based fabric discovery to automatically discover the fabric switch nodes, assign the infrastructure TEP addresses, and install the firmware on the switches.

The TEP address pool is a critical part of the configuration. You should choose a non-overlapping address space for this IP range. Read the section titled “Preparing the Fabric Infrastructure” for more details.

**Figure 31** shows how boot up and auto provisioning works for the Cisco ACI nodes. The node gets an IP address from the APIC. Then it asks to download the firmware through an HTTP GET request.

**Figure 31.** Leaf or Spine Boot up Sequence
Cisco APIC Teaming

APICs are equipped with two 10 Gigabit NICs for fabric connectivity. These NICs should be connected to different leaf nodes for redundancy. APIC connectivity is automatically configured for active-backup teaming, which means that only one interface is active at any given time. You can verify (but not modify) this configuration from the Bash shell under /proc/net/bonding.

APIC software creates bond0 and bond0infiniBand VLAN interfaces for in-band connectivity to the Cisco ACI leaf switches. It also creates bond1 as an Out-Of-Band (OOB) management port.

Assuming that the infrastructure VLAN ID is 4093 (not recommended), the network interfaces are as follows:

- **Bond0**: This is the NIC bonding interface for in-band connection to the leaf switch. No IP address is assigned for this interface.
- **Bond0.4093**: This sub interface connects to the leaf switch. The VLAN ID 4093 is specified during the initial APIC software configuration. This interface obtains a dynamic IP address from the pool of TEP addresses specified in the setup configuration.
- **bond1**: This is the NIC bonding interface for OOB management. No IP address is assigned. This interface is used to bring up another interface called oobmgmt.
- **oobmgmt**: This OOB management interface allows users to access the APIC. The IP address is assigned to this interface during the APIC initial configuration process in the dialog box.

You can also see the interfaces in the GUI, as shown in Figure 32.

Figure 32. Cisco APIC Interfaces

In-Band and Out-of-Band Management of Cisco APIC

When bringing up the APIC, you enter the management IP address for OOB management as well as the default gateway. The APIC is automatically configured to use both the OOB and the in-band management networks. If later you add an in-band management network, the APIC will give preference to the in-band management network connectivity.

You can control whether APIC prefers in-band or out-of-band connectivity by configuring APIC Connectivity Preferences under Fabric>Fabric Policies>Global Policies.
Cluster Sizing and Redundancy

To support greater scale and resilience, Cisco ACI uses a concept known as data sharding for data stored in the APIC. The basic theory behind sharding is that the data repository is split into several database units, known as shards. Data is placed in a shard, and that shard is then replicated three times, with each replica assigned to an APIC appliance, as shown in Figure 33.

Figure 33. Cisco APIC Data Sharding

Figure 33 shows that the policy data, topology data, and observer data are each replicated three times on a cluster of three APICs.

In an APIC cluster, there is no one APIC that acts as a leader for all shards. For each replica, a shard leader is elected, with write operations occurring only on the elected leader. Therefore, requests arriving at an APIC are redirected to the APIC that carries the shard leader.

After recovery from a “split-brain” condition (in which APICs are no longer connected to each other), automatic reconciliation is performed based on timestamps.

The APIC can expand and shrink a cluster by defining a target cluster size.

The target size and operational size may not always match. They will not match when:

- The target cluster size is increased
- The target cluster size is decreased
- A controller node has failed

When an APIC cluster is expanded, some shard replicas shut down on the old APICs and start on the new one to help ensure that replicas continue to be evenly distributed across all APICs in the cluster.

When you add a node to the cluster, you must enter the new cluster size on an existing node.

If you need to remove an APIC node from the cluster, you must remove the appliance at the end. For example, you must remove node number 4 from a 4-node cluster; you cannot remove node number 2 from a 4-node cluster.
Each replica in the shard has a use preference, and write operations occur on the replica that is elected leader. Other replicas are followers and do not allow write operations.

If a shard replica residing on an APIC loses connectivity to other replicas in the cluster, that shard replica is said to be in a minority state. A replica in the minority state cannot be written to (that is, no configuration changes can be made). A replica in the minority state can, however, continue to serve read requests. If a cluster has only two APIC nodes, a single failure will lead to a minority situation. However, because the minimum number of nodes in an APIC cluster is three, the risk that this situation will occur is extremely low.

**Note:** When bringing up the Cisco ACI fabric, you may have a single APIC or two APICs before you have a fully functional cluster. This is not the desired end state, but Cisco ACI lets you configure the fabric with one APIC or with two APICs because the bootstrap is considered an exception.

The APIC is always deployed as a cluster of at least three controllers, and at the time of this writing, the cluster can be increased to five controllers for one Cisco ACI pod or to up to seven controllers for multiple pods. You may want to configure more than three controllers, primarily for scalability reasons.

This mechanism helps ensure that the failure of an individual APIC will not have an impact because all the configurations saved on an APIC are also stored on the other two controllers in the cluster. In that case, one of the remaining two backup APICs will be promoted to primary.

If you deploy more than three controllers, not all shards will exist on all APICs. In this case, if three out of five APICs are lost, no replica may exist. Some data that is dynamically generated and is not saved in the configurations may be in the fabric but not on the remaining APIC controllers. To restore this data without having to reset the fabric, you can use the fabric ID recovery feature.

**Standby Controller**

The standby APIC is a controller that you can keep as spare, ready to replace any active APIC in a cluster in one click. This controller doesn’t participate in policy configurations or fabric management. No data is replicated to it, not even administrator credentials (use the rescue-user login).

While standing by, the controller is in standby mode and has for instance node ID of 4, but you can make the controller active as node ID 2 if you want to replace the APIC that was previously running with node ID 2.

**Fabric Recovery**

If all the fabric controllers are lost and you have a copy of the configuration, you can restore the VNI data that is not saved as part of the configuration by reading it from the fabric, and you can merge it with the last-saved configuration by using fabric ID recovery.

In this case, you can recover the fabric with the help of the Cisco Technical Assistance Center (TAC).

The fabric ID recovery feature recovers all the TEP addresses that are assigned to the switches and node IDs. Then this feature reads all the IDs and VTEPs of the fabric and reconciles them with the exported configuration.

The recovery can be performed only from an APIC that is already part of the fabric.
**Summary of Cisco APIC Design Considerations**

Design considerations associated with APICs are as follows:

- Each APIC should be dual-connected to a pair of leaf nodes (vPC is not used, so you can connect to any two leaf nodes).
- Ideally, APIC servers should be spread across multiple leaf nodes.
- Adding more than three controllers does not increase high availability, because each database component (shard) is replicated a maximum of three times. However, increasing the number of controllers increases control-plane scalability.
- Consider using a standby APIC.
- You should consider the layout of the datacenter to place the controllers in a way that reduces the possibility that the remaining controllers will be in read-only mode, or that you will have to perform fabric ID recovery.
- You should periodically export the entire XML configuration file. (This backup copy does not include data such as the VNIs that have been allocated to bridge domains and VRF instances. Run-time data is regenerated if you restart a new fabric, or it can be rebuilt with fabric ID recovery.)

**Preparing the Fabric Infrastructure**

The Cisco ACI fabric is an IP-based fabric that implements an integrated overlay, allowing any subnet to be placed anywhere in the fabric and supporting a fabric wide mobility domain for virtualized workloads. The Spanning Tree Protocol is not required in the Cisco ACI fabric and leaf switches.

The Cisco ACI fabric has been designed with a cloud provisioning model. This model defines two main administrator roles:

- **Infrastructure administrator**: This administrator has a global view of the system, like a super user. This administrator configures the resources shared by multiple tenants and also creates the tenants.
- **Tenant administrator**: This administrator configures the resources dedicated to a particular tenant.

You do not need to create two separate roles to administer Cisco ACI. The same person can perform both roles.

This section describes the network connectivity preparation steps normally performed by the infrastructure administrator prior to handing over tenant administration to individual tenant administrators.

**Defining a Naming Convention**

Cisco ACI is built using a managed object model, where each object requires a name. A clear and consistent naming convention is therefore important to aid with manageability and troubleshooting. It is highly recommended that you define the policy naming convention **before** you deploy the Cisco ACI fabric to help ensure that all policies are named consistently.
Sample naming conventions are shown in Table 1.

<table>
<thead>
<tr>
<th>Policy Name</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tenants</td>
<td></td>
</tr>
<tr>
<td>[Function]</td>
<td>Production</td>
</tr>
<tr>
<td></td>
<td>Development</td>
</tr>
<tr>
<td>VRFs</td>
<td></td>
</tr>
<tr>
<td>[Function]</td>
<td>Trusted</td>
</tr>
<tr>
<td></td>
<td>Untrusted</td>
</tr>
<tr>
<td></td>
<td>Production</td>
</tr>
<tr>
<td></td>
<td>Development</td>
</tr>
<tr>
<td>Bridge Domains</td>
<td></td>
</tr>
<tr>
<td>[Function]</td>
<td>Web</td>
</tr>
<tr>
<td></td>
<td>App</td>
</tr>
<tr>
<td></td>
<td>App-Tier1</td>
</tr>
<tr>
<td></td>
<td>App-Tier2</td>
</tr>
<tr>
<td>Endpoint Groups</td>
<td></td>
</tr>
<tr>
<td>[Function]</td>
<td>Web</td>
</tr>
<tr>
<td></td>
<td>App</td>
</tr>
<tr>
<td></td>
<td>App-Tier1</td>
</tr>
<tr>
<td></td>
<td>App-Tier2</td>
</tr>
<tr>
<td>Attachable Access Entity Profiles</td>
<td></td>
</tr>
<tr>
<td>[Function]</td>
<td>VMM</td>
</tr>
<tr>
<td></td>
<td>Bare-Metal-Hosts</td>
</tr>
<tr>
<td></td>
<td>L3Out_N7K</td>
</tr>
<tr>
<td>VLAN Pools</td>
<td></td>
</tr>
<tr>
<td>[Function]</td>
<td>VMM</td>
</tr>
<tr>
<td></td>
<td>Bare-Metal-Hosts</td>
</tr>
<tr>
<td></td>
<td>L3Out_N7K</td>
</tr>
<tr>
<td>Domains</td>
<td></td>
</tr>
<tr>
<td>[Function]</td>
<td>Bare-Metal-Hosts</td>
</tr>
<tr>
<td></td>
<td>VMM</td>
</tr>
<tr>
<td></td>
<td>L2DCI</td>
</tr>
<tr>
<td></td>
<td>L3DCI</td>
</tr>
<tr>
<td>Contracts</td>
<td>[Prov]<em>to</em>[cons]</td>
</tr>
<tr>
<td></td>
<td>Web_to_App</td>
</tr>
<tr>
<td>Subjects</td>
<td>[Rulegroup]</td>
</tr>
<tr>
<td></td>
<td>WebTraffic</td>
</tr>
<tr>
<td>Filters</td>
<td>[Resource-Name]</td>
</tr>
<tr>
<td></td>
<td>HTTP</td>
</tr>
<tr>
<td>Application Profiles</td>
<td>[Function]</td>
</tr>
<tr>
<td></td>
<td>SAP</td>
</tr>
<tr>
<td></td>
<td>Exchange</td>
</tr>
<tr>
<td>Interface Policies</td>
<td>[Type] [Enable</td>
</tr>
<tr>
<td></td>
<td>CDP_Enable</td>
</tr>
<tr>
<td></td>
<td>CDP_Disable</td>
</tr>
<tr>
<td></td>
<td>LLDP_Disable</td>
</tr>
</tbody>
</table>
### Interface Policy Groups

<table>
<thead>
<tr>
<th>Type</th>
<th>Functionality</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>vPC_ESXi-Host1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PC_ESXi-Host1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PORT_ESXi-Host1</td>
<td></td>
</tr>
</tbody>
</table>

### Interface Profiles

<table>
<thead>
<tr>
<th>Node1</th>
<th>Node2</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>102</td>
</tr>
</tbody>
</table>

Although some naming conventions may contain a reference to the type of object (for instance, a tenant may be called Production_TNT or similar), these suffixes are often felt to be redundant, for the simple reason that each object is of a particular class in the Cisco ACI fabric. However, some customers may still prefer to identify each object name with a suffix to identify the type.

#### Objects with Overlapping Names in Different Tenants

The names you choose for VRF instances, bridge domains, contracts, and so on are made unique by the tenant in which the object is defined. Therefore, you can reuse the same name for objects that are in different tenants except for those in tenant Common.

Tenant Common is a special Cisco ACI tenant that can be used to share objects such as VRF instances and bridge domains across multiple tenants. For example, you may decide that one VRF instance is enough for your fabric, so you can define the VRF instance in tenant Common and use it from other tenants.

Objects defined in tenant Common should have a unique name across all tenants. This approach is required because Cisco ACI has a resolution framework that is designed to automatically resolve relationships when an object of a given name is not found in a tenant by looking for it in tenant Common. See [https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/1-x/aci-fundamentals/b_ACI-Fundamentals/b_ACI-Fundamentals_chapter_010001.html](https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/1-x/aci-fundamentals/b_ACI-Fundamentals/b_ACI-Fundamentals_chapter_010001.html), which states:

"In the case of policy resolution based on named relations, if a target MO with a matching name is not found in the current tenant, the ACI fabric tries to resolve in the Common tenant. For example, if the user tenant EPG contained a relationship MO targeted to a bridge domain that did not exist in the tenant, the system tries to resolve the relationship in the Common tenant. If a named relation cannot be resolved in either the current tenant or the Common tenant, the ACI fabric attempts to resolve to a default policy. If a default policy exists in the current tenant, it is used. If it does not exist, the ACI fabric looks for a default policy in the Common tenant. Bridge domain, VRF, and contract (security policy) named relations do not resolve to a default."

If you define objects with overlapping names in tenant Common and in a regular tenant, the resolution mechanism may choose the object regardless of your intended configuration.

#### Interface Policy-Groups: Defining Port Channels and vPCs

You can create interface policy groups under Fabric > Access Policies > Interface Profiles > Policy Groups> Leaf Policy Groups.

A policy group can be for a single interface for a port channel or for a vPC.

The name that you give to a policy-group of the port-channel type is equivalent to the NX-OS command `channel-group channel-number`.

The name that you give to a policy group of the vPC type is equivalent to the `channel-group channel-number` and `vpc-number` definitions.
When you assign the same policy group to multiple interfaces of the same leaf switches or of two different leaf switches, you are defining the way that all these interfaces should be bundled together.

In defining the name for the policy group, consider that you need one policy-group name for every port channel and for every vPC.

**Infrastructure VLAN**

When provisioning a Cisco ACI fabric at the APIC’s first bootstrap, the system will ask for a VLAN number to be used as the infrastructure VLAN. This VLAN is used for internal control communication between fabric nodes (leaf and spine nodes and APICs). In a scenario in which the infrastructure VLAN is extended beyond the Cisco ACI fabric (for example, when using AVS or OpenStack integration with OpFlex protocol), this VLAN may need to traverse other (not Cisco ACI) devices, as shown in Figure 34.

**Figure 34.** Infrastructure VLAN Considerations

Some platforms (for example, Cisco Nexus 9000, 7000, and 5000 Series Switches) reserve a range of VLAN IDs: typically 3968 to 4095. To avoid conflicts, it is highly recommended that you choose an infrastructure VLAN that does not fall within the reserved range of other platforms: for example, choose VLAN 3967.

**VTEP IP Address Pool**

When provisioning a Cisco ACI fabric at the APIC’s first bootstrap, you have to define a TEP range for the addressing of all Cisco ACI nodes.

Although TEPs are located inside the fabric, there are some scenarios where the TEP range may be extended beyond the fabric. When you use AVS, fabric TEP addresses are allocated to the virtual switch. Therefore, it is not advisable to overlap the internal TEP range with an external network in your data center.
The number of addresses required for the TEP address pool depends on a number of factors, including the number of APICs, number of leaf and spine nodes, number of AVS instances, and number of vPCs required. To avoid issues with address exhaustion in the future, it is strongly recommended that customers consider allocating a /16 or /17 range if possible. If this is not possible, a /22 range should be considered the absolute minimum. However, this may not be sufficient for larger deployments. It is critical to size the TEP range appropriately as it cannot be easily modified later.

Fabric Access Policy Configuration
Fabric access policies are concerned with access to the Cisco ACI fabric from the outside world and include VLAN pools; domains; interface-related configurations such as LACP, LLDP, and Cisco Discovery Protocol; and port channels and vPCs.

Access policy configuration generally follows the workflow shown in Figure 35.

Figure 35. Access Policy Configuration Workflow

VLAN Pools and Domains
In the Cisco ACI fabric, a VLAN pool is used to define a range of VLAN numbers that will ultimately be applied on specific ports on one or more leaf nodes. A VLAN pool can be configured either as a static or a dynamic pool. Static pools are generally used for hosts and devices that will be manually configured in the fabric: for example, bare-metal hosts or L4-L7 devices attached using traditional services insertion. Dynamic pools are used when the APIC needs to allocate VLANs automatically: for instance, when using VMM integration or automated services insertion (service graphs).

It is a common practice to divide VLAN pools into functional groups, as shown in Table 2.

Table 2. VLAN Pool Example

<table>
<thead>
<tr>
<th>VLAN Range</th>
<th>Type</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 - 1100</td>
<td>Static</td>
<td>Bare-metal hosts</td>
</tr>
<tr>
<td>1101 - 1200</td>
<td>Static</td>
<td>Firewalls</td>
</tr>
<tr>
<td>1201 - 1300</td>
<td>Static</td>
<td>External WAN routers</td>
</tr>
<tr>
<td>1301 - 1400</td>
<td>Dynamic</td>
<td>Virtual machines</td>
</tr>
</tbody>
</table>

A domain is used to define the scope of VLANs in the Cisco ACI fabric: in other words, where and how a VLAN pool will be used. There are a number of domain types: physical, virtual (VMM domains), external Layer 2, and external Layer 3. It is common practice to have a 1:1 mapping between a VLAN pool and a domain.

Attachable Access Entity Profiles
The Attachable Access Entity Profile (AAEP) is used to map domains (physical or virtual) to interface policies, with the end goal of mapping VLANs to interfaces. Configuring an AAEP is roughly analogous to configuring switchport access vlan x on an interface in a traditional NX-OS configuration. In addition, AAEPs allow a one-to-many relationship (if desired) to be formed between interface policy groups and domains, as shown in Figure 36.
In the example in Figure 36, an administrator needs to have both a VMM domain and a physical domain (that is, using static path bindings) on a single port or port channel. To achieve this, the administrator can map both domains (physical and virtual) to a single AAEP, which can then be associated with a single interface policy group representing the interface and port channel.

**Interface Policies**

Interface policies are responsible for the configuration of interface-level parameters, such as LLDP, Cisco Discovery Protocol, LACP, port speed, storm control, and Miscabling Protocol (MCP). Interface policies are brought together as part of an interface policy group (described in the next section).

Each type of interface policy is preconfigured with a default policy. In most cases, the feature or parameter in question is set to **disabled** as part of the default policy.

It is highly recommended that you create explicit policies for each configuration item rather than relying on and modifying the default policy. For example, for LLDP configuration, it is highly recommended that you configure two policies, titled **LLDP_Enabled** and **LLDP_Disabled** or similar, and use these policies when either enabling or disabling LLDP. This helps prevent accidental modification of the default policy, which may have a wide impact.

**Note:** You should not modify the **Fabric Access Policy LLDP default** policy because this policy is used by spines and leaf nodes for bootup and to look for an image to run. If you need to create a different default configuration for the servers, you can create a new LLDP policy and give it a name, and then use this one instead of the policy called **default**.
Cisco Discovery Protocol, LLDP, and Policy Resolution
In Cisco ACI VRF instances and bridge domains, SVIs are not rendered on a leaf device unless the resolution can be completed. The resolution that determines whether these resources are required on a given leaf is based on Cisco Discovery Protocol, LLDP, or OpFlex (when the server supports it).

Therefore, the Cisco Discovery Protocol or LLDP configuration is not just for operational convenience, but it is necessary for forwarding to work correctly.

Be sure to configure Cisco Discovery Protocol or LLDP on the interfaces that connect to virtualized servers.

If you are using fabric extenders in the Cisco ACI fabric, note that support for Cisco Discovery Protocol has been added in Cisco ACI Release 2.2. If you have a design with fabric extenders and you are running an older version of Cisco ACI, you should configure LLDP for fabric extender ports.

For more information, please refer to the section "Resolution and Deployment Immediacy of VRF Instances, Bridge Domains, EPGs, and Contracts" later in this document.

Note: If virtualized servers connect to the Cisco ACI fabric through other devices such as blade switches using a Cisco UCS fabric interconnect, be careful when changing the management IP address of these devices. A change of the management IP address may cause flapping in the Cisco Discovery Protocol or LLDP information, which could cause traffic disruption while Cisco ACI policies are being resolved.

Port Channels and Virtual Port Channels
In a Cisco ACI fabric, port channels and vPCs are created using interface policy groups. The interface policy group ties together a number of interface policies, such as Cisco Discovery Protocol, LLDP, LACP, MCP, and storm control. When creating interface policy groups for port channels and vPCs, it is important to understand how policies can and cannot be reused. Consider the example shown in Figure 37.
In this example, two servers are attached to the Cisco ACI leaf pair using vPCs. In this case, two separate interface policy groups must be configured, associated with the appropriate interface profiles (used to specify which ports will be used), and assigned to a switch profile. A common mistake is to configure a single interface policy group and attempt to reuse it for multiple port channels or vPCs on a single leaf node. However, using a single interface policy group and referencing it from multiple interface profiles will result in additional interfaces being added to the same port channel or vPC, which may not be the desired outcome.

A general rule is that a port channel or vPC interface policy group should have a 1:1 mapping to a port channel or vPC. Administrators should not try to reuse port-channel and vPC interface policy groups for more than one port channel or vPC. Note that this rule applies only to port channels and vPCs. For access port interface policy groups, these policies can be reusable.

It may be tempting for administrators to use a numbering scheme for port channels and vPCs: for example, PC1, PC2, vPC1, and so on. However, this is not recommended because Cisco ACI allocates an arbitrary number to the port channel or vPC when it is created, and it is unlikely that this number will match, which could lead to confusion. Instead, it is recommended that you use a descriptive naming scheme: for example, Firewall_Prod_A.
Interface Overrides
Consider an example where an interface policy group is configured with a certain policy, such as a policy to enable LLDP. This interface policy group is associated with a range of interfaces, for example 1/1-2, which is then applied to a set of switches (for example, 101 to 104). The administrator now decides that interface 1/2 on a specific switch only (104) must run Cisco Discovery Protocol rather than LLDP. To achieve this, interface override policies can be used.

An interface override policy refers to a port on a specific switch (for example, port 1/2 on leaf node 104) and is associated with an interface policy group. In the example here, an interface override policy for interface 1/2 on the leaf node in question can be configured and then associated with an interface policy group that has Cisco Discovery Protocol configured, as shown in Figure 38.

Figure 38. Interface Overrides

Interface overrides are configured in the Interface Policies section under Fabric Access Policies, as shown in Figure 39.
Figure 39.  Interface Override Configuration

Note that if the interface override refers to a port channel or vPC, a corresponding port channel or vPC override policy must be configured and then referenced from the interface override.

**Miscabling Protocol**

Unlike traditional networks, the Cisco ACI fabric does not participate in the Spanning Tree Protocol and does not generate BPDUs. BPDUs are instead transparently forwarded through the fabric between ports mapped to the same EPG on the same VLAN. Therefore, Cisco ACI relies to a certain degree on the loop-prevention capabilities of external devices.

Some scenarios, such as the accidental cabling of two leaf ports together, are handled directly using LLDP in the fabric. However, there are some situations where an additional level of protection is necessary. In those cases, enabling MCP can help.

MCP, if enabled, provides additional protection against misconfigurations that would otherwise result in loop situations. MCP is a lightweight protocol designed to protect against loops that cannot be discovered by either Spanning Tree Protocol or LLDP. It is recommended that you enable MCP on all ports facing external switches or similar devices. MCP must first be enabled globally through the Global Policies section of the Fabric > Access Policies tab, as shown in Figure 40.
Figure 40. MCP Configuration

<table>
<thead>
<tr>
<th>System</th>
<th>Tenants</th>
<th>Fabric</th>
<th>VM Networking</th>
<th>L4-L7 Services</th>
<th>Admin</th>
<th>Operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Policies</td>
<td></td>
<td>Inventory</td>
<td>Fabric Policies</td>
<td>Access Policies</td>
<td></td>
<td>MCP</td>
</tr>
</tbody>
</table>

The following parameters can be changed here:

- **Key**: The key to uniquely identify MCP packets within the fabric
- **Initial Delay**: The delay time before MCP begins taking action
- **Loop Detect Multiplication Factor**: The number of continuous packets that a port must receive before declaring a loop

Note that the action to disable the port upon loop detection can also be enabled here.

MCP must also be enabled on individual ports and port channels through the interface policy group configuration, as shown in Figure 41.

Note that prior to Cisco ACI Release 2.0(2f), MCP detected loops at the link level by sending MCP PDUs untagged. Software Release 2.0(2f) added support for Per-VLAN MCP. With this improvement, Cisco ACI sends MCP PDUs tagged with the VLAN ID specified in the EPG for a given link. Therefore, now MCP can be used to detect loops in non-native VLANs. If MCP is configured to disable the link, and if a loop is detected in any of the VLANs present on a physical link, MCP then disables the entire link.
Figure 41. Associating MCP with the Policy Group

Storm Control
Storm control is a feature used to monitor the levels of broadcast, multicast, and unknown unicast traffic and to suppress this traffic if a user-configured threshold is reached. Storm control on the Cisco ACI fabric is configured by opening the Fabric > Access Policies menu and choosing Interface Policies.

Storm control takes two values as configuration input:

- **Rate**: Defines a rate level against which traffic will be compared during a 1-second interval. The rate can be defined as a percentage or as the number of packets per second.
- **Max Burst Rate**: Specifies the maximum traffic rate before storm control begins to drop traffic. This rate can be defined as a percentage or the number of packets per second.

Storm control can behave differently depending on the flood settings configured at the bridge domain level. If a bridge domain is set to use hardware proxy for unknown unicast traffic, the storm control policy will apply to broadcast and multicast traffic. If, however, the bridge domain is set to flood unknown unicast traffic, storm control will apply to broadcast, multicast, and unknown unicast traffic.

Port Tracking
The port-tracking feature (first available in Release 1.2(2g)) addresses a scenario where a leaf node may lose connectivity to all spine nodes in the Cisco ACI fabric and where hosts connected to the affected leaf node in an active-standby manner may not be aware of the failure for a period of time. (Figure 42).
Figure 42. Loss of Leaf Connectivity in an Active-Standby NIC Teaming Scenario

The port-tracking feature detects a loss of fabric connectivity on a leaf node and brings down the host-facing ports. This allows the host to fail over to the second link, as shown in Figure 43.

Figure 43. Active-Standby NIC Teaming with Port Tracking Enabled

Except for very specific server deployments, servers should be dual homed, and port tracking should always be enabled.

**Endpoint Loop Protection**

Endpoint loop protection is a feature that takes action if the Cisco ACI fabric detects an endpoint moving more than a specified number of times during a given time interval. Endpoint loop protection can take one of two actions if the number of endpoint moves exceeds the configured threshold:

- Disable endpoint learning within the bridge domain.
- Disable the port to which the endpoint is connected.

The recommendation is to enable endpoint loop protection using the default parameters:

- Loop detection interval: 60
- Loop detection multiplication factor: 4
- Action: Port Disable
These parameters state that if an endpoint moves more than four times within a 60-second period, the endpoint loop-protection feature will take the specified action (disable the port).

The endpoint loop-protection feature is enabled by choosing Fabric > Access Policies > Global Policies, as shown in Figure 44.

**Figure 44.** Enabling Endpoint Loop Protection

If the action taken during an endpoint loop-protection event is to disable the port, the administrator may wish to configure automatic error disabled recovery; in other words, the Cisco ACI fabric will bring the disabled port back up after a specified period of time. This option is configured by choosing Fabric > Access Policies > Global Policies and choosing the Frequent EP Moves option, as shown in Figure 45.

**Figure 45.** Error Disabled Recovery Configuration

**Spanning-Tree Considerations**

The Cisco ACI fabric does not run Spanning Tree Protocol natively.

The flooding scope for BPDUs is different than the flooding scope for data traffic. The unknown unicast traffic and broadcast traffic are flooded within the bridge domain; spanning-tree BPDUs are flooded within a specific VLAN encapsulation (in many cases, an EPG corresponds to a VLAN, but that does not necessarily have to be the case).

**Figure 46** shows an example in which external switches connect to the fabric.
The interactions between the Cisco ACI fabric and the Spanning Tree Protocol are controlled by the EPG configuration. You can find more details and recommendations in the section “Designing the Tenant Network,” under “EPG Design Considerations.”

**Fabric Policy Configuration**

Fabric policies are found on the main Fabric tab in the APIC GUI and are concerned with configuration of the fabric itself (for example, IS-IS, management access, MP-BGP, and fabric MTU). Many of the policies found in this section should not be modified under normal circumstances, and the recommendation is to not change them. However, there are a number of policies that deserve consideration and are therefore covered in this section.

**BGP Route-Reflector Policy**

The BGP route-reflector policy controls whether MP-BGP runs within the fabric and which spine nodes should operate as BGP reflectors. When a Cisco ACI fabric is initially provisioned, MP-BGP is not enabled inside the fabric. However, in the majority of cases, MP-BGP must be enabled to allow the distribution of external routing information throughout the fabric.

It is important to note that the BGP Autonomous System (AS) number is a fabric wide configuration setting that applies across all Cisco ACI pods that are managed by the same APIC cluster (multipod).

To enable and configure MP-BGP within the fabric, modify the **BGP Route Reflector default** policy under Pod Policies on the Fabric Policies tab. The default BGP route-reflector policy should then be added to a Pod policy group and Pod profile to make the policy take effect, as shown in Figure 47.
COOP Group Policy

COOP is used within the Cisco ACI fabric to communicate endpoint mapping information between spine nodes. Starting with software Release 2.0(1m), the Cisco ACI fabric has the ability to authenticate COOP messages.

The COOP group policy (found under Pod Policies on the Fabric tab) controls the authentication of COOP messages. Two modes are available: Compatible Mode and Strict Mode. Compatible Mode accepts both authenticated and non-authenticated connections, provides backward compatibility, and is the default option. Strict Mode allows MD5 authentication connections only. The two options are shown in Figure 48.
It is recommended that you enable Strict Mode in production environments to help ensure the most secure deployment.

**Designing the Tenant Network**

The Cisco ACI fabric uses VXLAN-based overlays to provide the abstraction necessary to share the same infrastructure across multiple independent forwarding and management domains, called tenants. Figure 49 illustrates the concept.

**Figure 49.** Tenants Are Logical Divisions of the Fabric

Tenants primarily provide a management domain function (a tenant is a collection of configurations that belong to an entity), such as the development environment in Figure 49, that keeps the management of those configurations separate from those contained within other tenants.

By using VRF instances and bridge domains within the tenant, the configuration also provides a data-plane isolation function. Figure 50 illustrates the relationship among the building blocks of a tenant.
Tenant Network Configurations

In a traditional network infrastructure, the configuration steps consist of the following:

- Define a number of VLANs at the access and aggregation layers.
- Configure access ports to assign server ports to VLANs.
- Define a VRF instance at the aggregation-layer switches.
- Define an SVI for each VLAN and map these to a VRF instance.
- Define Hot Standby Router Protocol (HSRP) parameters for each SVI.
- Create and apply ACLs to control traffic between server VLANs and from server VLANs to the core.

A similar configuration in Cisco ACI requires the following steps:

- Create a tenant and a VRF instance.
- Define one or more bridge domains, configured either for traditional flooding or using the optimized configuration available in Cisco ACI.
- Create EPGs for each server security zone (these may map one to one with the VLANs in the previous configuration steps).
- Configure the default gateway (known as a subnet in Cisco ACI) as part of the bridge domain or the EPG.
- Create contracts.
- Configure the relationship between EPGs and contracts.
Implementing a Simple Layer 2 Topology

If you need to implement a simple Layer 2 topology without any routing, you can create one or more bridge domains and EPGs. You can then configure the bridge domains for hardware-proxy mode or for flood-and-learn mode (see the section “Bridge Domain Design Considerations” for more details).

**Note:** When you create any configuration or design in Cisco ACI, for objects to be instantiated and programmed into the hardware, they must meet the requirements of the object model. If a reference is missing, the object will not be instantiated.

In the Cisco ACI object model, the bridge domain is a child of the VRF instance, so even if you require a pure Layer 2 network, you must still create a VRF instance and associate the bridge domain with that VRF instance.

In summary, to create a simple Layer 2 network in Cisco ACI, perform the following steps:

1. Create a VRF instance.
2. Create a bridge domain and associate the bridge domain with the VRF instance.
3. You do not have to enable unicast routing in the bridge domain but you may still want to enable it to activate learning of IP address information.
4. You do not have to give the bridge domain a subnet IP address(SVI IP address) but you may still want to configure one to make sure ACI maintains an up to date mapping database.
5. Configure the bridge domain for either optimized switching (also called hardware-proxy mode: that is, using the mapping database for Layer 2 unknown unicast MAC addresses) or the traditional flood-and-learn behavior (preferable in case there are non-IP silent hosts).
6. Create EPGs and make sure that they have a relationship with a bridge domain (EPGs and bridge domains do not necessarily need a one-to-one relationship; you can have multiple EPGs in the same bridge domain).
7. Create contracts between EPGs as necessary, or if you want all EPGs to be able to talk to each other without any filtering, you can set the VRF instance to “unenforced.”

The rest of this section describes the design considerations if you want to create a more complex configuration.

**VRF Design Considerations**

The VRF is the data-plane segmentation element for traffic within or between tenants. Routed traffic uses the VRF as the VNID. Even if Layer 2 traffic uses the bridge domain identifier, the VRF is always necessary in the object tree for a bridge domain to be instantiated.

Therefore, you need either to create a VRF in the tenant or refer to a VRF in the Common tenant.

There is no 1:1 relationship between tenants and VRFs:

- A tenant can rely on a VRF from the Common tenant.
- A tenant can contain multiple VRFs.

A popular design approach in multitenant environments where you need to share a L3Out connection is to configure bridge domains and EPGs in individual user tenants, while referring to a VRF residing in the Common tenant.
Shared L3Out connections can be simple or complex configurations, depending on the option that you choose. This section covers the simple and recommended options of using a VRF from the Common tenant. The use of VRF leaking is a more advanced option that is covered in the section “Designing for Multitenancy and Shared Services”.

When creating a VRF, you must consider the following choices:

- Whether you want the traffic for all bridge domains and EPGs related to a VRF to be filtered according to contracts
- The policy control enforcement direction (ingress or egress) for all EPGs to outside filtering

**Note:** Each tenant can include multiple VRFs. The current number of supported VRFs per tenant is documented in the Verified Scalability guide on Cisco.com ([https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/2- x/verified_scalability/b_Verified_Scalability_2_3_1x_and_12_3_1x.html](https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/2-x/verified_scalability/b_Verified_Scalability_2_3_1x_and_12_3_1x.html)). Regardless of the published limits, it is good practice to distribute VRFs across different tenants to have better control-plane distribution on different APICs.

**VRFs and Bridge Domains in the Common Tenant**

In this scenario, you create the VRF instance and bridge domains in the Common tenant and create EPGs in the individual user tenants. You then associate the EPGs with the bridge domains of the Common tenant. This configuration can use static or dynamic routing (Figure 51).

The configuration in the Common tenant is as follows:

1. Configure a VRF under the Common tenant.
2. Configure an L3Out connection under the Common tenant and associate it with the VRF.
3. Configure the bridge domains and subnets under the Common tenant.
4. Associate the bridge domains with the VRF instance and L3Out connection.

The configuration in each tenant is as follows:

1. Under each tenant, configure EPGs and associate the EPGs with the bridge domain in the Common tenant.
2. Configure a contract and application profile under each tenant.

**Figure 51.** Shared L3Out Connection Through the Common Tenant with a VRF Instance and Bridge Domains in the Common Tenant
This approach has the following advantages:

- The L3Out connection can be configured as dynamic or static.
- Each tenant has its own EPGs and contracts.

This approach has the following disadvantages:

- Each bridge domain and subnet is visible to all tenants.
- All tenants use the same VRF instance. Hence, they cannot use overlapping IP addresses.

**VRFs in the Common Tenant and Bridge Domains in User Tenants**

In this configuration, you create a VRF in the Common tenant and create bridge domains and EPGs in the individual user tenants. Then you associate the bridge domain of each tenant with the VRF instance in the Common tenant (Figure 52). This configuration can use static or dynamic routing.

Configure the Common tenant as follows:

1. Configure a VRF instance under the Common tenant.
2. Configure an L3Out connection under the Common tenant and associate it with the VRF instance.

Configure the individual tenants as follows:

1. Configure a bridge domain and subnet under each customer tenant.
2. Associate the bridge domain with the VRF in the Common tenant and the L3Out connection.
3. Under each tenant, configure EPGs and associate the EPGs with the bridge domain in the tenant itself.
4. Configure contracts and application profiles under each tenant.

**Figure 52.** Shared L3Out Connection with the VRF Instance in the Common Tenant

The advantage of this approach is that each tenant can see only its own bridge domain and subnet. However, there is still no support for overlapping IP addresses.

**Ingress Versus Egress Filtering Design Recommendations**

Whether to apply policy enforcement at ingress or egress (in relation to the L3Outconnection and border leaf) is covered in the section “Ingress or Egress Policy Filtering for External EPG.” This section provides a simpler guide to making the choice at the time of VRF deployment.

The ingress policy enforcement feature improves policy CAM utilization on the border leaf nodes by distributing the filtering function across all regular leaf nodes.
The use of ingress policy for VRF instances and attachment of endpoints to border leaf switches are fully supported when all leaf switches in the Cisco ACI fabric are second-generation leaf switches such as the Cisco Nexus 9300-EX and Cisco 9300-FX platform switches.

If the topology contains first-generation leaf switches, and regardless of whether the border leaf is a first- or second-generation leaf switch, some caveats apply when servers connect to the border leaf nodes. If you deploy a topology that connects to the outside through border leaf switches that are also used as computing leaf switches, follow these guidelines:

- If you configure the VRF instance for ingress policy (which is the default and recommended approach), make sure that the software is Cisco ACI Release 2.2(2e) or later. Also configure the option to disable endpoint learning on the border leaf switches. You can disable remote IP address endpoint learning on the border leaf from Fabric>Access Policies>Global Policies>Fabric Wide Setting Policy by selecting Disable Remote EP Learn.
- You can configure the VRF instance for egress policy by selecting the Policy Control Enforcement Direction option Egress under Tenants>Networking>VRFs.

**Note:** The Disable Remote EP Learn configuration option disables the learning of remote endpoint IP addresses for VTEP mapping, and only on border leaf switches that have a VRF instance with ingress policy enabled. This configuration option doesn't change the learning on the MAC addresses of the endpoints.
Bridge Domain Design Considerations

The main bridge domain configuration options that should be considered to tune the bridge domain behavior are as follows:

- Whether to use hardware proxy or unknown unicast flooding
- Whether to enable or disable ARP flooding
- Whether to enable or disable unicast routing
- Whether or not to define a subnet
- Whether to define additional subnets in the same bridge domain
- Whether to constrain the learning of the endpoints to the subnet address space
- Whether to configure the endpoint retention policy

You can configure the bridge domain forwarding characteristics as optimized or as custom, as follows:

- If ARP flooding is enabled, ARP traffic will be flooded inside the fabric as per regular ARP handling in traditional networks. If this option is disabled, the fabric will attempt to use unicast to send the ARP traffic to the destination. Note that this option applies only if unicast routing is enabled on the bridge domain. If unicast routing is disabled, ARP traffic is always flooded.
- Hardware proxy for Layer 2 unknown unicast traffic is the default option. This forwarding behavior uses the mapping database to forward unknown unicast traffic to the destination port without relying on flood-and-learn behavior, as long as the MAC address is known to the spine (which means that the host is not a silent host).
- With Layer 2 unknown unicast flooding, that is, if hardware proxy is not selected, the mapping database and spine proxy are still populated with the MAC-to-VTEP information. However, the forwarding does not use the spine-proxy database. Layer 2 unknown unicast packets are flooded in the bridge domain using one of the multicast trees rooted in the spine that is scoped to the bridge domain.

The Layer 3 Configurations tab allows the administrator to configure the following parameters:

- Unicast Routing: If this setting is enabled and a subnet address is configured, the fabric provides the default gateway function and routes the traffic. Enabling unicast routing also instructs the mapping database to learn the endpoint IP-to-VTEP mapping for this bridge domain. The IP learning is not dependent upon having a subnet configured under the bridge domain.
- Subnet Address: This option configures the SVI IP addresses (default gateway) for the bridge domain.
- Limit IP Learning to Subnet: This option is similar to a unicast reverse-forwarding-path check. If this option is selected, the fabric will not learn IP addresses from a subnet other than the one configured on the bridge domain.

On a Layer 2–only bridge domain, you can still have unicast routing enabled to enable IP learning and to take advantage of the additional troubleshooting features that Cisco ACI offers.

It is possible for unicast routing to be enabled under a Layer 2–only bridge domain because traffic forwarding in Cisco ACI operates as follows:

- Cisco ACI routes traffic destined for the router MAC address.
- Cisco ACI bridges traffic that is not destined for the router MAC address.
**Note:** Many bridge domain configuration changes require removal of the entries from the mapping database and from the hardware tables of the leafs, so they are disruptive. This is true, for example, for changes from hardware proxy to unknown unicast flooding mode, and as of Cisco ACI Release 2.3, also when you enable Limit IP Learning to Subnet. When changing the bridge domain configuration, be sure to keep in mind that this change can cause traffic disruption.

**Connecting to Existing Networks and Migration Designs**

When connecting to an existing Layer 2 network, you should consider deploying a bridge domain in flood-and-learn mode. This means enabling flooding for Layer 2 unknown unicast traffic and ARP flooding.

Consider the topology of Figure 53. The reason for using unknown unicast flooding instead of hardware proxy in the bridge domain is that it Cisco ACI may take a long time to learn the MAC addresses and IP addresses of the hosts connected to the existing network (switch A and switch B). Servers connected to leaf1 and leaf2 may trigger the learning of the MAC addresses of the servers connected to switch A and B because they would ARP for them, which would then make hardware proxy a viable option. Now imagine that the link connecting switch A to leaf 3 goes down, and that the link connecting switch B to leaf 4 becomes a forwarding link. All the endpoints learned on leaf 3 are now cleared from the mapping database. Servers connected to leaf 1 and leaf2 still have valid ARP entries for the hosts connected to switch A and switch B, so they won't implement ARP immediately. If the servers connected to leaf 1 and leaf 2 send frames to the servers connected to switch A and switch B, these will be dropped up until the servers connected to switch A and switch B send out some traffic that updates the entries on leaf 4. Switches A and B may not flood any traffic to the Cisco ACI leaf switches until the MAC entries expire in the existing network forwarding tables. The servers in the existing network may not send an ARP request until the ARP caches expire. Therefore, to avoid traffic disruption you should set the bridge domain that connects to switches A and B for unknown unicast flooding.

**Figure 53.** Using Unknown Unicast Flooding for Bridge Domains Connected to Existing Network Infrastructure
When using the bridge domain configured for Layer 2 unknown unicast flooding, you may also want to select the option called Clear Remote MAC Entries. Clear Remote MAC Entries helps ensure that when the leaf ports connected to the active Layer 2 path go down, the MAC address entries of the endpoints are cleared both on the local leaf (as for leaf 3 in the previous example) and associated remote endpoint entries in the tables of the other leaf switches in the fabric (as for leaf switches 1, 2, 4, 5, and 6 in the previous example). The reason for this setting is that the alternative Layer 2 path between switch B and leaf 4 in the example may be activated, and clearing the remote table on all the leaf switches prevents traffic from becoming black-holed to the previous active Layer 2 path (leaf 3 in the example).

**Optimizing a Bridge Domain for Flooding Reduction of Layer 2 Unknown Unicast**

If you want to reduce flooding in the bridge domain that is caused by Layer 2 unknown unicast frames, you should configure the following options:

- Configure hardware proxy to remove unknown unicast flooding.
- Configure unicast routing to enable the learning of endpoint IP addresses (regardless of whether you need to route traffic).
- Configure a subnet to enable the bridge domain to use ARP to resolve endpoints when the endpoint retention policy expires, and also to enable the bridge domain to perform ARP gleaning for silent hosts. When configuring a subnet, you also should enable the option Limit IP Learning to Subnet.
- Define an endpoint retention policy. This is important if the ARP cache timeout of hosts is longer than the default timers for MAC entries on the leaf and spine switches. With an endpoint retention policy defined, you can either tune the timers to last longer than the ARP cache on the servers, or if you have defined a subnet IP address and IP routing on the bridge domain, Cisco ACI will send ARP requests to for the hosts before the timer has expired in which case the tuning may not be required.

**Warning:** The change from unknown unicast flooding mode to hardware proxy mode is disruptive to the traffic in the bridge domain.

**ARP Flooding**

If ARP flooding is disabled, a Layer 3 lookup occurs for the target IP address of the ARP packet. ARP behaves like a Layer 3 unicast packet until it reaches the destination leaf switch.

ARP flooding is required when you need Gratuitous ARP (GARP) requests to update host ARP caches or router ARP caches. This is the case when an IP address may have a different MAC address (for example, with clustering of failover of load balancers and firewalls).

You should enable ARP flooding and also GARP-based detection.

**Note:** GARP-based detection helps in a variety of scenarios in both first- and second-generation Cisco ACI leaf switches. In first-generation switches, this option was useful primarily when a host connected to a Cisco ACI leaf through an intermediate switch changed the MAC address for the same IP address: for instance, because of a floating IP address. In second-generation Cisco ACI leaf switches, this option is still useful primarily if you need to deselect data-plane learning. Do not change the data-plane learning configuration except for service graph redirect configurations. For more information, refer to the "Data-Plane Learning" section of this document.
Endpoint Learning Considerations

Endpoint learning in the mapping database can be used by Cisco ACI to optimize traffic forwarding (in the case of Layer 2 entries), to implement routing of the traffic (for Layer 3 entries), and to perform advanced troubleshooting for applications (for example, using iTrace route, the troubleshooting wizard, and the endpoint tracker).

If routing is disabled under the bridge domain:

- Cisco ACI learns the MAC addresses of the endpoints in the mapping database.
- Cisco ACI floods ARP requests (regardless of whether ARP flooding is selected).

If routing is enabled under bridge domain:

- Cisco ACI learns MAC addresses for Layer 2 traffic in the mapping database (this happens with or without ip routing).
- Cisco ACI learns MAC and IP addresses for Layer 3 traffic in the mapping database.

In the bridge domain, you configure a subnet to define a default gateway for servers. See the section “Default Gateway (Subnet) Design Considerations.”

Make sure to use the Limit IP Learning to Subnet option to help ensure that only endpoints that belong to the bridge domain subnet are learned.

**Warning:** At the time of this writing, enabling Limit IP Learning to Subnet is disruptive to the traffic in the bridge domain.

Endpoint Aging

If no activity occurs on an endpoint, the endpoint information is aged out dynamically based on the setting on an idle timer. If no activity is detected from a local host after 75 percent of the idle timer value has elapsed, the fabric checks whether the endpoint is still alive by sending a probe to it.

If the endpoint does not actively send traffic for the configured idle time interval, a notification is sent to the mapping database using COOP to indicate that the endpoint should be deleted from the database.

For Cisco ACI to be able to maintain an updated table of endpoints, it is preferable to have the endpoints learned using the IP address (that is, they are not just considered to be Layer 2 hosts) and to have a subnet configured under a bridge domain.

A bridge domain can learn endpoint information with IP routing enabled and without any subnet. However, if a subnet is configured, the bridge domain can send an ARP request for the endpoint whose endpoint retention policy is about to expire, to see if it is still connected to the fabric.

If you are using the hardware-proxy option, always define the endpoint retention policy in one of these two ways:

- If the bridge domain is not configured with a subnet IP address and if IP routing is disabled, make sure the endpoint retention policy is defined with a timer that is longer than the ARP cache of the servers. If the endpoint retention policy is too aggressive, upon expiration of the MAC entry in the mapping database, the spine will drop the packet, even if the leaf nodes send traffic to the spines for an unknown MAC unicast destination.
- If the bridge domain has a subnet IP address and if IP routing enabled, the endpoint retention policy configuration makes sure that Cisco ACI sends ARP requests for the host before the entry expires. This updates the mapping database for both the MAC address and the IP address of the endpoint.
If there are multiple IP addresses for the same MAC address as in the case of a device that performs NAT, and if you want to age out the IP addresses individually, you need to enable the IP Aging option under Fabric>Access Policies>Global Policies>IP Aging Policy.

**Floating IP Address Considerations**

In some deployments, an IP address may exist on multiple servers and, as a result, be associated with multiple MAC addresses. For example:

- For a firewall or load-balancer high-availability pair: Upon failover, the firewall IP address may change the MAC address and move to a different port. The firewall or load-balancing device may send a gratuitous ARP request to update the ARP caches of all other hosts in the same bridge domain. This requires ARP flooding to be enabled in the bridge domain.
- NIC teaming and bonding: For this, see the section “NIC Teaming Design Considerations.”
- In the case of clustering, an IP address may move from one server to another, thus changing the MAC address and announcing the new mapping with a GARP request. This notification must be received by all hosts that had the IP request cached in their ARP tables.

As described earlier, Cisco ACI learns the MAC and IP address association for a leaf node, bridge domain, and port number based on data-plane or control-plane traffic.

The design for clustered servers is outside the scope of this document, and it may require configuration changes for data-plane learning.

**Data-Plane Learning**

The data-plane learning option in the bridge domain offers the capability to change the way that Cisco ACI learns endpoint IP addresses and MAC addresses. At the time of this writing this option is designed to be used primarily with the service graph redirect bridge domain.

All other use cases require an understanding of the way that Cisco ACI forwarding works because if you deselect data-plane learning in the bridge domain, Cisco ACI does the following:

- It disables the learning of endpoint IP addresses as a result of routing. That is, the learning of the endpoint IP address is based on ARP, and GARP-based detection would have to be enabled. In first-generation Cisco ACI leaf switches, this selection also affects the learning of endpoints that send traffic encapsulated in VXLAN to the leaf (in the case of the use of AVS, for instance).
- It disables the learning of the IP-to-VTEP mapping, which means that traffic from a remote leaf would be looked up in the spine. Policy enforcement would occur on the egress leaf. As a result, multicast routing won’t work.
- It disables the learning of the MAC-to-VTEP mapping in remote leaf switches, so you would have to configure hardware proxy mode to avoid flooding all Layer 2 traffic in this bridge domain. Policy enforcement would occur on the egress leaf.

For these reasons, if the bridge domain was previously configured with data-plane learning and this option is changed later, the administrator has to clear the remote entries in all the leaf switches in which this bridge domain is present. A better approach would be to create a new bridge domain and configure it with data-plane learning disabled from the beginning.

At the time of this writing, the recommended approach is to not configure a bridge domain with data-plane learning deselected except for the specific case of service graph redirection.
NIC Teaming Design Considerations

No virtualized servers can be connected to the Cisco ACI fabric through NIC teaming in several ways. The most commonly used NIC teaming options are:

- Active-standby NIC teaming
- vPC

This section describes the design considerations for the first two options. When you use vPC, no special tuning is required on the bridge domain because the vPC interface is logically equivalent to a single interface.

With active-standby NIC teaming, one interface is active and one or more is in a standby state. There are different implementations of the failover process depending on the bonding implementation:

- The MAC address of the active interface stays identical after a failover, so there is no need to remap the IP address of the server to a new MAC address.
- When a failover happens, the newly active interface uses its own MAC address to send traffic. In this case, the IP-to-MAC mapping must be updated on all the servers in the same Layer 2 domain. Therefore, with this type of implementation, the server sends a GARP request after a failover.

With the first implementation, the bridge domain configuration does not require any change if the newly active interface starts sending traffic immediately after the failover. The MAC-to-VTEP mapping is automatically updated in the mapping database, and as a result, the IP-to-VTEP mapping is updated, so everything works correctly.

With the second implementation, the bridge domain must be configured for ARP flooding in order for the GARP request to reach the servers in the bridge domain. The GARP packet also triggers an update in the mapping database for the IP-to-MAC mapping and IP-to-VTEP mapping, regardless of whether ARP flooding is enabled.

Changing Bridge Domain Settings in a Production Network

When changing bridge domain settings in a production network, use caution because endpoints that had been learned in the mapping database may be then flushed after the change. This is because in the current implementation the VNID used by the same Bridge Domain configured for unknown unicast flooding or for hardware-proxy is different.

When changing bridge domain settings from unknown unicast flooding to hardware-proxy mode, use caution because hardware-proxy mode relies on the mapping database for forwarding unknown unicast MAC addresses.

Consider an example where the hosts in the bridge domain are not sending traffic periodically but only when an event requires it: for example, a VMware kernel (VMkernel) interface for vMotion. Such hosts may have a cached ARP entry. If you change the bridge domain settings to hardware proxy and the ARP entry on the hosts doesn't expire immediately afterward, when the host tries to send traffic to another host, that host will effectively be generating unknown unicast MAC address traffic.

This traffic in hardware-proxy mode is not flooded, but sent to the spine proxy. The spine proxy does not have an updated mapping database unless the destination host has spoken after you changed the bridge domain settings. As a result, this traffic will be dropped.

Similar considerations apply if you change the Limit IP Learning to Subnet configuration in a bridge domain.

Because of this, if you change the settings of a bridge domain, you should help ensure that either the hosts in the bridge domain are not silent or that their ARP caches are refreshed after the change.
Summary of Bridge Domain Recommendations

The recommended Bridge Domain configuration that works in most scenarios consists of the following settings:

- For Layer 2 domains consisting of endpoints directly connected to the Cisco ACI fabric, the bridge domain should be configured for hardware proxy. This setting not only reduces the flooding due to Layer 2 unknown unicast, but it is also more scalable because the fabric uses more the spine-proxy table capacity instead of just relying on the hardware tables on the individual leaf switches.

- For bridge domains connected to existing Layer 2 networks, you should configure the bridge domain for unknown unicast flooding and select the Clear Remote MAC Entries option.

- Use ARP flooding with GARP-based detection enabled. Because of the variety of teaming implementations and the potential presence of floating IP addresses, ARP flooding often is required.

- You may consider enabling IP routing for the reasons previously described related to the need to maintain an up-to-date mapping database. Make sure that learning of endpoint IP addresses in a Layer 2 bridge domain does not create an unwanted direct path to this bridge domain and the number of endpoints learned per leaf and per port is compatible with the verified scalability limits.

- Configure a subnet and select the Limit IP Learning to Subnet option.

- Configure endpoint retention policy. This configuration helps ensure that the forwarding tables are up-to-date even if the ARP cache on the servers has expired.

- If you have devices that forward multiple IP addresses with the same MAC address and you want to age the IP addresses individually, configure Fabric>Access Policies>Global Policies>IP Aging Policy to enable IP aging.

You should use caution if you change the bridge domain configuration from unknown unicast flooding to hardware proxy for the reasons described previously: the ARP cache on the servers may take long to expire, and in the presence of silent hosts the forwarding tables may not yet be populated.

Default Gateway (Subnet) Design Considerations

Pervasive Gateway

The Cisco ACI fabric operates as an anycast gateway for the IP address defined in the bridge domain subnet configuration. This is known as a pervasive gateway.

The pervasive SVI is configured on a leaf switch wherever the bridge domain of the tenant is present.

Virtual Versus Custom MAC Address for the SVI with Fabrics Connected Using Layer 2 Extension

The bridge domain lets you configure two different MAC addresses for the subnet:

- Custom MAC address
- Virtual MAC address

The primary use case for this feature is related to Layer 2 extension of a bridge domain if you connect two fabrics at Layer 2 in order for each fabric to have a different custom MAC address.
**Note:** The reason for this is the following: Imagine that there are two bridge domains, BD1 and BD2, both present in Fabric 1 and Fabric 2. Imagine that these bridge domains are extended between these two fabrics through some Layer 2 extension technology (for example, EoMPLS). Imagine that a host in Fabric 1 on BD1 sends traffic to Host 2 in Fabric 2 on BD2. If the endpoint information of Host 2 has not yet been learned by Fabric 1, when Host 1 sends a packet, Fabric 1 performs gleaning of the IP address of Host 2. This ARP request is generated by a leaf of Fabric 1, it is transported through Layer 2 (EoMPLS) to Fabric 2, and it carries as a source MAC address the MAC address of BD2. If the MAC address of BD2 is identical in Fabric 1 and Fabric 2, the ARP reply from Host 2 reaches Fabric 2, but it never makes it to Fabric 1. If instead each fabric has a unique MAC address for each bridge domain, the reply to the ARP request is forwarded to the leaf of Fabric1 that connects to the leaf of Fabric 2.

If you configure a unique custom MAC address per fabric, you will also want to configure a virtual MAC address that is identical in both fabrics to help ensure a transparent vMotion experience.

When the fabric sends an ARP request from a pervasive SVI, it uses the custom MAC address and the physical IP address of the SVI.

When the server sends ARP requests for its default gateway (the virtual IP address for the subnet), the MAC address that it gets in the ARP response is the virtual MAC address.

**Note:** In the Cisco Nexus 93128TX, 9372PX and TX, and 9396PX and TX platforms, when the virtual MAC address is configured, traffic is routed only if it is sent to the virtual MAC address. If a server chooses to send traffic to the custom MAC address, this traffic cannot be routed.

**Virtual Versus Primary IP address for the SVI with Fabrics Connected Using Layer 2 Extension**

When interconnecting a pair of Cisco ACI fabrics at Layer 2, it is often a requirement to provision the same pervasive gateway on both fabrics, to be used as the default gateway for hosts. This capability is known in Cisco ACI as a common pervasive gateway.

When the fabric sends an ARP request from a pervasive SVI, it uses the primary, non-virtual IP address as the source IP address. Servers should be configured to use the virtual IP address as the default gateway. The server would then receive the virtual MAC address as the gateway MAC address.

**Note:** For the Cisco Nexus 93128TX, 9372PX and TX, and 9396PX and TX platforms, when the virtual MAC address is configured, traffic is routed only if it is sent to the virtual MAC address. If a server chooses to send traffic to the custom MAC address, this traffic cannot be routed.

**Subnet Configuration: Under Bridge Domain or EPG**

When connecting servers to Cisco ACI, you should set the servers’ default gateway as the subnet IP address of the bridge domain.

Cisco ACI also lets you enter the subnet IP address at the EPG level for designs that require VRF leaking. In Cisco ACI releases earlier than Release 2.3, the subnet defined under an EPG that is the provider of shared services had to be used as the default gateway for the servers. Starting with Cisco ACI Release 2.3, you can enter a subnet under the EPG for cross-VRF filtering purposes only by selecting the No Default SVI Gateway option.
The differences between a subnet under the bridge domain and a subnet under the EPG are as follows:

- **Subnet under the bridge domain**: If you do not plan any route leaking among VRF instances and tenants, the subnets should be placed only under the bridge domain. If Cisco ACI provides the default gateway function, the IP address of the SVI providing the default gateway function should be entered under the bridge domain.

- **Subnet under the EPG**: If you plan to make servers on a given EPG accessible from other tenants (such as in the case of shared services), you must configure the provider-side subnet also at the EPG level. This is because a contract will then also place a route for this subnet in the respective VRF instances that consume this EPG. If you configure subnets at the EPG level, they must be non-overlapping, which means that if EPG1 is configured for one subnet, EPG2 of the same VRF must be configured with a non-overlapping subnet. At the time of this writing, when you define a subnet under an EPG that differs from the subnet defined under the bridge domain (for instance, if the prefix is longer than that of the bridge domain subnet), the default gateway for the servers that belong to this EPG should be configured with the subnet used under the bridge domain. Make sure that the No Default SVI Gateway option is selected under the EPG subnet.

**Note**: There are more design choices related to VRF leaking. These are fully discussed in the section “Designing for Multitenancy and Shared Services.”

Subnets can have these properties:

- **Advertised Externally**: This option indicates that this subnet should be advertised to an external router by a border leaf (through an L3Out connection).
- **Private to VRF**: This option indicates that this subnet is contained within the Cisco ACI fabric and is not advertised to external routers by the border leaf.
- **Shared Between VRF Instances**: This option is for shared services. It is used to indicate that this subnet should be leaked to one or more VRFs. The shared subnet attribute is applicable to both public and private subnets.

**Summary of Best Practices for Bridge Domain Subnet Configuration**

The following list summarizes the best practices for bridge domain subnet configuration:

- Define one subnet as primary.
- Do not configure more than one MAC address unless you need to for the purposes of Layer 2 extension.
- The subnet should always be configured at the bridge domain level.
- If you are entering subnet information under EPGs that provide shared services for cross-VRF contracts, make sure to select the No Default SVI Gateway option under the EPG subnet.

**EPG Design Considerations**

Traffic from endpoints is classified and grouped into EPGs based on various configurable criteria, some of which are hardware dependent, and some of which are software dependent.

Cisco ACI can classify three types of endpoints:

- Physical endpoints
- Virtual endpoints
- External endpoints (endpoints that send traffic to the Cisco ACI fabric from the outside)
You must distinguish how the hardware classifies traffic and how the transport on the wire keeps traffic.

The administrator may configure classification based on virtual machine attributes, and depending on the combination of software and hardware, that may translate into a VLAN-based classification or MAC-based classification.

Hardware (depending on the ASIC model) can classify traffic as follows:

- Based on VLAN or VXLAN encapsulation
- Based on port and VLAN or port and VXLAN
- Based on network and mask or IP address for traffic originated outside the fabric: that is, traffic that is considered as part of the Layer 3 external traffic
- Based on source IP address or subnet (with Cisco Nexus E platform leaf nodes, Cisco Nexus 9300-EX or Cisco 9300-FX platform switches)
- Based on source MAC address (with Cisco Nexus 9300-EX or Cisco 9300-FX platform switches)

From the administrator perspective, it is possible to configure classification of the incoming traffic to the leaf as follows:

- Based on VLAN encapsulation
- Based on port and VLAN
- Based on network and mask or IP address for traffic originating outside the fabric: that is, traffic that is considered as part of the Layer 3 external traffic
- Based on explicit virtual NIC (vNIC) assignment to a port group: At the hardware level, this translates into a classification based on a dynamic VLAN or VXLAN negotiated between Cisco ACI and the VMM.
- Based on source IP address or subnet: For virtual machines, this function does not require any specific hardware if you are using AVS. For physical machines, this function requires the hardware to support source IP address classification (Cisco Nexus E platform leaf nodes and later platforms).
- Based on source MAC address: For virtual machines, this function does not require specific hardware if you are using AVS. For physical machines, this requires the hardware to support MAC-based classification and ACI version 2.1 or higher.
- Based on virtual machine attributes: This option assigns virtual machines to an EPG based on attributes associated with the virtual machine. At the hardware level, this translates into a classification based on VLAN or VXLAN (if using AVS software on the virtualized host or, more generally, if using software that supports OpFlex protocol on the virtualized host) or based on MAC addresses (Cisco Nexus 9000 EX platform with VMware vDS).

You can assign a workload to an EPG as follows:

- Map an EPG statically to a port and VLAN.
- Map an EPG statically to a VLAN switchwide on a leaf.
- Map an EPG to a VMM domain (followed by the assignment of vNICs to the associated portgroup).
- Map a base EPG to a VMM domain and create microsegments based on virtual machine attributes (followed by the assignment of vNICs to the base EPG).
**Note:** If you configure EPG mapping to a VLAN switchwide (using a static leaf binding), Cisco ACI configures all leaf ports as Layer 2 ports. If you then need to configure an L3Out connection on this same leaf, these ports cannot then be configured as Layer 3 ports. This means that if a leaf is both a computing leaf and a border leaf, you should use EPG mapping to a **port** and **VLAN**, not switchwide to a VLAN.

**VLANs and Cisco ACI**
Cisco ACI is designed so that the user can focus on the logical or abstracted design in terms of distributed bridge domains, distributed VRFs, and so on, and not worry about maintaining VLAN mapping information.

Networking equipment managed by Cisco ACI uses VXLAN tunnels internally to the fabric, and can use VLANs or VXLANs to communicate with external devices. The VLANs used to communicate with servers or switches have local significance on the leaf. VLANs are simply a tag on the wire to segment traffic.

**Connecting EPGs to External Switches**
If two external switches are connected to two different EPGs within the fabric, you must ensure that those external switches are not directly connected outside the fabric. It is strongly recommended in this case that you enable BPDU guard on the access ports of the external switches to help ensure that any accidental direct physical connections are blocked immediately.

Consider Figure 54 as an example.

**Figure 54.** Switches Connected to Different EPGs in the Same Bridge Domain

In this example, VLANs 10 and 20 from the outside network are stitched together by the Cisco ACI fabric. The Cisco ACI fabric provides Layer 2 bridging for traffic between these two VLANs. These VLANs are in the same flooding domain. From the perspective of the Spanning Tree Protocol, the Cisco ACI fabric floods the BPDU s within the EPG (within the same VLAN ID). When the Cisco ACI leaf receives the BPDU s on EPG 1, it floods them to all leaf ports in EPG 1, and it does not send the BPDU frames to ports in other EPGs. As a result, this flooding behavior can break the potential loop within the EPG (VLAN 10 and VLAN 20). You should ensure that VLANs 10 and 20 do not have any physical connections other than the one provided by the Cisco ACI fabric. Be sure to turn on the BPDU guard feature on the access ports of the outside switches. By doing so, you help ensure that if someone mistakenly connects the outside switches to each other, BPDU guard can disable the port and break the loop.
Working with Multiple Spanning Tree

Additional configuration is required to help ensure that Multiple Spanning Tree (MST) BPDU frames flood properly. BPDU frames for Per-VLAN Spanning Tree (PVST) and Rapid Per-VLAN Spanning Tree (RPVST) have a VLAN tag. The Cisco ACI leaf can identify the EPG on which the BPDU frames need to be flooded based on the VLAN tag in the frame.

However, for MST (IEEE 802.1s), BPDU frames don’t carry a VLAN tag, and the BPDU frames are sent over the native VLAN. Typically, the native VLAN is not used to carry data traffic, and the native VLAN may not be configured for data traffic on the Cisco ACI fabric. As a result, to help ensure that MST BPDU frames are flooded to the desired ports, the user must create an EPG (an MST EPG) for VLAN 1 as native VLAN to carry the BPDU frames. This EPG connects to the external switches that run MST.

In addition, the administrator must configure the mapping of MST instances to VLANs to define which MAC address table must be flushed when a topology change notification (TCN) occurs. When a TCN event occurs on the external Layer 2 network, this TCN reaches the leafs where it connects to via the MST EPG and it flushes the local endpoint information associated with these VLANs on these leafs and as result these entries are removed from the spine-proxy mapping database.

Configuring Trunk and Access Ports

In Cisco ACI, you can configure ports that are used by EPGs in one of these ways:

- Trunk or tagged (classic IEEE 802.1q trunk): Traffic for the EPG is sourced by the leaf with the specified VLAN tag. The leaf also expects to receive traffic tagged with that VLAN to be able to associate it with the EPG. Traffic received untagged is discarded.
- Access (untagged): Traffic for the EPG is sourced by the leaf as untagged. Traffic received by the leaf as untagged or with the tag specified during the static binding configuration is associated with the EPG.
- Access (IEEE 802.1p): If only one EPG is bound to that interface, the behavior is identical to that in the untagged case. If other EPGs are associated with the same interface, then traffic for the EPG is sourced with an IEEE 802.1q tag using VLAN 0 (IEEE 802.1p tag).

If you are using Cisco Nexus 9300-EX or Cisco 9300-FX platform switches, you can have different interfaces on the same leaf bound to a given EPG in both the trunk and access (untagged) modes at the same time. This configuration was not possible with previous-generation leaf switches. Therefore, for first-generation leaf switches it used to be a good practice to select the Access (IEEE 802.1p) option to connect an EPG to a bare-metal host because that option allowed access and trunk ports in the same EPG.

Using the access (IEEE 802.1p) EPG binding for access ports also works for most servers, but this setting sometimes is incompatible with hosts using the pre-boot execution environment (PXE) and non-x86 hosts. This is the case because traffic from the leaf to the host may be carrying a VLAN tag of 0. Whether or not an EPG with access ports configured for access (IEEE 802.1p) has a VLAN tag of 0 depends on the configuration.

If a port on a leaf node is configured with multiple EPGs, where one of those EPGs is in access (IEEE 802.1p) mode and the others are in trunk mode, the behavior differs depending on the switch hardware in use:

- If using Cisco Nexus 93180YC-EX or 93108TC-EX Switches, traffic from the EPG in IEEE 802.1p mode will exit the port untagged.
- If using other than Cisco Nexus 9000 EX platform switches, traffic from the EPG in IEEE 802.1p mode will exit the port tagged as VLAN 0.
In summary, if you are using a Cisco Nexus 9300-EX or Cisco 9300-FX platform switches as a leaf, you should configure access ports with static binding of type Access (untagged), and you can have a mix of access (untagged) and trunk (tagged) ports in the same EPG.

If you are using first-generation leaf switches, you can have EPGs with both access and trunk ports by configuring access ports as type Access (IEEE 802.1p).

You can also define an EPG binding to a VLAN on a leaf without specifying a port. This option is convenient, but it has the disadvantage that if the same leaf is also a border leaf, you cannot configure Layer 3 interfaces because this option changes all the leaf ports into trunks, so if you have an L3Out connection, you will then have to use SVI interfaces.

**EPG-to-VLAN Mapping**

In general, VLANs in Cisco ACI have local significance on a leaf switch. If per-port VLAN significance is required, you must configure a physical domain that is associated with a Layer 2 interface policy that sets the VLAN scope to port local.

The rules of EPG-to-VLAN mapping with a VLAN scope set to global are as follows:

- You can map an EPG to a VLAN that is not yet mapped to another EPG on that leaf.
- Regardless of whether two EPGs belong to the same or different bridge domains, on a single leaf you cannot reuse the same VLAN used on a port for two different EPGs.
- The same VLAN number can be used by one EPG on one leaf and by another EPG on a different leaf. If the two EPGs are in the same bridge domain, they share the same flood domain VLAN for BPDUs and they share the broadcast domain.

The rules of EPG-to-VLAN mapping with the VLAN scope set to local are as follows:

- You can map two EPGs of different bridge domains to the same VLAN on different ports of the same leaf, if the two ports are configured for different physical domains.
- You cannot map two EPGs of the same bridge domain to the same VLAN on different ports of the same leaf.

Figure 55 illustrates these points.
EPG and VLAN Recommendations

The recommendation is to use unique VLANs per EPG within a bridge domain and across leaf nodes, to be able to scope flooding and BPDUs within the EPG if so desired.

Within an EPG, another recommendation is to use VLANs for connectivity to a switched network other than the VLAN used for endpoints directly attached to the leaf of the fabric. This approach limits the impact of spanning-tree TCN events only to clearing the endpoints learned on the switched network.

EPG and Endpoint Learning

You can verify the endpoint learning in Cisco ACI by viewing the Client Endpoints field on the EPG Operational tab.

The learning source field will typically display one (or both) of the following learning source types:

A learning source of vmm is relevant for the purposes of Resolution and Deployment Immediacy settings in the presence of virtualized hosts. This learning source type indicates that the VMM and APIC have resolved to which leaf node and port a virtual machine is attached by correlating Cisco Discovery Protocol and LLDP or by using the OpFlex protocol.

- A learning source of learned indicates that the endpoint has been learned through the data plane and exists in the mapping database.

You can find these values next to each endpoint MAC and IP address:

- vmm: This value is learned from a VMM such as vCenter or SCVMM. This is not an indication of an entry learned through the data plane, but instead indicates that vCenter or SCVMM, etc. have communicated to the APIC the location of the virtual machine endpoint, and depending on the Resolution and Deployment Immediacy settings that you configured, this may have triggered the instantiation of the VRF, bridge domain, EPG, and contract on the leaf where this virtual machine is active.
- vmm, learn: This means that both the VMM and the data plane (both real data plane and ARP) provided this entry information.
• **learn**: The information is from ARP spoofing or data-plane forwarding.
• **static**: The information is manually entered.
• **static, learn**: The information is manually entered, plus the entry is learned in the data plane.

**Contract Design Considerations**

A contract is a policy construct used to define communication between EPGs. Without a contract between EPGs, no communication is possible between those EPGs (unless the VRF instance is configured as "unenforced"). Within an EPG, a contract is not required to allow communication (although communication can be prevented with microsegmentation features). Figure 56 shows the relationship between EPGs and contracts.

**Figure 56.** EPGs and Contracts

An EPG provides or consumes a contract (or provides **and** consumes a contract). For example, the Web EPG in the example in Figure 56 provides a contract that the App EPG consumes. Similarly, the App EPG provides separate contracts, which are consumed by the Web and DB EPGs.

The use of contracts in Cisco ACI has the following goals:

- Define an ACL to allow communications between security zones.
- Define route leaking between VRFs or tenants.

Figure 57 shows how contracts are configured between EPGs: for instance, between internal EPGs and external EPGs.

**Figure 57.** Contracts Between Internal and External EPGs

**Security Contracts Are ACLs Without IP Addresses**

You can think of security contracts as ACLs between EPGs. As Figure 58 illustrates, the forwarding between endpoints is based on routing and switching as defined by the configuration of VRF instances and bridge domains. Whether the endpoints in the EPGs can communicate depends on the filtering rules defined by contracts.
Figure 58. Contracts Are ACLs

Note: Contracts can also control more than just the filtering. If contracts are used between EPGs in different VRF instances, they are also used to define the VRF route-leaking configuration.

Filters and Subjects
A filter is a rule specifying fields such as the TCP port and protocol type, and it is referenced within a contract to define the communication allowed between EPGs in the fabric.

A filter contains one or more filter entries that specify the rule. The example in Figure 59 shows how filters and filter entries are configured in the APIC GUI.

Figure 59. Filters and Filter Entries

A subject is a construct contained within a contract and typically references a filter. For example, the contract Web might contain a subject named Web-Subj that references a filter named Web-Filter.
Concept of Direction in Contracts
As you can see from the previous section, filter rules have a direction, similar to ACLs in a traditional router. ACLs are normally applied to router interfaces. In the case of Cisco ACI, the ACLs (contract filters) differ from classic ACLs in the following ways:

- The interface to which they are applied is the connection line of two EPGs.
- The directions in which filters are applied are the consumer-to-provider and the provider-to-consumer directions.
- The ACLs do not include IP addresses because traffic is filtered based on EPG (or source group or class-ID, which are synonymous).

Understanding the Bidirectional and Reverse Filter Options
When you create a contract, two options are typically selected by default:

- Apply Both Directions
- Reverse Filter Ports

The Reverse Filter Ports option is available only if the Apply Both Directions option is selected (Figure 60).

Figure 60. Apply Both Directions and Reverse Filter Ports Option Combinations

An example clarifies the meaning of these options. If you require client-EPG (the consumer) to consume web services from port 80 on server-EPG (the provider), you must create a contract that allows source Layer 4 port “any” (“unspecified” in Cisco ACI terminology) to talk to destination Layer 4 port 80. You must then consume the contract from the client EPG and provide the same contract from the server EPG (Figure 61).

Figure 61. The Filter Chain of a Contract Is Defined in the Consumer-to-Provider Direction

The effect of enabling the Apply Both Directions option is to program two TCAM entries: one that allows source port “unspecified” to talk to destination port 80 in the consumer-to-provider direction, and one for the provider-to-consumer direction that allows source port “unspecified” to talk to destination port 80 (Figure 62).
Figure 62. Apply Both Directions Option and the Filter Chain

As you can see, this configuration is not useful because the provider (server) would generate traffic from port 80 and not to port 80.

If you enable the option Reverse Filter Ports, Cisco ACI reverses the source and destination ports on the second TCAM entry, thus installing an entry that allows traffic from the provider to the consumer from Layer 4 port 80 to destination port "unspecified" (Figure 63).

Figure 63. Apply Both Directions and Reverse Filter Ports Options

Cisco ACI by default selects both options: Apply Both Directions and Reverse Filter Ports.

Configuring a Single Contract Between EPGs

An alternative method for configuring filtering rules on a contract is to manually create filters in both directions: consumer to provider and provider to consumer.

With this configuration approach, you do not use Apply Both Directions or Reverse Filter Ports, as you can see in Figure 64.
Figure 64. Configuring Contract Filters at the Subject Level

The configuration of the contract in this case consists of entering filter rules for each direction of the contract. Figure 65 provides a graphical representation of the contract and the interface between consumer and provider.

Figure 65. Configuring Filters for Consumer-to-Provider and Provider-to-Consumer Directions

Note: As you can see from this example, more than one contract between any two EPGs is not generally required. Instead, edit one contract and enter additional rules as needed.
Contract Scope
The scope of a contract defines the EPGs to which the contract can be applied:

- **VRF**: EPGs associated with the same VRF instance can use this contract.
- **Application profile**: EPGs in the same application profile can use this contract.
- **Tenant**: EPGs in the same tenant can use this contract even if the EPGs are in different VRFs.
- **Global**: EPGs throughout the fabric can use this contract.

Contracts and Filters in the Common Tenant
In Cisco ACI, the Common tenant provides resources that are visible and can be used from other tenants. This may mean not having to define the same filter multiple times for the same protocol.

Although it is convenient to use filters from the Common tenant, it is not necessarily a good idea to use contracts from the Common tenant:

- **On reason** is that the name used for contracts in tenant Common should be unique across all tenants. If a tenant is using a contract called for instance “from A to B” from tenant Common (common/from A to B), and you define a new contract with the same name inside of the tenant itself (mytenant/from A to B), ACI will change the EPG relations that were previously associated with common/from A to B to be associated to the locally defined contract mytenant/from A to B.
- **Another reason** is that if multiple tenants provide and consume the same contract from the Common tenant, you are effectively establishing a relationship with other tenants.

For instance, imagine that in the Common tenant you have a contract called web-to-app and you want to use it in Tenant A and in Tenant B to allow EPG-web of Tenant A to talk to EPG-app of Tenant A. Imagine that you also want to allow EPG-web of Tenant B to talk to EPG-app of Tenant B with the same contract. If you use the contract in this way, you are likely to also enable EPG-web of Tenant A to talk to EPG-app of Tenant B.

This is by design because you are telling Cisco ACI that EPGs in both tenants are providing and consuming the same contract.

To implement a design where the web EPG talks to the application EPG of its own tenant, you should configure the contract web-to-app in each individual tenant.

You could also define contracts from the Common tenant to implement the same design. To do this, verify that you set the scope of the contract correctly at the time of creation. For example, set the contract scope in the Common tenant to Tenant. Cisco ACI will then scope the contract to each tenant where it would be used as if the contract had been defined in the individual tenant.

Configuring an Unenforced VRF Instance and Preferred Groups
In certain deployments, all EPGs associated with a VRF instance may need to be able to communicate freely. In this case, you can configure the VRF instance with which they are associated as “unenforced.”

You can also use a VRF instance as "enforced," but you need to organize EPGs into two groups:

- **EPG members of the preferred group**: The endpoints in these EPGs can communicate without contracts even if they are in different EPGs. If one of two endpoints that need to communicate is part of the preferred group and the other is not, a contract is required.
- **EPGs that are not in the preferred group**: These are regular EPGs.
Using vzAny

vzAny, referred to in the Cisco ACI GUI as “EPG collection for context,” is a special object that represents all EPGs associated with a given VRF instance, including the Layer 3 external EPG.

This concept is useful when a configuration has contract rules that are common across all the EPGs under the same VRF. In this case, you can place the rules that are common across the VRF into a contract associated with vzAny.

When using vzAny, you must understand how vzAny interacts with VRF route leaking and with L3Out.

One common use of the vzAny object relates to consumption of the same set of shared services provided by an EPG in a different VRF. vzAny can only be a consumer of shared services, not a provider.

For more details about vzAny restrictions, please refer to this document:


Note When using vzAny with shared services contracts, vzAny is supported only as a shared services consumer, not as a shared services provider.

An additional consideration when using vzAny is the fact that it includes the Layer 3 external connection of the VRF. If vzAny is consuming a contract provided by an EPG within a different VRF, the subnets defined under this EPG may be announced from the L3Out interface. For example, if you have vzAny from VRF1 consuming a contract provided by an EPG from a different VRF (VRF2), the subnets of VRF1 that are marked as public will be announced through the L3Out interface of VRF2.

Summary of Best Practices for Contracts

The following list summarizes best practices for contracts:

- Define contract scope correctly according to requirements.
- Define contracts in individual tenants rather than in tenant Common.
- You can reuse filters defined in tenant Common.
- If you want to reuse "contracts" make sure you understand the scope of the contract as described in the section "Contracts and Filters in the Common Tenant".
- If you are using VRF route leaking, determine which side is the provider and which is the consumer. Depending on this definition, you may have to configure the subnet under the EPG or the bridge domain.
- Consider the number of filter rules and the impact they may have on leaf node hardware entries.
- Use vzAny to reduce the consumption of entries in the TCAM. Be sure that you understand how vzAny interacts with VRF leaking and how it influences L3Out routing announcements.

Resolution and Deployment Immediacy of VRF Instances, Bridge Domains, EPGs, and Contracts

Cisco ACI optimizes the use of hardware and software resources by programming the hardware with VRFs, bridge domains, SVIs, EPGs, and contracts only if endpoints are present on a leaf that is associated with these.

Policy deployment immediacy is configurable for EPGs.
These optimizations are configurable through two options:

- **Resolution Immediacy**: This option controls when VRF, bridge domains, and SVIs are pushed to the leaf nodes.
- **Deployment Immediacy**: This option controls when contracts are programmed in the hardware.

Resolution and Deployment Immediacy are configuration options that are configured when an EPG is associated with a domain. A domain represents either a VMM vDS for a given datacenter or a set of VLANs mapped to a set of leaf switches and associated ports.

The options for Resolution Immediacy (that is, for programming of the VRF, bridge domain, and SVI) are as follows:

- **Pre-Provision**: This option means that the VRF, bridge domain, SVI, and EPG VLAN mappings are configured on the leaf nodes based on where the domain (or to be more precise, the attachable access entity profile) is mapped within the fabric access configuration. If EPG1 is associated with VMM domain1, the bridge domain and the VRF to which EPG1 refers are instantiated on all the leaf nodes where the VMM domain is configured. If EPG2 is also associated with VMM domain1, the bridge domain and VRF that EPG2 refers to are also instantiated on all the leaf nodes where this VMM domain is configured.
- **Immediate**: This option means that the VRF, bridge domain, SVI, and EPG VLAN mappings are configured on a leaf as soon as a hypervisor connected to this leaf is attached to an APIC VMM virtual switch. A discovery protocol such as Cisco Discovery Protocol and LLDP (or the OpFlex protocol for AVS) is used to form the adjacency and discover to which leaf the virtualized host is attached. If EPGs 1 and 2 are associated with VMM domain1, the bridge domains and the VRFs to which these EPGs refer are instantiated on all leaf nodes where Cisco ACI leaf nodes have discovered the host.
- **On-Demand**: This option means that the VRF, bridge domain, SVI, and EPG VLAN mappings are configured on a leaf switch only when a hypervisor connected to this leaf is connected to a virtual switch managed by the APIC and at least one virtual machine on the host is connected to a portgroup and EPG that is associated with this vnic and leaf. If a virtual machine vNIC is associated with an EPG whose vnic is connected to leaf1, only the VRF, bridge domain and EPG VLAN related to this EPG are instantiated on that leaf.

The options for Deployment Immediacy (that is, for programming of the policy CAM) are as follows:

- **Immediate**: The policy CAM is programmed on the leaf as soon as the policy is resolved to the leaf (see the discussion of resolution immediacy) regardless of whether the virtual machine on the virtualized host has sent traffic.
- **On-Demand**: The policy CAM is programmed as soon as first data-plane packet reaches the switch.
Table 3 illustrates the configuration options.

Table 3. Resolution and Deployment Immediacy

<table>
<thead>
<tr>
<th>Resolution</th>
<th>Pre-Provision</th>
<th>Immediate</th>
<th>Immediate</th>
<th>Immediate</th>
<th>Immediate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deploymen t</td>
<td>On-Demand</td>
<td></td>
<td>On-Demand</td>
<td></td>
<td>On-Demand</td>
</tr>
<tr>
<td>Hardware resource</td>
<td>VRF, bridge domain, and SVI</td>
<td>Policy CAM</td>
<td>VRF, bridge domain, and SVI</td>
<td>Policy CAM</td>
<td>VRF, bridge domain, and SVI</td>
</tr>
<tr>
<td>Domain associated to EPG</td>
<td>On leaf nodes where AEP and domain are present</td>
<td>On leaf nodes where AEP and domain are present</td>
<td>On leaf nodes where AEP and domain are present</td>
<td>On leaf where host is connected</td>
<td>On leaf where host is connected</td>
</tr>
<tr>
<td>Host discovered on leaf through Cisco Discovery Protocol</td>
<td>Same as above</td>
<td>Same as above</td>
<td>Same as above</td>
<td>Same as above</td>
<td>Same as above</td>
</tr>
<tr>
<td>Virtual machine associated with port group</td>
<td>Same</td>
<td>Same</td>
<td>Same</td>
<td>Same</td>
<td>Same</td>
</tr>
<tr>
<td>Virtual machine sending traffic</td>
<td>Same</td>
<td>On leaf where virtual machine sends traffic</td>
<td>Same</td>
<td>On leaf where virtual machine sends traffic</td>
<td>Same</td>
</tr>
</tbody>
</table>

The use of the On-Demand option optimizes the consumption of VRFs, bridge domains, and TCAM entries.

For example, consider a topology consisting of two leaf nodes (leaf1 and leaf2). A cluster of ESX hosts are connected. Some hosts are attached to leaf1 and some to leaf 2. An EPG, EPG1, is associated with BD1, which in turn is associated with VRF1. A virtual machine, VM1, is attached to EPG1 through a portgroup. If the virtual machine is hosted on a server attached to leaf1 and no virtual machine attached to EPG1 is hosted on servers connected to leaf2, you will see VRF1, BD1, and the contracts rules for EPG1 on leaf1 only, and not on leaf2. On leaf1, you will also see contract rules in the TCAM that are relevant for this EPG. If the virtual machine moves to a server that is connected to leaf2, you will see that VRF1, BD1, and the contract rules for EPG1 are programmed on leaf 2.
**Note:** The On-Demand option is compatible with vMotion migration of Virtual Machines and requires coordination between APIC and the VMM. If all APICs in a cluster are down, vMotion movement of a virtual machine from one virtual host connected one leaf node to another virtual host connected to a different leaf node can occur, but the virtual machine may not have connectivity on the destination leaf: for instance, if a virtual machine moves from a leaf node where the VRF, bridge domain, EPG, and contracts were instantiated to a leaf node where these objects have not yet been pushed. APIC must be informed by the VMM about the move to configure the VRF, bridge domain, and EPG on the destination leaf node.

If no APIC is present due to multiple failures, if the On-Demand option is enabled and if no other virtual machine was already connected to the same EPG on the destination leaf node, the VRF, bridge domain, and EPG cannot be configured on this leaf node. In most deployments, the advantages of On-Demand option for resource optimization outweigh the risk of vMotion movement during the absence of all APICs.

You can choose to use the Pre-Provision option for Resolution Immediacy when you need to help ensure that resources on which the resolution depends are allocated immediately. This setting may be needed, for instance, when the management interface of a virtualized host is connected to the Cisco ACI fabric leaf.


This helps the situation where management traffic for hypervisors/VM controllers are also using the virtual switch associated to APIC VMM domain (VMM switch).

“Deploying a VMM policy such as VLAN on ACI leaf switch requires APIC to collect CDP/LLDP information from both hypervisors via VM controller and ACI leaf switch. However if VM Controller is supposed to use the same VMM policy (VMM switch) to communicate with its hypervisors or even APIC, the CDP/LLDP information for hypervisors can never be collected because the policy required for VM controller/hypervisor management traffic is not deployed yet.”

The microsegmentation feature requires resolution to be immediate.

For all other virtual machines, use of the On-Demand option saves hardware resources.

**Virtual Infrastructure Design Considerations**

The VMM domain provides Cisco ACI connectivity to data center hypervisors. The hypervisors can be Microsoft Hyper-V, VMware ESXi, or OpenStack with Kernel-based Virtual Machine (KVM).

A VMM domain is defined as the virtual machine manager information and the pool of VLANs or multicast addresses for VXLANs that this VMM uses to send traffic to the leaf switches. The VLAN pool should be configured as dynamic, to allow the APIC to allocate VLANs to EPGs and port groups as needed. A VLAN pool can consist of both dynamic and static ranges. This may be required if you need to define a static binding to a specific VLAN used by the same virtualized host that is part of a VMM domain. The VMM domain is associated with an AAEP and a policy group.

This section discusses configurations and design considerations for deployment of Cisco ACI with a virtualized environment and, in particular, with VMware vSphere.
The steps that the APIC uses to integrate with vSphere are as follows:

1. The administrator creates a VMM domain in the APIC with the IP address and credentials for connecting to vCenter.
2. The APIC connects to vCenter and creates a new vDS under vCenter.
3. The vCenter administrator adds the ESX host or hypervisor to the vDS controlled by the APIC and assigns the ESXi host hypervisor ports as uplinks on the vDS. These uplinks must connect to the Cisco ACI leaf switches.
4. The APIC learns to which leaf port the hypervisor host is connected using LLDP or Cisco Discovery Protocol.
5. The tenant administrator creates and associates application EPGs with the VMM domains.
6. The APIC automatically creates port groups in vCenter under the vDS. The EPG is automatically mapped to port groups. This process provisions the network policy in vCenter.
7. The vCenter administrator creates virtual machines and assigns them to port groups. The APIC learns about the virtual machine placements based on the vCenter events.

For this configuration to work, the APIC must be able to reach the vCenter IP address.

If both in-band and OOB paths are available to external servers, APIC chooses the in-band path.

**Design Choices for Virtual Infrastructure Deployments**

**Using Manual (Static) Binding or VMware vDS Managed by Cisco APIC or AVS (Cisco OpFlex Protocol)**

Cisco ACI can be integrated with vSphere using either static bindings or through a VMM domain:

- With static bindings, the VLAN assignment to port groups is static: that is, defined by the administrator.
- When you use a VMM domain, the VLAN allocation is dynamic and maintained by the APIC. The resolution in this case is also dynamic, so the allocation of VRF, bridge domain, EPG, etc. on a leaf is managed by the APIC through the discovery of a virtualized host attached to a leaf port. This dynamic allocation of resources works if one of the following control-plane protocols is in place between the virtualized host and the leaf: Cisco Discovery Protocol, LLDP, or OpFlex protocol.

Cisco ACI can be integrated with vSphere in several ways:

- By manually allocating VLANs to port groups and matching them using static EPG mapping: The traffic encapsulation is based on VLANs in this case, and virtual machine discovery is based on the transmission of traffic by the virtual machines. In this case, the configuration in Cisco ACI is equivalent to having physical hosts attached to the leaf. Therefore, the Cisco ACI fabric configuration is based on the definition of a physical domain.
- By integrating the APIC and vCenter and using the standard vDS: Traffic encapsulation is based on VLANs, and virtual machine discovery is based on a combination of communication between vCenter and the APIC and the use of LLDP or Cisco Discovery Protocol. The use of Cisco Discovery Protocol or LLDP between the ESXi host and the Cisco ACI leaf enables Cisco ACI to resolve the location of an ESXi host, and as a consequence the location where virtual machines that are on that host are connected. When integrating Cisco ACI with vDS, you can use features such as classification based on virtual machine attributes only if you are using this in conjunction with Cisco Nexus 9300 EX platform switches.
- By integrating the APIC and vCenter and using AVS: The traffic encapsulation is based on VLANs or VXLANs in this case, and virtual machine discovery is based on the OpFlex protocol. This option enables the use of virtual machine–based classification with any Cisco Nexus 9000 Series leaf.
Note: The OpFlex protocol is an extensible policy-resolution protocol that can be used to distribute policies to physical or virtual network devices. The OpFlex protocol is used between the Cisco ACI fabric and AVS to provide the configurations to AVS. It is used by AVS to give Cisco ACI information about which endpoints are attached to servers and to AVS.

**Local and No Local (or Fabric Extender) Switching Modes**

When configuring virtualized servers with Cisco ACI, you can choose whether virtual machine-to-virtual machine traffic within the same EPG on the same host is switched locally (that is, within the hypervisor), or switched by the upstream leaf or ToR switch.

The two modes of switching are categorized as follows:

- **Local Switching mode**: Virtual machine-to-virtual machine traffic in the same EPG on the same virtualized server is switched within the server (hypervisor).
- **Non-Local Switching mode (fabric extender [FEX] mode)**: Virtual machine-to-virtual machine traffic within the same EPG on the same virtualized server is switched by the upstream leaf node.

The segmentation and transport mechanism for the two modes varies:

- **Local Switching mode**: Requires the use of VLAN or VXLAN to segment the EPGs
- **Non-Local Switching mode (FEX mode)**: Can be used with AVS in VXLAN mode.

*Figure 66* illustrates these options.

![Local and Non-Local Switching Modes](image)

**Design Considerations for Remotely Attached Hosts**

If you plan to integrate virtualized hosts that connect to the Cisco ACI fabric through intermediate network devices (for instance, in the case of blade switches or the use of Cisco Unified Computing System™ (Cisco UCS), or for migration purposes), and you require dynamic provisioning of policies on the leaf nodes, two options are available:

- If there is only one Layer 2 hop between the ESXi host and the leaf node, you can use vDS (not managed by the APIC) with static binding, vDS managed by Cisco ACI, or AVS.
- If there is more than one hop between the ESXi host and the leaf, vDS (not managed by the APIC) with static bindings or AVS can be used.
**Note:** When using vDS managed by Cisco ACI, you can have only one Layer 2 hop between the leaf and the virtualized host due to the way policy resolution works. Cisco ACI optimizes the allocation of VRFs, bridge domains, and EPGs on a leaf based on control-plane information (Cisco Discovery Protocol color LLDP or the OpFlex protocol). Cisco ACI leaf nodes communicate information to the APIC about which virtualized host is connected to a given port. The APIC then provisions VRFs, bridge domains, and EPGs accordingly. If there is a single layer of switching between the virtualized host and the leaf node, Cisco ACI can still derive the information by correlating the Cisco Discovery Protocol color LLDP information from the virtualized host and vCenter and the Cisco Discovery Protocol color LLDP information from the leaf nodes. With additional Layer 2 hops between the virtualized host and the leaf node, Cisco Discovery Protocol and LLDP will not be able to assist with discovery, and OpFlex protocol is required. OpFlex protocol provides the communication channel between the virtualized host and the leaf node.

Figure 67 illustrates the deployment possibilities.

With AVS, it is possible to deploy virtualized hosts without having to attach those hosts directly to the leaf nodes.

**Figure 67.** Remotely Attached Hosts

When virtualized servers are not directly attached to the leaf nodes, and if you are using VLAN as the segmentation option, you must configure the transit switching infrastructure (such as Cisco UCS fabric interconnects) to trunk all VLANs that are part of the dynamic pool used by the VMM domain.

Using VLAN or VXLAN for Segmentation

You have the choice of using either VLAN or VXLAN for EPG segmentation on virtualized servers. In both cases, the VLAN and VXLAN tags have local significance between the server and the leaf. The traffic is then encapsulated by the leaf into a new VXLAN frame before being forwarded in the fabric.

If using VLAN encapsulation, you must predefine a range of dynamic VLANs for the APIC to orchestrate with the VMM.

If using VXLAN with AVS, you must define a range of multicast addresses, where each multicast address will be used by one EPG. These addresses don't need to be different than the multicast range used by the ACI fabric for VXLAN purposes.

**Note:** In addition to this you need to define a multicast address called "AVS Fabric-Wide Multicast Address" which is also called a "VMM GiPo".

The choice of VLAN or VXLAN depends on several factors. One is simply what vDS or AVS supports. For example, when you use AVS and forwarding with Local Switching mode, you can choose between VLAN or VXLAN encapsulation to segment the EPGs. When you use AVS in FEX mode, you must use VXLAN encapsulation.

You also should use VXLAN as the transport for virtualized servers that are not directly attached, as depicted in Figure 68.

If VXLAN mode is used, Cisco ACI uses a VXLAN overlay over the infrastructure VLAN between the leaf and host for each EPG. This simplifies the deployment, because the only VLAN that needs to be enabled in the transit network is the infrastructure VLAN.
Figure 68. Using VXLAN to Connect Remote Hosts to the Fabric

Using VLAN and VXLAN on the Same AVS Device
At the time of this writing, when you use L4-L7 virtual services you need to use AVS with VLAN-based segmentation. You may also want to limit the number of VLANs to trunk on the links between servers and the leaf, and hence use VXLAN-based segmentation for all traffic that is not for L4-L7 services.

AVS supports a mode of operation that is called AVS mixed encap mode. This mode requires the following configurations:

- Define a VMM domain based on AVS with VXLAN as the encapsulation and also define a VLAN range.
- Configure the VMM domain based on AVS for local switching mode.

If you decide to deploy L4-L7 services in the VMM domain, Cisco ACI will automatically create shadow EPGs based on the VLANs.

Uplink Load-Balancing Options
Redundancy and traffic sharing on multiple uplinks from the virtualized host must take into account the following:

- vPC pairs of leaf nodes: A common method of deployment is to use port channels from the virtualized hosts to the leaf nodes, so be sure that vPC leaf pairs are configured.
- Allocation of different a policy group for each ESXi host: Define which pair of interfaces on different leaf nodes should form a port channel.
• Consistent policy-group and vSwitch policy configuration: The policy-group configuration defines the configuration on the leaf ports. The vSwitch policy defines the configuration on the virtualized host uplinks. Depending on whether the ESXi host and the leaf nodes are directly attached, they may need to match.

• Bridge domain settings: Traffic forwarding based on MAC address pinning may require tuning of the bridge domain to take into account that only one uplink port is designed to handle broadcast packets such as ARP requests. These settings depend on the switch vendor implementation. When using VMM domains with vDS, with load balancing based on the virtual portID, you can keep bridge domain optimizations enabled.

VMware vDS Instantiation
For each VMM domain defined in Cisco ACI, the APIC creates a vDS in the hypervisor. In Figure 69, the user configured two VMM domains with the same vCenter but with different data centers, and as a result the APIC creates two vDS instances.

Figure 69. For Each VMM Domain, Cisco APIC Creates a VMware vDS in the Hypervisor

Management Connectivity from APIC to Virtual Machine Manager
The most common deployment option for virtualization is to deploy Cisco ACI with integration with the VMM using vDS or AVS. With this deployment mode, the APIC creates port groups on a vDS and helps ensure that the VLAN or VXLAN tagging is dynamically assigned.

For this integration, you must provision management connectivity from the APIC to vCenter.

In some cases, the VMM (for example, vCenter) may be connected directly to the Cisco ACI fabric. If this is the case, there may be a management EPG whose deployment depends on vCenter, while vCenter itself requires that management EPG to communicate with the APIC.

In this scenario, you should use pre provision Resolution Immediacy so that the leaf is preprogrammed with the necessary EPGs to help ensure management connectivity.
If you are using AVS as the virtual switch, the management information is provided to AVS by vCenter using opaque data. In this case, you need to help ensure that the ESXi host management connectivity is based on something other than AVS: for instance, on a vDS managed by the APIC (you can have more than one vDS on the same host, both managed by the APIC, with one based on AVS).

The Resolution Immediacy Pre-Provision option is not supported by AVS. AVS requires vCenter to provide the opaque data to establish the OpFlex channel, and pre provisioning the EPG, bridge domain, and VRF on the leaf would not be sufficient.

**Configuring VMware vDS Uplinks**

The vCenter administrator assigns the host and its uplinks to a vDS controlled by Cisco ACI. At the same time, the administrator also assigns the uplinks to a vDS uplink port group. For instance, the administrator might assign vCenter2-DVUplinks-95.

You should configure vDS through the APIC to use LLDP or Cisco Discovery Protocol:

- vDS supports only one discovery protocol—either Cisco Discovery Protocol or LLDP—not both simultaneously.
- LLDP takes precedence if both LLDP and Cisco Discovery Protocol are defined.
- To use Cisco Discovery Protocol, disable LLDP and enable Cisco Discovery Protocol.

You can configure a vDS through the APIC for uplink load balancing as follows:

- **LACP:** You can enable or disable LACP for the uplink port group. You can use this option when the corresponding ports on the leaf are configured for vPC.
- **MAC address pinning:** This option is equivalent to virtual port ID load balancing. When you use this option in conjunction with vDS, it does not require configuration of vPC or port channels on the leaf to which the ports are connected.

**Note:** If you use MAC address pinning with AVS for servers directly connected to Cisco ACI leaf nodes, you should enable vPC on the leaf nodes, because AVS needs to create two OpFlex channels, one per leaf and vPC, for the two channels to be established.

**Using Cisco AVS**

This section describes how Cisco ACI functions when you configure AVS with VXLAN. Cisco ACI uses the infrastructure VLAN to extend tenant Infra to the AVS with which it needs to communicate. This VLAN must be enabled in the AAEP configuration.

The AVS component communicates with the rest of the Cisco ACI infrastructure through a VMkernel interface that is automatically associated with a portgroup named vtep in the vDS.

**Figure 70** illustrates the interface created on the host and the association with the vtep portgroup.
The uplink portgroup includes the ports that you selected from each host as the uplinks for this vDS that is based on AVS (Figure 71).

Figure 71. Ports and Uplinks in the VMware vDS Based on Cisco AVS
The following configuration sequence brings up the OpFlex protocol that is used by Cisco ACI to communicate with the AVS:

- You add AVS under the VMM.
- AVS receives switch opaque data (SOD).
- A host is added to the vDS that is based on AVS.
- The uplink is brought up.
- The VTEP is created.
- The VTEP on the ESX host receives the IP address through DHCP from the tenant Infra range. This used to require the configuration of DHCP in Cisco ACI; however, this configuration step is no longer required.

**VTEP Interface on Cisco AVS**
When you use AVS, the APIC creates a VTEP port group that is associated with a VMkernel interface.

The IP address of this VMkernel interface is taken from the same range used for the PTEP addresses that are defined when you provision Cisco ACI. The vtep portgroup is connected to the infrastructure VLAN. This VLAN must be trunked from an Cisco ACI leaf so that the AVS can be part of the Cisco ACI Infra tenant.

The VTEP interface is used for OpFlex, VXLAN, and ERSPAN traffic.

Do not delete or change any parameters for the virtual machine kernel NIC (vmknic) created for the OpFlex channel.

**Using Multiple VMware vDS Devices**
In most cases, a single vDS with multiple port groups provides sufficient isolation. In some cases, however, multiple vDSs may be required for administrative reasons. The main restriction is that there can only be one vDS based on AVS per host.

You can have multiple vDSs on the same ESXi host (either APIC controlled or static) as long as they use different uplink vmnic interfaces.

**Note:** If you assign the same ESXi host to multiple VMM domains, you should define a non-overlapping range of VLANs for each VMM domain.

You can have vDSs of different types. For instance, one could be a vSphere vDS and another could be a VMM-based vDS.

The following are examples of supported deployment scenarios if each vDS uses a different set of uplink vmnics:

- vDS (unmanaged by APIC) and vDS (managed by APIC) on the same host: This is a common scenario for migrating from a deployment other than Cisco ACI to Cisco ACI.
- vDS (unmanaged by APIC) and AVS (managed by APIC) on the same host: This is another common scenario for migration.
- vDS (managed) and vDS (managed)
- vDS (managed) and AVS (managed)

There may be maximum and tested limits depending on the vendor and the software release.
Policy Resolution with Virtual Machines

In Cisco ACI, the deployment of bridge domains, VRFs, and of the policy CAM on a given leaf can be optimized based on the presence of virtual machines that require these resources. This was previously described in the section “Resolution and Deployment Immediacy of VRFs, Bridge Domains, EPGs, and Contracts”.

If the resolution is set to Pre-Provision, the bridge domains and VRFs are configured on all leaf nodes where the virtual domain is configured through an AAEP.

If the resolution is set to Immediate, the bridge domains and VRFs are deployed as soon as the host on which the EPG is present is discovered. This requires the host to run Cisco Discovery Protocol or LLDP so that the leaf can detect the host.

If the resolution is set to On-Demand, the bridge domains and VRFs are configured on the leaf nodes where at least one virtual machine is attached to the EPG. This also requires Cisco Discovery Protocol or LLDP so that the leaf can correlate the information about which host the virtual machine runs on and on which leaf this host is connected.

A special case is the use of a virtualized host that is not directly attached to a leaf, but through an intermediate Layer 2 switch. To implement the Resolution Immediacy On-Demand or Immediate option, Cisco ACI also uses Cisco Discovery Protocol or LLDP, but correlates the Cisco Discovery Protocol or LLDP information from the leaf and from the VMM (for example, vCenter). By correlating this information, Cisco ACI can resolve which port on which leaf the remote virtualized host is sending traffic to, and as a result it can deploy policies in an optimized way.

**Note:** The options with the virtualized host not directly connected to a leaf are described in the section “Design Considerations for Remotely Attached Hosts.”

Cisco Discovery Protocol and LLDP Configuration with VMware vDS

The ability to perform Cisco Discovery Protocol or LLDP discovery of the host requires the configuration on the leaf ports and on the ESXi host uplinks to be consistent. For this you need to configure the policy group for the interface (Figure 72) with the same LLDP or Cisco Discovery Protocol configuration as the vSwitch uplinks configuration in the VMM domain (Figure 73).
Cisco OpFlex Protocol
When using a device that supports OpFlex protocol, you do not need Cisco Discovery Protocol or LLDP for host discovery. OpFlex protocol takes care of this.

The OpFlex protocol communication happens in the infrastructure VRF and VLAN through TCP and SSL, and it is binary encoded. The AVS device gets a DHCP IP address in the infrastructure VRF from the APIC that is acting as the DHCP server.
The OpFlex protocol runs on port 8000 or 8002. You can view this information by logging into a virtualized server and issuing the command `vemcmd show OpFlex`.

```
[root@esx6-3:~] vemcmd show opflex
```

```
Status: 12 (Active)
[...]
Remote IP: 10.0.0.30 Port: 8000
Infra vlan: 4093
FTEP IP: 10.0.0.32
Switching Mode: NS
Encap Type: VXLAN
NS GIPO: 224.1.1.1
```

Although there may be no need for Cisco Discovery Protocol or LLDP for policy resolution, you may still want to enable Cisco Discovery Protocol and LLDP in the vSwitch policies and for the policy group for the leaf ports for troubleshooting purposes.

**Cisco OpFlex Protocol with vPC**

OpFlex protocol is optimized for use with vPC between two leaf nodes.

When the OpFlex discovery message is received from the OpFlex agent by the OpFlex element running on the leaf, it checks to see if the interface on which the discovery was received is part of a vPC. If so, two IP addresses are sent back: one for the local PTEP and the other for the vPC peer PTEP. The OpFlex agent on the AVS sets up two OpFlex channels: one to each PTEP.

When an endpoint is attached to an EPG, this is signaled to both leaf nodes on both channels. This causes endpoint policy to be downloaded on each of the two leaf nodes.

OpFlex protocol chooses one of the leaf nodes as primary to use for receiving policy updates.

**EPG Segmentation Options**

**Using VLANs to Segment EPGs**

You can use VLAN-based segmentation as the transport mechanism between the virtualized host and the leaf. With VLAN-based segmentation, forwarding uses Local Switching mode.

Whether you use vDS or AVS with VLAN-based segmentation, the configuration looks very similar. You must have a VLAN domain with a dynamic range of VLANs and this must be associated with the Cisco ACI leaf ports using an AAEP.

In addition to the range of VLANs used by the EPG when you use AVS, you must ensure that the infrastructure VLAN is enabled in the AAEP for the AVS to communicate with the rest of the Cisco ACI fabric (Figure 74).
Using VXLANs to Segment EPGs

You can use VXLAN-based segmentation as the transport mechanism between the virtualized host and the leaf. You can also use VXLAN-based segmentation with AVS or with vDS with VMware vShield.

When using Local Switching mode, you can choose to use VXLAN to segment the EPGs. When using Non-Local Switching (or FEX mode), you have to use VXLAN to segment the EPGs.

When running in VXLAN mode, each EPG is allocated a multicast address. For the configuration of AVS with VXLAN, you need to assign a pool of multicast addresses with a large enough range to be sufficient for all the EPGs.

**Note:** You do not have to specify a VXLAN range because Cisco ACI maintains a reserved range of VNIDs that are assigned to AVS.

Each multicast address is used by one EPG. These addresses don't need to be different than the multicast range used by the ACI fabric for VXLAN purposes because the traffic from the AVS to the leaf is decapsulated at the leaf and reencapsulated before being sent to the fabric. The reason for using one multicast address per EPG is to conform to the RFC specifications for VXLAN. When there is multidestination traffic on a Bridge Domain the leaf replicates the multidestination traffic sending one copy per EPG with the destination set to the EPG multicast address.

In addition to this you need to define a multicast address called “AVS Fabric-Wide Multicast Address” which is also referred to as a “VMM GIPO”. This address represents all the leafs that the AVS connects to and it is used for AVS-to-leaf traffic as if it was an any cast VTEP.

The AVS VTEP encapsulates traffic with a VNI and a corresponding multicast address (which in Cisco ACI terminology is called a GIPO').

After the packet reaches the ingress leaf switch, the leaf removes the GIPO’, and it adds the GIPO of the bridge domain and sends the packet over the fabric.
In the egress direction from the leaf switch to the virtualized host, Cisco ACI removes the bridge domain GIPo and adds the EPG GIPo’, and then sends the packet to VTEPs.

Even if there is a dedicated multicast address for each EPG, the forwarding behavior of broadcast, unknown unicast, and multicast traffic is not different than it is in the bridge domain, in that all EPGs within the bridge domain receive it even if they have a different multicast IP address.

The administrator must configure the multicast range as well as the multicast IP address that represents all the EPGs on the AVS. The reason for allowing the administrator to choose this range is that it is also possible to have other switches in the path between the virtualized server and the leaf, in which case the administrator must choose a range of multicast IP addresses that do not overlap with the existing ones.

When it comes to preparing the virtual infrastructure, the main differences between AVS and vDS with vShield are as follows:

- If you use VXLAN with AVS, Cisco ACI uses a predefined pool of VNIs to help ensure that the same VNID is not assigned to another EPG.
- If you use VXLAN with vDS and vShield, the administrator must ensure that the range of VNIDs is not overlapping or duplicated in different vShield controllers or domains. If the administrator assigns domains that are overlapping to the same leaf, the leaf raises a fault.

**Micro-segmentation Considerations**

“Micro-segmentation” is an umbrella term used to refer to a set of features that includes two main categories of functions:

- Isolation of the endpoints within an EPG, similar to the private VLAN feature in traditional NX-OS switches
- The capability to automatically classify workloads into an EPG based on attributes of the virtual machine: for instance, the hostname, the operating system, etc.

The first option consists of constraining the communication between endpoints of the same EPG. This option is similar to the isolated VLANs feature in Cisco NX-OS. You can also create intra-EPG contracts to allow endpoints within the same EPG to communicate only on ports defined by the intra-EPG contract.

At the time of this writing, you can classify physical servers and virtual machines based on several characteristics:

- MAC address: Applicable to both physical and virtual machines
- IP address: Applicable to both physical and virtual machines
- DNS name: Applicable to both physical and virtual machines
- Virtual machine custom attribute
- Virtual machine operating system
- Virtual machine datacenter
- Virtual machine identifier
- Virtual machine name
- vNIC Dn
- Virtual machine - tag

If you are using Cisco Nexus 9300EX or a newer platform, you can deploy micro-segmentation with a standard vDS and also with the AVS software switch.
If you are using first-generation Cisco ACI leaf switches, you may have to use the AVS software switch to configure certain micro-segmentation options. Table 4 clarifies this point.

Table 4 provides a reference showing which classification feature is applicable to which type of endpoint, from which software it is supported, what hardware it requires, and whether it demands the use of a VLAN or VXLAN transport.

**Note:** EPGs configured with micro-segmentation require resolution to be immediate.

### Table 4. Micro-segmentation Options

<table>
<thead>
<tr>
<th>Classification Feature</th>
<th>Type of Endpoint</th>
<th>Software</th>
<th>Hardware</th>
<th>Host Software Switch</th>
<th>VLAN or VXLAN Requirement</th>
<th>Deployment Immediacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>VLAN-based</td>
<td>Physical</td>
<td>X</td>
<td>X</td>
<td>Release 1.0 and later</td>
<td>All Cisco Nexus 9300 platforms</td>
<td>No requirements</td>
</tr>
<tr>
<td></td>
<td>Virtual</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VXLAN based</td>
<td>Physical</td>
<td>X</td>
<td></td>
<td>Release 1.0 and later</td>
<td>All Cisco Nexus 9300 platforms</td>
<td>No requirements</td>
</tr>
<tr>
<td></td>
<td>Virtual</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IP-based classification: /22 and subnets</td>
<td>Physical</td>
<td>X</td>
<td>Release 1.2</td>
<td>Cisco Nexus 9300 E and EX platforms</td>
<td>N/A</td>
<td>VLAN</td>
</tr>
<tr>
<td></td>
<td>Virtual</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAC-based classification</td>
<td>Physical</td>
<td>X</td>
<td></td>
<td>Release 2.1</td>
<td>Cisco Nexus E (Donner-C) or EX platform hardware</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Virtual</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPG isolation</td>
<td>Physical</td>
<td>X</td>
<td></td>
<td>Release 2.0</td>
<td>All Cisco Nexus 9300 platforms</td>
<td>Hyper-V with OpFlex</td>
</tr>
<tr>
<td></td>
<td>Virtual</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 4. EPG isolation

<table>
<thead>
<tr>
<th>Classification Feature</th>
<th>Type of Endpoint</th>
<th>Software</th>
<th>Hardware</th>
<th>Host Software Switch</th>
<th>VLAN or VXLAN Requirement</th>
<th>Deployment Immediacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virtual machine attributes classification</td>
<td>Physical</td>
<td>X</td>
<td>Release 1.1(1j)</td>
<td>All Cisco Nexus 9300 platforms</td>
<td>AVS</td>
<td>VLAN and VXLAN</td>
</tr>
<tr>
<td></td>
<td>Virtual</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPG isolation</td>
<td>Physical</td>
<td>X</td>
<td></td>
<td>Release 1.2(2h)</td>
<td>All Cisco Nexus 9300 platforms</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Virtual</td>
<td>X</td>
<td></td>
<td>Release 1.2(2h)</td>
<td>All Cisco Nexus 9300 platforms</td>
<td>vDS</td>
</tr>
<tr>
<td></td>
<td>Virtual</td>
<td>X</td>
<td></td>
<td>Release 1.3</td>
<td>All Cisco Nexus 9300 platforms</td>
<td>AVS</td>
</tr>
</tbody>
</table>

© 2018 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.
Uplink Load-Balancing Options
Cisco ACI offers several load-balancing options for teaming purposes from the virtualized host to the leaf nodes, as illustrated in Figure 75.

Figure 75. Uplink Load-Balancing Options

Note that not all modes work for all virtualized host deployments:

- LACP Passive mode is not supported for directly connected hosts. Ports using LACP Passive mode do not initiate an LACP handshake. We recommend that you always use LACP Active rather than LACP Passive on leaf ports. LACP Passive mode can be used on leaf policy groups when there is an intermediate Layer 2 device between the leaf and the AVS and where the Layer 2 device ports are using LACP Active mode.
- MAC Pinning-Physical-NIC-load mode is not supported for Cisco AVS
- MAC Pinning mode with AVS is meaningful only when you have a Layer 2 device between the leaf and the virtualized host. If you use MAC pinning with an AVS-based vDS directly attached to a leaf, you should also enable vPC on the leaf nodes.
- The Minimum Number of Links and Maximum Number of Links options are relevant only when LACP is used.

Policy Group Versus vSwitch Policy
Cisco ACI can configure both the fabric leaf ports and the virtual switching infrastructure within the virtualized server.

For this purpose, there are two configuration constructs: one focused on the fabric ports and another focused on the virtualized server:

- Fabric Access-> Interface Policies-> Policy Group: By creating a policy group of type leaf access, port channel, and vPC, you can define the way that you want to configure the leaf ports that connect to the virtualized host.
- VM Networking> VMM Domain> vSwitch policies: This configuration controls the vSwitch or vDS on the server side that connects to the Cisco ACI fabric (or to an intermediate network infrastructure).

Figure 76 presents an example of servers connected to Cisco ACI through an intermediate network and shows which network is configured by which option.

**Figure 76.** Servers Connected to Cisco ACI Through an Intermediate Network

In some cases, the vDS uplink configuration is automatically derived from the policy-group configuration. For instance, if you configure a policy group of type vPC with LACP, Cisco ACI will program uplink load balancing on the vDS accordingly without any vSwitch configuration.

In other cases, it may be necessary to have a more specific configuration on the vDS uplinks, and if you configure the vSwitch uplink policy, you can override the policy-group configuration that has been copied to the vDS.

In yet other cases, you may simply need to have two very different configurations on the leaf compared than on the vDS, because the leaf may connect to other switches that run port channels, for instance, whereas the hosts may connect to these other switches with a different uplink configuration such as MAC address pinning.

**VMware vDS Teaming Options: MAC Address Pinning**

If you deploy a vDS controlled by the APIC, and if the ESXi hosts are dual-homed to the leaf nodes, you can configure all leaf ports connected to the ESX hosts identically with a leaf access policy group (a regular access-policy group rather than vPC). You should ensure that either Cisco Discovery Protocol or LLDP is enabled.

You should also configure the vSwitch policy in the VMM domain for MAC pinning and for Cisco Discovery Protocol or LLDP.
Resolution and Deployment Immediacy can be selected as desired. If one vDS EPG is providing management connectivity for vCenter, you should configure Resolution Immediacy as Pre-Provision.

The EPGs will show that the same endpoints are learned on both leaf nodes to which the vmnics are attached (indicated as “vmm”) for the purposes of the Resolution and Deployment Immediacy settings, but they are learned only on the leaf where the virtual machine is pinned for the purposes of the mapping database and data-plane forwarding.

Bridge domain settings with vDS can be set to optimized mode: using hardware proxy with ARP flooding disabled unless silent hosts are present. vDS with MAC pinning forwards multidestination traffic coming from the leaf to a vmnic, to the virtual machines that are associated with that vmnic. This behavior is consistent with the way that Cisco ACI optimizes ARP, because Cisco ACI forwards ARP requests to the vmnic where it discovered the virtual machine.

Figure 77 illustrates the configuration of vDS with MAC address pinning.

Figure 77. MAC Address Pinning Configuration

**VMware vDS Teaming Options: Static Port Channel with vPC**

If you deploy a vDS controlled by an APIC, and if the ESXi hosts are directly attached to the leaf nodes, you should configure as many vPC policy groups as the number of ESXi hosts and assign them to pairs of interfaces on two leaf nodes. The policy group should also have Cisco Discovery Protocol or LLDP enabled.

You should ensure that the interfaces to which each policy group is assigned match the vmnics of each ESX host.

The vSwitch policy configuration would not be required in this case, because it is derived from the policy-group configuration.

Resolution and Deployment Immediacy options can be selected as desired. If one vDS EPG is providing management connectivity for vCenter, you should configure Resolution Immediacy as Pre-Provision.
The EPGs will show that each endpoint is learned on the vPC to which the vmnics are attached (indicated as "vmm") for the purposes of the Resolution and Deployment Immediacy settings, and learned from the vPC as well for the purposes of the mapping database and data-plane forwarding.

**Figure 78** illustrates the configuration of vDS deployment with vPC.

**Figure 78.** VMware vDS Deployment with vPC

---

**Cisco AVS Teaming Options with Directly Attached Servers: vPC with LACP**

When using AVS with hosts directly attached to the leaf nodes, you should use vPC with LACP.

If you use AVS configured with MAC pinning with directly attached hosts, this will work only if vPC is also configured on the leaf nodes. In contrast to vDS, AVS uses the OpFlex protocol to communicate with the leaf nodes and to indicate which endpoint is present. If you are using MAC pinning without vPC on the leaf nodes, the AVS will be able to establish only one channel to one leaf, and endpoints will appear as if they were all connected to that single leaf.

If you use vPC instead, Cisco ACI is able to program AVS to have a channel to each vPC peer. When a new endpoint is discovered, the information about the endpoint is synchronized to both vPC peers.

With vPC configured on the leaf nodes, you then have the choice of using LACP on the AVS and host side, which is the preferred option.

With this configuration, you should configure as many vPC policy groups as the number of ESXi hosts and assign them to pairs of interfaces on two leaf nodes. The policy group can also have Cisco Discovery Protocol or LLDP enabled, but this is not mandatory.

You should ensure that the interfaces to which each policy group is assigned match the vmnics of each ESXi host.

The vSwitch policy configuration would not be required in this case, because it is derived from the policy-group configuration.

Resolution and Deployment Immediacy settings can be selected as desired. The policy resolution Pre-Provision option is not supported with AVS.
The bridge domain settings can use Cisco ACI optimizations such as hardware proxy with ARP flooding disabled if applicable.

The EPGs will show that each endpoint is learned on the vPC to which the vmnics are attached (indicated as “vmm”) for the purposes of the Resolution and Deployment Immediacy settings, and learned from the vPC as well for the purposes of the mapping database and data-plane forwarding.

You can verify the status of the port channel from the AVS as in the following example:

```
[root@esx-6-2:] vemcmd show lacp
LACP Offload is Enabled

---------------------------------------------------
LACP Offload Config for LTL 20
---------------------------------------------------
LACP Bit Set : Yes
LACP Port state : Channeling

---------------------------------------------------
LACP Offload Config for LTL 21
---------------------------------------------------
LACP Bit Set: Yes
LACP Port state: Channeling
```

Figure 79 shows the design with AVS configured for LACP.

**Figure 79.** Design with Cisco AVS Configured for LACP
Cisco AVS Teaming Options with Remotely Attached Servers

This section refers to the configuration where there is a switching infrastructure between the virtualized host and the leaf nodes and where AVS is used to create the overlay to connect virtual machines to the leaf nodes.

In this case, you should configure the policy group for connectivity to the switching infrastructure from the leaf, and the vSwitch policy for connectivity of the virtualized host to the switching infrastructure.

With the MAC pinning option on the vSwitch policy, each VMkernel interface (VTEP) in the AVS is associated with one of the two available vmnics, and if one vmnic fails, the virtual machines that were using it start using the remaining ones.

VMkernel Connectivity

vMotion should be configured on a separate vmknic with a dedicated EPG. vMotion should not be configured on the same vmknic used for OpFlex protocol, which is connected by default to the vtep port group.

It may be desirable to set Resolution Immediacy for the EPG for VMkernel to Pre-Provision. This is also desirable if you are using NFS access to store the virtual machine disk.

For these reasons, for vMotion it may be desirable to use a separate vDS controlled by the APIC and configured with the Resolution Immediacy Pre-Provision option.

Because virtualized hosts are normally shared by multiple tenants, it may also be desirable to configure the infrastructure for vMotion in the Common tenant:

- A dedicated bridge domain
- A dedicated VRF

The VMkernel interface is silent for the most part, so the bridge domain configuration should take this into account:

- If the bridge domain is configured for unknown unicast flooding and ARP flooding, the bridge domain does not need to have any routing enabled, nor a subnet.
- If the bridge domain is configured for hardware proxy, it may be advisable to also configure routing, disable ARP flooding, and configure a subnet IP address on the same subnet as the VMkernel interface used for vMotion purposes.

In the first case, the VMkernel communication happens between host H1 and H2, even if they never sent out any packets before, because the ARP request from H1 on the VMkernel interface is flooded on bridge domain BD1 (Figure 80).
In the second case, the VMkernel communication happens between host H1 and H2, even if they never sent out any packets before, because Cisco ACI performs ARP gleaning. When the VMkernel interface of H1 sends an ARP request for the VMkernel IP address of H2, Cisco ACI instructs all the SVIs on BD1 to send ARP requests for the unknown endpoint IP address. The response to the ARP requests not only provides the ARP information to H1, but also updates the mapping database with the IP and MAC address information for the VMkernel interface of H2.

**Note:** If you make configuration changes to the bridge domain, the MAC and IP address information learned in the mapping database is cleared, and if the ARP table on H1 has not expired, H1 may try to send traffic to H2 without sending any ARP request. As a result, vMotion for virtual machines running on H1 may fail.

**Error! Reference source not found.** shows the EPG operational view for client endpoints. It illustrates the fact that it is possible that the IP address and MAC address of the VMkernel interface is not learned until the first vMotion movement happens.

The setup used in Figure 81 consists of a bridge domain configured to provide both NFS connectivity and vMotion support. Figure 82 shows the MAC and IP addresses learned in the EPG. The IP addresses and MAC addresses of the VMkernel interfaces used for NFS are learned (learning source: vmm, learned). The MAC and IP addresses of the VMkernel interfaces used for vMotion instead are not learned (learning source: vmm; vCenter provided this information, but the endpoint is not in the mapping database yet, because it is not learned).

The IP addresses and MAC addresses of the VMkernel interfaces used for vMotion will be learned during the first vMotion event.
Summary of Best Practices for the Virtual Infrastructure

The following list summarizes some of the best practices for deployment with a virtual infrastructure:

- Consider the Resolution and Deployment Immediacy On-Demand options for most workloads to limit the consumption of hardware resources.
- Consider the Resolution Immediacy Pre-Provision option for management connectivity to the virtualized hosts or to the VMM system if it is connected to Cisco ACI leaf switches for management.
- Consider vPC for the uplink connectivity of the virtualized servers and the use of LACP.
- Consider the use of AVS for remote servers if you want to limit the number of VLANs to be trunked.
- Consider the use of AVS for micro segmentation on first-generation leaf switches.
- Remember to include the infrastructure VLAN in the AAEP for the AVS.
- If you are using AVS, the recommended approach is mixed encapsulation, with VXLAN as the default encapsulation and with local switching mode. This configuration reduces the number of VLANs to be trunked and gives you the flexibility to use mixed encapsulation at some point if this feature is needed.

Designing External Layer 3 Connectivity

This section explains how Cisco ACI can connect to outside networks using Layer 3 routing. It explains the route exchange between Cisco ACI and the external routers, and how to use dynamic routing protocols between the Cisco ACI border leaf switch and external routers. It also explores the forwarding behavior between internal and external endpoints and the way that policy is enforced for the traffic flow between them.

Cisco ACI refers to external Layer 3 connectivity as a L3Out connection. This section provides additional details about the implementation and use of the L3Out connection by the tenant administrator.

In a regular configuration, route peering and static routing are performed on a per-VRF basis, in a manner similar to the use of VRF-lite on traditional routing platforms. External prefixes that are learned on a per-VRF basis are redistributed to the leaf nodes where the tenant endpoints reside. Routes are installed in the forwarding tables of the leaf nodes only if the specific VRF is deployed on that leaf.

Alternatively, shared Layer 3 connectivity can be provided in two ways: using shared L3Out connections, or using an MP-BGP and EVPN plus VXLAN connection to an external device (such as a Cisco Nexus 7000 Series Switch with appropriate hardware and software). This has the advantage of not requiring separate L3Out policies for each individual tenant and VRF. The section “Designing for Multitenancy and Shared Services” provides more details.
Layer 3 Outside and External Routed Networks

In a Cisco ACI fabric the bridge domain is not meant for the connectivity of routing devices, and this is why you cannot configure static or dynamic routes directly on a bridge domain. You instead need to use a specific construct for routing configurations: the L3Out.

L3Out policy is used to configure interfaces, protocols, and protocol parameters necessary to provide IP connectivity to external routing devices. An L3Out connection is always associated with a VRF. L3Out connections are configured using the External Routed Networks option on the Networking menu for a tenant.

Part of the L3Out configuration involves also defining an external network (also known as an external EPG) for the purpose of access-list filtering. The external network is used to define which subnets are potentially accessible through the Layer 3 routed connection. In Figure 82, the networks 50.1.0.0/16 and 50.2.0.0/16 are accessible outside the fabric through an L3Out connection. As part of the L3Out configuration, these subnets should be defined as external networks. Alternatively, an external network could be defined as 0.0.0.0/0 to cover all possible destinations.

Figure 82. External Network

After an external network has been defined, contracts are required between internal EPGs and the external networks in order for traffic to flow. When defining an external network, check the box External Subnets for the External EPG, as shown in Figure 83. The other checkboxes are relevant for transit and shared services scenarios and are described later in this section.
Figure 83. Defining Traffic Filtering for Outside Traffic

L3Out Simplified Object Model

Figure 84 shows the object model for L3Out. This helps in understanding the main building blocks of the L3Out model.

Figure 84. Object Model for L3Out
The L3Out policy is associated with a VRF and consists of the following:

- Logical node profile: This is the leaf wide VRF routing configuration, whether it is dynamic or static routing. For example, if you have two border leaf nodes, the logical node profile consists of two leaf nodes.
- Logical interface profile: This is the configuration of Layer 3 interfaces or SVIs on the leaf defined by the logical node profile. The interface selected by the logical interface profile must have been configured with a routed domain in the fabric access policy. This routed domain may also include VLANs if the logical interface profile defines SVIs.
- External network and EPG: This the configuration object that classifies traffic from the outside into a security zone.

The L3Out connection must be referenced by the bridge domain whose subnets need to be advertised to the outside.

L3Out policies, or external routed networks, provide IP connectivity between a VRF and an external IP network. Each L3Out connection is associated with one VRF instance only. A VRF may not have an L3Out connection if IP connectivity to the outside is not required.

A L3Out configuration always includes a router ID for each leaf as part of the node profile configuration, regardless of whether the L3Out connection is configured for dynamic routing or static routing.

**Route Announcement Options for the Layer 3 Outside**

This section describes the configurations needed to specify which bridge domain subnets are announced to the outside routed network and which outside routes are imported into the Cisco ACI fabric.

When specifying subnets under a bridge domain or an EPG for a given tenant, you can specify the scope of the subnet:

- Advertised Externally: This subnet is advertised to the external router by the border leaf.
- Private to VRF: This subnet is contained within the Cisco ACI fabric and is not advertised to external routers by the border leaf.
- Shared Between VRF Instances: This option is for shared services. It indicates that this subnet needs to be leaked to one or more private networks. The shared-subnet attribute applies to both public and private subnets.

For subnets defined in the bridge domain to be announced to the outside router, the following conditions must be met:

- The subnets need to be defined as advertised externally.
- The bridge domain must have a relationship with the L3Out connection (in addition to its association with the VRF instance).
- A contract must exist between the Layer 3 external EPG (external subnets for the external EPG) and the EPG associated with the bridge domain. If this contract is not in place, the announcement of the subnets cannot occur.
- When defining a L3Out configuration, the route control export option is automatically selected. You define which subnets are announced from the L3Out to the outside by configuring the default-export route map. **Figure 85** shows the bridge domain and subnet configuration with the relationships to a L3Out.
You can control which of the outside routes learned through the L3Out are imported into the Cisco ACI fabric. You do this using the default-import route-map configuration under the L3Out (Figure 86).

These route maps apply to all routes:

- Directly connected subnets
- Static routes
- Transit routes

Figure 86. L3Out Configuration to Control Imported and Exported Routes
External Network (External EPG) Configuration Options

The external endpoints are assigned to an external EPG (which the GUI calls an external network). For the L3Out connections, the external endpoints can be mapped to an external EPG based on IP prefixes or host addresses.

**Note:** EPGs for external or outside endpoints are sometimes referred to as prefix-based EPGs if defined as networks and masks, or IP-based EPGs if defined as /32. IP-based EPG is also the terminology used to define EPG classification based on the IP address for hosts directly attached to the leaf nodes.

For each L3Out connection, the user has the option to create one or multiple external EPGs based on whether different policy treatments are needed for different groups of external endpoints.

Under the Layer 3 external EPG configurations, the user can map external endpoints to this EPG by adding IP prefixes and network masks. The network prefix and mask don’t need to be the same as the ones in the routing table. When only one external EPG is required, simply use 0.0.0.0/0 to assign all external endpoints to this external EPG.

After the external EPG has been created, the proper contract can be applied between the external EPG and other EPGs.

The main function of the external network configuration (part of the overall L3Out configuration) is to classify traffic from the outside to an EPG to establish which outside and inside endpoints can talk. However, it also controls a number of other functions such as import and export of routes to and from the fabric.

This is a summary of the options that for the external network configuration and the functions they perform:

- **Subnet:** This defines the subnet that is primarily used to define the external EPG classification.
- **Export Route Control Subnet:** This configuration controls which of the transit routes (routes learned from another L3Out) should be advertised. This is an exact prefix and length match. This item is covered in more detail in the “Transit Routing” section.
- **Import Route Control Subnet:** This configuration controls which of the outside routes learned through BGP should be imported into the fabric. This is an exact prefix and length match.
- **External Subnets for the External EPG:** This defines which subnets belong to this external EPG for the purpose of defining a contract between EPGs. This is the same semantics as for an ACL in terms of prefix and mask.
- **Shared Route Control Subnet:** This indicates that this network, if learned from the outside through this VRF, can be leaked to the other VRFs (if they have a contract with this external EPG).
- **Shared Security Import Subnets:** This defines which subnets learned from a shared VRF belong to this external EPG for the purpose of contract filtering when establishing a cross-VRF contract. This configuration matches the external subnet and masks out the VRF to which this external EPG and L3Out belong.
- **Aggregate Export:** This option is used in conjunction with Export Route Control Subnet and allows the user to export all routes from one L3Out to another without having to list each individual prefix and length. This item is covered in more detail in the “Transit Routing” section.
• **Aggregate Import**: This allows the user to import all the BGP routes without having to list each individual prefix and length. You achieve the same result by not selecting Route Control Enforcement Input in the L3Out (which is the default). This option is useful if you have to select Route Control Enforcement Input to then configure action rule profiles (to set BGP options for instance), in which case you would then have to explicitly allow BGP routes by listing each one of them with Import Route Control Subnet. With Aggregate Import, you can simply allow all BGP routes. The only option that can be configured at the time of this writing is 0.0.0.0/0.

**Border Leaf Switch Designs**

Border leaf switches are Cisco ACI leaf switches that provide Layer 3 connections to outside networks. Any Cisco ACI leaf switch can be a border leaf, and there is no limitation on the number of leaf switches that can be used as border leaf switches. The border leaf can also be used to connect to computing, IP storage, and service appliances. In large-scale design scenarios, for greater scalability, it may be beneficial to separate border leaf switches from the leaf switches that connect to computing and service appliances.

Border leaf switches support three types of interfaces to connect to an external router:

- Layer 3 (routed) interface
- Sub interface with IEEE 802.1Q tagging: With this option, multiple sub interfaces can be configured on the main physical interface, each with its own VLAN identifier.
- Switched virtual interface: With an SVI, the same physical interface that supports Layer 2 and Layer 3 can be used for Layer 2 connections as well as an L3Out connection.

In addition to supporting routing protocols to exchange routes with external routers, the border leaf applies and enforces policy for traffic between internal and external endpoints.

Cisco ACI supports the following routing mechanisms:

- Static routing (supported for IPv4 and IPv6)
- OSPFv2 for regular, stub, and not-so-stubby-area (NSSA) areas (IPv4)
- OSPFv3 for regular, stub, and NSSA areas (IPv6)
- EIGRP (IPv4 only)
- iBGP (IPv4 and IPv6)
- eBGP (IPv4 and IPv6)

Through the use of sub interfaces or SVIs, border leaf switches can provide L3Out connectivity for multiple tenants with one physical interface.

**Outside Bridge Domains**

L3Out can be configured with Layer 3 interfaces, sub interfaces, or SVIs. When configuring an SVI on L3Out, you can specify a VLAN encapsulation. Specifying the same VLAN encapsulation on multiple border leaf nodes in the same L3Out results in the configuration of an external bridge domain.

Compared to a bridge domain inside the fabric, there is no mapping database for the L3Out, and the forwarding of traffic at Layer 2 is based on flood and learn over VXLAN.

If the destination MAC is the SVI MAC address, the traffic is routed in the fabric as already described.
An L3Out connection is instantiated immediately on the leaf because it is not associated with the discovery of endpoints.

**L3Out with vPC**

You can configure dynamic routing protocol peering over a vPC for an L3Out connection by specifying the same SVI encapsulation on both vPC peers as illustrated in Figure 87. The SVI configuration instantiates a bridge domain (which in the figure has a VNID of 5555). The external router peers with the SVI on each leaf device. In addition, the SVIs on the two leaf devices peer with each other.

If static routing to the fabric is required, you must specify the same secondary IP address on both vPC peer devices’ SVIs.

**Figure 87.** Dynamic Routing: Peering over vPC

Additional design considerations are necessary when using a L3Out based on vPC with more than two border leaf switches.

**Considerations When Using More Than Two Border Leaf Switches**

Depending on the hardware used for the leaf switches and on the software release, the use of more than two border leaf switches as part of the same L3Out in Cisco ACI may be restricted. Restriction occurs in the following cases:

- The L3Out consists of more than two leaf switches with the SVI in the same encapsulation (VLAN)
- The border leaf switches are configured with static routing to the external device
- The connectivity from the outside device to the fabric is vPC based
These restrictions occur because traffic may be routed from one datacenter to the local L3Out and then bridged on the external bridge domain to the L3Out in another datacenter.

Figure 88 shows, on the left, a topology that works with both first- and second-generation leaf switches. The topology on the right works with only Cisco Nexus 9300-EX and Cisco 9300-FX platform switches. In the topologies, Cisco ACI is configured for static routing to an external active-standby firewall pair. The L3Out uses the same encapsulation on all the border leaf switches to allow static routing from any border leaf to the active firewall. The dotted lines indicate the border leaf switches.

Figure 88. Design Considerations with Static Routing L3Out with SVI and vPC

With topologies consisting of more than two border leaf switches, the preferred approach is to use dynamic routing and to use a different VLAN encapsulation for each vPC pair on the L3Out SVI. This approach is preferred because the fabric can route the traffic to the L3Out interface that has reachability to the external prefix without the need to perform bridging on an outside bridge domain. Figure 89 illustrates this point.

Figure 89 shows four border leaf switches: two in each datacenter. There are two L3Outs or a single L3Out that uses different VLAN encapsulations for DC1 and DC2. The L3Out is configured for dynamic routing with an external device.

For this design, there are no specific restrictions related to routing to the outside.

Figure 89. Design Considerations with Dynamic Routing L3Out with SVI and vPC
Considerations When Connecting Servers to Border Leaf Switches

Dedicated border leaf switches offer increased availability. For example, failure scenarios related to routing protocols don’t affect server-to-server connectivity, and a computing leaf failure doesn’t affect outside reachability from another computing leaf.

However, for smaller scale datacenters it is sometimes necessary also to use border leaf switches to connect workloads as depicted in Figure 90.

The recommendations related to this design take into account the policy CAM filtering optimization called ingress filtering, which is controlled by the configurable Policy Control Enforcement Direction option in the VRF configuration.

![Figure 90. Design Considerations When Attaching Endpoints to Border Leaf Switches]

The following considerations apply to this design:

- For best practices related to the use of external routers or firewalls connected in vPC mode to the border leaf, please refer to the previous section.

- Attachment of endpoints to border leaf switches is fully supported when the leaf switches in DC1 and DC2 are all Cisco Nexus 9300-EX and Cisco 9300-FX platform switches. In presence of L3Out with SVI and external bridge domain stretched between border leaves you should disable remote IP address endpoint learning on the border leaf from Fabric > Access Policies > Global Policies > Fabric Wide Setting Policy by selecting Disable Remote EP Learn.

- If the computing leaf switches—that is, the leaf switches to which the servers are connected—are first-generation leaf switches, you need to consider the following options:
  - If VRF ingress policy is enabled (which is the default and recommended configuration), you need to make sure that the software is Cisco ACI Release 2.2(2e) or later. You also should configure the option to disable endpoint learning on the border leaf switches. You can disable remote IP address endpoint learning on the border leaf from Fabric>Access Policies>Global Policies>Fabric Wide Setting Policy by selecting Disable Remote EP Learn.
  - You can also configure the VRF instance for egress policy by selecting the Policy Control Enforcement Direction option Egress under Tenants>Networking>VRFs.
Gateway Resiliency with L3Out

Cisco ACI uses a pervasive gateway as the default gateway for servers. The pervasive gateway is configured as a subnet under the bridge domain.

Some design scenarios may require gateway resiliency on the L3Out. For example, external services devices (such as firewalls) may require static routing to subnets inside the Cisco ACI fabric, as shown in Figure 91.

For L3Outs configured with static routing, Cisco ACI provides multiple options for a resilient next hop:

- Secondary IP: This option is available on routed interfaces, sub interfaces, and SVIs, but it is used primarily with SVIs.
- Hot Standby Routing Protocol (HSRP): This option is available on routed interfaces and on sub interfaces (and not on SVIs). It is used primarily in conjunction with an external switching infrastructure that helps ensure Layer 2 connectivity between the sub interfaces.

Figure 91. L3Out Secondary Address Configuration

In the example in Figure 91, a pair of Cisco ASA firewalls (running in active-standby mode) are attached to the Cisco ACI fabric. On the fabric side, L3Out is configured to connect to the firewalls. The Layer 2 connectivity for subnet 192.168.1.0/24 is provided by the Cisco ACI fabric by using SVIs with the same encapsulation on both leaf switches. On the firewalls, a static route exists pointing to internal Cisco ACI subnets through the 192.168.1.254 address. This .254 address is configured on the fabric as a shared secondary address under the L3Out configuration. When configuring the interface profile under L3Out, configuration options exist for Side A, Side B, and secondary addresses, as shown in Figure 92.
In this example, 192.168.1.254 is configured as the shared secondary address, which is then used as the next hop for the static route configured on the firewall.

**Route Distribution Within the Cisco ACI Fabric**

MP-BGP is implemented between leaf and spine switches to propagate external routes within the Cisco ACI fabric. BGP route reflectors are deployed to support a large number of leaf switches within a single fabric. All the leaf and spine switches are in one single BGP autonomous system (including all pods in a Cisco ACI multipod deployment).

After the border leaf learns the external routes, it redistributes the external routes from a given VRF instance to an MP-BGP VPNv4 address family instance. MP-BGP maintains a separate BGP routing table for each VRF instance. Within MP-BGP, the border leaf advertises routes to a spine switch, which is a BGP route reflector. The routes are then propagated to all the leaf switches on which the VRF instances are instantiated. **Figure 93** illustrates the routing protocol within the Cisco ACI fabric and the routing protocol between the border leaf and external router using VRF-lite.
Advertisement of Bridge Domain Subnets

Border leaf switches are the location at which tenant (bridge domain) subnets are injected into the protocol running between the border leaf switches and external routers.

Announcing bridge domain subnets to the outside requires the configurations previously described: a subnet under the bridge domain defined as Advertised Externally, a reference to the L3Out from the bridge domain, and a contract between the external EPG and internal EPGs and the definition of a default export policy as a route profile.

Administrators determine which tenant subnets they want to advertise to the external routers. When specifying subnets under a bridge domain or an EPG for a given tenant, the user can specify the scope of the subnet:

- **Advertised Externally**: This subnet is advertised to the external router by the border leaf using the associated L3Out.
- **Private to VRF**: This subnet is contained within the Cisco ACI fabric and is not advertised to external routers by the border leaf.
- **Shared Between VRFs**: This option is used for shared services. It indicates that this subnet needs to be leaked to one or more private networks. The shared-subnet attribute applies to both public and private subnets.
Importing Routes
External prefixes learned by an L3Out may or may not be automatically redistributed to MP-BGP, depending on the configuration of the Route Control Enforcement import option in the L3Out.

If L3Out Route Control Enforcement is not selected, all networks learned from the outside are redistributed to MP-BGP.

You can control which routes are imported if, under L3Out, you choose the Route Control Enforcement option and select Import.

This option applies to OSPF, EIGRP, and BGP.

You can specify the prefixes that are redistributed by configuring the default import route profile under the L3Out.

**Note:** You can also define which routes are imported by configuring subnets under the Layer 3 external network and selecting Import Route Control Subnet for each network. This configuration is a specific match (that is, a match of the prefix and prefix length).

BGP Autonomous System Number Considerations
The Cisco ACI fabric supports one Autonomous System (AS) number. The same AS number is used for internal MP-BGP and for the BGP session between the border leaf switches and external routers. The BGP AS number is configured as described previously in the “Preparing the Fabric Infrastructure” and “BGP Route Reflector Policy” sections.

It is possible for an administrator to override the global AS number configuration using the local AS number found under the BGP peer connectivity profile when configuring each L3Out. This can be used if the administrator wants the Cisco ACI fabric to appear as a different AS number to the one configured globally. This configuration is shown in Figure 94.
Figure 94. L3Out BGP Configuration

BGP Route Reflector Placement
In the Cisco ACI fabric, BGP route reflectors are used for two purposes:

- Regular BGP route reflectors are used for traditional L3Out connectivity through leaf nodes.
- BGP EVPN route reflectors are used for multipod and EVPN WAN connectivity.

For traditional L3Out connectivity (that is, through leaf nodes), it is recommended to configure a pair of route reflectors per Cisco ACI pod for redundancy, as shown in Figure 95.

Figure 95. BGP Route-Reflector Placement
Route reflectors are configured using the Route Reflector Policy configuration under Fabric Policies > Pod Policies using the Route Reflector Nodes configuration (not External Route Reflector Nodes).

**BGP Maximum Path**

As with any other deployment running BGP, it is good practice to limit the number of AS paths that Cisco ACI can accept from a neighbor. This setting can be configured under Tenant>Networking>Protocol Policies> BGP> BGP Timers by setting the Maximum AS Limit value.

**Transit Routing**

The transit routing function in the Cisco ACI fabric enables the advertisement of routing information from one L3Out to another, allowing full IP connectivity between routing domains through the Cisco ACI fabric. The configuration consists of specifying which of the imported routes from a L3Out should be announced to the outside through another L3Out, and which external EPG can talk to which external EPG. You specify this configuration through the definition of contracts provided and consumed by the external network under the L3Out.

To configure transit routing through the Cisco ACI fabric, you need to allow the announcement of routes either by configuring the route profiles (default export and default import) or by marking the subnets in question with the Export Route Control option when configuring external networks under the L3Out. An example is shown in Figure 96.

**Figure 96.** Export Route Control Operation

In the example in Figure 96, the desired outcome is for subnet 60.1.1.0/24 (which has been received from Router 1) to be advertised through the Cisco ACI fabric to Router 2. To achieve this, the 60.1.1.0/24 subnet must be defined on the second L3Out and allowed through a route profile. This configuration will cause the subnet to be redistributed from MP-BGP to the routing protocol in use between the fabric and Router 2.

It may not be feasible or scalable to define all possible subnets individually as export route control subnets. It is therefore possible to define an aggregate option that will mark all subnets for export. An example is shown in Figure 97.
In the example in Figure 97 there are a number of subnets received from Router 1 that should be advertised to Router 2. Rather than defining each subnet individually, the administrator can define the 0.0.0.0/0 subnet and set the Aggregate option. This option instructs the fabric that all transit routes should be advertised from this L3Out. Note that the Aggregate option does not actually configure route aggregation or summarization; it is simply a method to specify all possible subnets as exported routes.

In some scenarios, you may need to export static routes between L3Outs, as shown in Figure 98.

Supported Combinations for Transit Routing

Some limitations exist on the supported transit routing combinations through the fabric. In other words, transit routing is not possible between all possible routing protocols.

The latest matrix showing supported transit routing combinations is available at the following link:
Loop Prevention in Transit Routing Scenarios

When the Cisco ACI fabric advertises routes to an external routing device using OSPF or EIGRP, all advertised routes are tagged with the number 4294967295 by default. For loop-prevention purposes, the fabric will not accept routes inbound with the 4294967295 tag. This may cause issues in some scenarios where tenants and VRFs are connected together through external routing devices, or in some transit routing scenarios such as the example shown in Figure 99.

Figure 99. Loop Prevention with Transit Routing

In the example in Figure 99, an external route (30.1.0.0/16) is advertised in Cisco ACITenant 2, which is acting as a transit route. This route is advertised to the firewall through the second L3Out, but with a route tag of 4294967295. When this route advertisement reaches Cisco ACITenant 1, it is dropped due to the tag.

To avoid this situation, the default route tag value should be changed under the tenant VRF as shown in Figure 100.

Figure 100. Changing Route Tags
Route Summarization
Support for route summarization was introduced in Release 1.2(2) for BGP, EIGRP, and OSPF routing protocols. Summarization in Cisco ACI has the following characteristics:

- Route summarization occurs from the border leaf switches. Summaries are never carried inside the fabric.
- Summarization works for both tenant (bridge domain) routes and transit routes.
- Summary routes are installed in the routing table as routes to Null0.

Although there are some slight variations depending on the routing protocol in use, the general configuration method for route summarization is to configure a subnet entry in the External Networks section of the L3Out configuration. The configured subnet should be the actual summary address you wish to advertise. Additionally, the Route Summarization Policy (OSPF and BGP) or Route Summarization (EIGRP) option must be selected, along with the Export Route Control option.

The configurations for BGP, OSPF, and EIGRP summarization are shown in Figure 101, Figure 102, and Figure 103.

Figure 101. BGP Route Summarization Configuration

![BGP Route Summarization Configuration](image1)

Figure 102. OSPF Route Summarization Configuration

![OSPF Route Summarization Configuration](image2)
For BGP summarization, the AS-Set option can be configured. This option instructs Cisco ACI to include BGP path information with the aggregate route. If AS-Set is required, create a new BGP summarization policy, select the AS-Set option, and then associate this policy under the External Network configuration. Figure 104 shows the configuration of the AS-Set option under the BGP summarization policy.

Figure 104. BGP AS-Set Configuration

For OSPF route summarization, two options are available: external route summarization (equivalent to the summary-address configuration in Cisco IOS® Software and Cisco NX-OS Software) and inter-area summarization (equivalent to the area range configuration in Cisco IOS Software and NX-OS).

When tenant routes or transit routes are injected into OSPF, the Cisco ACI leaf node where the L3Oout resides is acting as an OSPF autonomous system boundary router (ASBR). In this case, the summary-address configuration (that is, external route summarization) should be used. This concept is shown in Figure 105.

Figure 105. OSPF Route Summarization Configuration
Figure 105. OSPF Summary-Address Operation

For scenarios where there are two L3Outs, each using a different area and attached to the same border leaf switch, the area range configuration will be used to summarize, as shown in Figure 106.

Figure 106. OSPF Area Range Operation

The OSPF route summarization policy is used to determine whether the summarization will use the area range or summary-address configuration, as shown in Figure 107.
Figure 107. OSPF Route Summarization

In the example in Figure 107, checking the Inter-Area Enabled box means that area range will be used for the summary configuration. If this box is unchecked, summary-address will be used.

Considerations for Multiple L3OutConnections
When configuring multiple connections from a border leaf, it is possible to use either a single L3Out connection or multiple L3Out connections. In some environments, it may be necessary to configure multiple L3Out connections in a single VRF (either with or without transit routing).

When deploying OSPF with a requirement for multiple networks, an administrator can choose to use either a single L3Out or separate L3Out instances for each connection.

An important point to consider is that the OSPF area is defined at the L3Out level. As a result, the following two rules apply:

- If you require the same border leaf to connect to multiple OSPF peer devices within the same area, you must use a single L3Out. It is not possible to configure multiple L3Out connections with the same OSPF area.
- If you require OSPF connections to two different areas from the same leaf node, separate L3Out connections must be used for this. Note that one of the L3Out connections must be part of area 0 in common with regular OSPF requirements.

Security Enforcement with Multiple L3Out Connections
External networks (also known as external EPGs) are used in L3Out configurations to define the external network destinations for the purposes of applying access controls (contracts). It is important to understand how this classification occurs and how this may affect security enforcement, particularly in an environment where multiple L3Out connections are associated with a single VRF and where overlapping external networks are configured.

Consider the example shown in Figure 108.
Figure 108. Security Enforcement with Multiple EPGs: Overlapping Subnet Classification

In this example, two L3Out connections are configured within the same VRF. The subnet 192.168.1.0/24 is accessible through one of the L3Out connections, and the subnet 172.31.1.0/24 is accessible through the other. From an Cisco ACI configuration perspective, both L3Out connections have an external network defined using the subnet 0.0.0.0/0. The desired behavior is to allow traffic between the Web EPG and the external network 192.168.1.0/24. Therefore, there is a contract in place permitting traffic between the Web EPG and L3Out 1.

This configuration has the side effect of also allowing traffic between the Web EPG and L3Out 2, even though no contract is configured for that communication flow. This happens because the classification takes place at the VRF level, even though external networks are configured under L3Out.

To avoid this situation, configure more specific subnets for the external EPGs under each L3Out, as shown in Figure 109.
Protocol Policies

Protocol policies are used to define various configuration parameters for a number of Layer 3 protocols, including BGP, EIGRP, OSPF, PIM, and IGMP. These policies can be configured on a per-tenant basis, as shown in Figure 110.
In many cases, common protocol policies will be required across all tenants and VRFs in the Cisco ACI fabric. In that case, protocol policies can be configured in the Common tenant and reused across all other tenants in the fabric.

**Bidirectional Forwarding Detection**

Cisco ACI Software Release 1.2(2g) added support for bidirectional forwarding detection (BFD). BFD is a software feature used to provide fast failure detection and notification to decrease the convergence times experienced in a failure scenario. BFD is particularly useful in environments where Layer 3 routing protocols are running over shared Layer 2 connections, or where the physical media does not provide reliable failure detection mechanisms.

In Cisco ACI, BFD is supported on L3Out interfaces only, where BGP, OSPF, EIGRP, or static routes are in use. BFD is not supported for fabric interfaces (that is, interfaces used to connect leaf and spine nodes together). BFD in Cisco ACI has the following characteristics:

- BFD Version 1 is used.
- Cisco ACI BFD uses asynchronous mode (that is, both endpoints send hello packets to each other).
- BFD is not supported for multihop BGP.

By default, a BGP global policy exists for both IPv4 and IPv6 sessions. The default timers specified in this policy have a 50-millisecond interval with a multiplier of 3, as shown in **Figure 111**.
This global default policy can be overridden if required by creating a new non default policy and assigning it to a switch policy group and then a switch profile.

BFD is also configurable on a per-tenant basis (under Networking >Protocol Policies) and will override the global BFD policy.

It is recommended to enable BFD on L3Out SVIs wherever possible to help ensure fast failure detection (assuming that the connected device supports it). For routed interfaces and sub interfaces, BFD may still be enabled, although physical interface mechanisms should ensure fast failure detection in most circumstances.

**L3Out Router ID Considerations**

When configuring a logical node profile under an L3Out configuration, you have to specify a router ID. An option exists to create a loopback address with the same IP address as that configured for the router ID.

It is recommended that the following best practices for L3Out router IDs be applied:

- Do not create a loopback interface with a router ID for OSPF, EIGRP, and static L3Out connections. This option is needed only for BGP when establishing BGP peering sessions from a loopback address.
- Create a loopback interface for BGP multihop peering between loopback addresses. It is possible to establish BGP peers sessions to a loopback address that is not the router ID. To achieve this, disable the Use Router ID as Loopback Address option and specify a loopback address that is different than the router ID.
- Each leaf switch should use a unique router ID. When configuring L3Out on multiple border leaf switches, each switch (node profile) should have a unique router ID.
- Use the same router ID value for all L3Out connections on the same node within the same VRF. Use of different router IDs is not supported, and a fault will be raised if different router IDs are configured for L3Out connections on the same node.
- A router ID for static L3Out connections must be specified even if no dynamic routing is used for the L3Out connection. The Use Router ID as Loopback Address option should be unchecked, and the same rules as outlined previously apply regarding the router ID value.

It is important to make sure that router IDs are unique within a routing domain. In other words, the router ID should be unique for each node within a VRF. The same router ID can be used on the same node within different VRFs. However, if the VRFs are joined to the same routing domain by an external device, then the same router ID should not be used in the different VRFs.
Summary of L3Out Design Recommendations
This list summarizes some of the design considerations when deploying a L3Out:

- Consider whether you want to build a L3Out connection using individual connections per tenant and VRF (VRF-lite) or with shared L3Out and MP-BGP EVPN VXLAN services.
- When you need to share an L3Out connection across multiple tenants, consider using a VRF instance from tenant Common and the associated L3Out without having to configure VRF for the individual tenants and VRF leaking.
- Each leaf switch should use a unique router ID for each VRF.
- Consider the use of an SVI with or without vPC for migration purposes from an existing network or in case you need to integrate with an L4-L7 device.
- Make sure you understand how to design the setup when using more than two border leaf switches with the same L3Out with vPC.
- Use the secondary IP address function to provide a shared gateway if you are using static routing to the fabric.
- When using multiple L3Outs for the same VRF instance, be sure to enter specific subnets in the External Networks Subnets (or External EPG) settings and not just 0.0.0.0/0.
- If you are configuring BGP peering, be sure to set the Max AS path limit.

Designing for Multitenancy and Shared Services
A common requirement of multitenant cloud infrastructures is the capability to provide shared services to hosted tenants. Such services include Active Directory, DNS, and filers. Figure 112 illustrates this requirement.

Figure 112. Shared Services Tenant
In Figure 112, Tenants 1, 2, and 3 have locally connected servers, respectively part of EPGs A, B, and C. Each tenant has an L3Out connection connecting remote branch offices to this data center partition. Remote clients for Tenant 1 need to establish communication with servers connected to EPG A. Servers hosted in EPG A need access to shared services hosted in EPG D in a different tenant. EPG D provides shared services to the servers hosted in EPGs A and B and to the remote users of Tenant 3.

In this design, each tenant has a dedicated L3Out connection to the remote offices. The subnets of EPG A are announced to the remote offices for Tenant 1, the subnets in EPG B are announced to the remote offices of Tenant 2, and so on. In addition, some of the shared services may be used from the remote offices, as in the case of Tenant 3. In this case, the subnets of EPG D are announced to the remote offices of Tenant 3.

Another common requirement is shared access to the Internet, as shown in Figure 113. In the figure, the L3Out connection of the Shared Services tenant (L3Out 4) is shared across Tenants 1, 2, and 3. Remote users may also need to use this L3Out connection, as in the case of Tenant 3. In this case, remote users can access L3Out 4 through Tenant 3.

Figure 113. Shared L3Out Connection
These requirements can be implemented in several ways:

- Use the VRF instance from the Common tenant and the bridge domains from each specific tenant.
- Use the equivalent of VRF leaking (which in Cisco ACI means configuring the subnet as shared).
- Provide shared services with outside routers connected to all tenants.
- Provide shared services from the Shared Services tenant by connecting it with external cables to other tenants in the fabric.

The first two options don’t require any additional hardware beyond the Cisco ACI fabric itself. The third option requires external routing devices such as additional Cisco Nexus 9000 Series Switches that are not part of the Cisco ACI fabric.

If you need to put shared services in a physically separate device, you are likely to use the third option.

The fourth option, which is logically equivalent to the third one, uses a tenant as if it were an external router and connects it to the other tenants through loopback cables.

If you have a specific constraint that make the first two options not viable, but if you don’t want to have an additional router to manage, then most likely you will want to use the fourth option.

**Inter-Tenant and Inter-VRF Communication**

In a Cisco ACI fabric, it is possible to configure communication between tenants, as well as communication between VRFs within a tenant, using the constructs available within the fabric (that is, avoiding the use of an external routing or security device to route between tenants and VRFs). This approach is analogous to VRF route leaking within a traditional routing and switching environment.

The example in Figure 114 shows a scenario where communication must occur between two EPGs across different VRFs within the same tenant.

**Figure 114.** Inter-VRF Communication
In the scenario in Figure 114, EPG 1 is providing a contract, which EPG 2 is consuming. The following are main points about the configuration of inter-VRF communication:

1. The scope of the contract used for the inter-VRF communication must be set to either Tenant or Global.
2. On the provider side, in addition to defining the subnet under the bridge domain, you need to enter the subnet (or a more specific subnet such as a /32) for VRF leaking under the EPG and mark it as Shared Between VRFs. You should also set the option No Default SVI Gateway.
3. On the consumer side, the subnet should be configured under the bridge domain marked as Shared Between VRFs.

In order for inter-VRF (and inter-tenant) traffic to flow, two factors must be addressed. First, routes must be leaked between the two VRFs in question. Second, the fabric must allow the communication based on the class-ID field carried in the VXLAN header. The class-ID normally has a locally significant value, but in certain configurations, such as with VRF-to-VRF traffic, Cisco ACI must use a global class-ID that is unique in the fabric.

Both of these factors are controlled using the contract object. When a consumer EPG is attached to a contract, the bridge domain subnet of that consumer EPG will automatically be leaked to the provider EPG’s VRF. For the provider-side subnet to be leaked to the consumer VRF instance, the same subnet as the bridge domain or a more specific one must also be configured at the provider EPG level and marked as shared. In the example, the provider EPG is configured with the exact IP address of the endpoint providing the shared service. Note that even if the VRFs are set to Unenforced mode, you will still need to configure a contract between provider and consumer EPGs in order for route leaking to occur.

The second example (shown in Figure 115) is for a scenario where communication between VRFs residing in different tenants is required.

Figure 115. Inter-Tenant Communication
In the scenario shown in Figure 115, the main difference from the inter-VRF example is that a global contract must be exported from Tenant A. On the EPG configuration within Tenant B, the contract is added as a consumed contract interface, selecting the contract that was previously exported. All other configurations (subnet under EPG, bridge domain, etc.) are identical to the configurations shown in the inter-VRF example.

**Note** For the scenario in Figure 115 you could also configure the contract in Tenant common and set it with a "global" scope.

**Configuration of the Subnet: When to Enter the Subnet Under the EPG**

One consideration that has to be made with shared services in Cisco ACI is when to configure the subnet (default gateway) under the EPG.

The general guidance is that a subnet used as the default gateway for servers should always be configured at the bridge domain level. This section aims to clarify the purpose of placing subnets under the EPG.

Cisco ACI optimizes route leaking between VRF instances by leaking the routes of the provider side only for EPGs that provide shared services. All subnet routes of the consumer-side bridge domain are instead leaked to the provider-side VRF instance.

A bridge domain can contain multiple subnets, so Cisco ACI must know which of the provider-side subnets should be leaked to the consumer VRF instance. For this optimization to occur, the subnets or the /32 that you enter on the provider-side EPG is leaked on the consumer-side VRF instance.

The definition of a subnet under the provider-side EPG is used only for the purpose of VRF leaking. You should configure this subnet **not** to provide the default gateway function by selecting the option No Default SVI Gateway. The subnet defined under the bridge domain is the default gateway for the servers on the provider-side EPGs.

In presence of VRF leaking, the classification information of which endpoint belongs to which EPG must be carried across VRFs. In order to optimize resource usage, Cisco ACI looks up traffic in the policy CAM table with the scope set to the consumer-side VRF only. This means that traffic filtering for provider EPG to consumer EPG and for the opposite direction happens in the context of the consumer-VRF. The classification information of the endpoints that belong to the provider-side VRF is then based on the subnet information that you enter in the provider-side EPGs.

The subnet defined on the provider-side EPG should be non overlapping with other subnets defined in the EPGs in the same bridge domain because the IP address specified in the EPG is used to derive the destination class ID when cross-VRF forwarding is performed.

To configure route leaking between any two tenants or VRF instances, Cisco ACI requires you to configure a contract with a scope of Tenant or Global (that is, not just VRF), and you should define subnets under the provider-side EPGs in addition to (or in alternative to) the bridge domain.

**Figure 116** Illustrates this configuration.
The subnet 20.20.20.6/32 defined under the provider EPG is configured as shared. The default gateway for the server 20.20.20.6 is the bridge domain subnet 20.20.20.1.

**Note:** If that subnet must also be announced to an L3Out connection, it should also be configured as advertised externally.

You need to make sure that all EPGs in VRF2 use disjoint subnets. For instance, if EPG2 is defined with 20.20.20.1/24 as a subnet, another EPG, such as EPG3 under VRF2, cannot also use 20.20.20.1/24. Otherwise, classification optimization cannot be performed. When traffic from the consumer-side VRF is destined to endpoints in the provider-side VRF with address in the 20.20.20.xrange, Cisco ACI would not know which provider-EPG they need to be associated with because all EPGs from the provider VRF would share the same subnet.

**Summary of Best Practices for Shared Services**

For shared services VRF configurations that provide the greatest scalability, follow these guidelines:

- Create a contract with global scope.
- Configure subnets under the bridge domain for both the provider and consumer sides and mark the consumer-side subnet as shared.
- In the provider-side EPG, enter the subnets or /32 that represents the server IP addresses that will be connected to this EPG. Configure these subnets or /32 as shared. Make sure that these subnets are configured **not** to provide the default gateway function.
- The subnets defined under the provider-EPGs that share the same VRF must be disjointed and not overlap.
- Configure a contract between the provider-side EPG (bridge domain and VRF) and the consumer-side EPG (bridge domain and VRF).
Shared L3Out Connections

It is a common approach for each tenant and VRF residing in the Cisco ACI fabric to have its own dedicated L3Out connection. However, an administrator may wish to use a single L3Out connection that can be shared by multiple tenants within the Cisco ACI fabric. This allows a single L3Out connection to be configured in a single, shared tenant (such as the Common tenant), with other tenants on the system sharing this single connection, as shown in Figure 117.

Figure 117. Shared L3Out Connections

A shared L3Out configuration is similar to the inter-tenant communication discussed in the previous section. The difference is that in this case, the routes are being leaked from the L3Out connection to the individual tenants, and vice versa. Contracts are provided and consumed between the L3Out connection in the shared tenant and the EPGs in the individual tenants.

To set up a shared L3Out connection, the L3Out connection can be defined as usual in the shared tenant (this tenant can be any tenant, not necessarily the Common tenant). The external network should be defined as usual. However, it should be marked with Shared Route Control Subnet and Shared Security Import Subnet. This means that the routing information from this L3Out connection can be leaked to other tenants, and subnets accessible through this L3Out connection will be treated as external EPGs for the other tenants sharing the connection (Figure 118).

Further information about these options follows:

- **Shared Route Control Subnet**: This option indicates that this network, if learned from the outside through this VRF, can be leaked to other VRFs (assuming that they have a contract with the external EPG).

- **Shared Security Import Subnets**: This option defines which subnets learned from a shared VRF belong to this external EPG for the purpose of contract filtering when establishing a cross-VRF contract. This configuration matches the external subnet and masks out the VRF to which this external EPG and L3Out connection belongs. This configuration requires that the contract filtering be applied at the border leaf.
In the example in Figure 119, the Aggregate Shared Routes option is checked. This means that all routes will be marked as Shared Route Control (in other words, all routes will be eligible for advertisement through this shared L3Out connection).

At the individual tenant level, subnets defined under bridge domains should be marked as both Advertised Externally and Shared Between VRFs, as shown in Figure 119.

Note: If you use vzAny on a VRF (for example, VRF1) to reduce the policy CAM consumption, be aware that vzAny also includes the Layer 3 external EPG of the L3Out connection of VRF1. As a result, if the vzAny of a VRF (VRF1) is a consumer of an EPG of a different VRF (VRF2), the EPG subnets of the second VRF (VRF2) are also announced to the L3Out connection of VRF1.
Layer 3 EVPN WAN Connectivity

Layer 3 EVPN over Fabric WAN (for more information, see the link https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-736899.html) is another option for external Layer 3 connectivity. It was introduced in Cisco ACI Release 2.0.

This option uses a single BGP session from the Cisco ACI spine switches to the external WAN device. All tenants are able to share this single connection, which dramatically reduces the number of tenant L3Out connections required. Layer 3 EVPN over fabric WAN is shown in Figure 120.

Figure 120. Layer 3 EVPN over Fabric WAN

Note that Layer 3 EVPN connectivity differs from regular tenant L3Out connectivity in that the physical connections are formed from the spine switches rather than leaf nodes. Layer 3 EVPN requires an L3Out connection to be configured in the Infra tenant. This L3Out connection will be configured to use BGP and to peer with the external WAN device. The BGP peer will be configured for WAN connectivity under the BGP peer profile, as shown in Figure 121.

Figure 121. BGP WAN Connectivity Configuration
The L3Out connection must also be configured with a provider label. Each individual tenant will then be configured with a local L3Out connection configured with a consumer label, which will match the provider label configured on the Infra L3Out connection.

In a small to medium-size environment (up to a few hundred tenants) and in environments where the external routing devices do not support EVPN connectivity, it is recommended to use individual L3Out connections per tenant. This is analogous to the use of VRF-lite in traditional environments, where each routed connection is trunked as a separate VLAN and sub interface on the physical links from the leaf nodes to the external device.

In a larger environment the recommended approach is to use Layer 3 EVPN over Fabric WAN to achieve multitenant Layer 3 connectivity from the fabric. This approach provides a greater scale and is also preferred over the shared L3Out approach described earlier in this document.

Scalability Considerations
The scalability factors of a Cisco ACI fabric include the following:

- The number of leaf switches and ports supported (read tested); this value is mainly a control-plane factor
- The number of EPGs supported
- The number of contracts supported (or the number of filter entries in the policy TCAM)
- The number of VRF instances supported per leaf and globally

Cisco publishes the numbers for each of these factors that have been validated by Quality Assurance (QA) testing on the page that contains the configuration guide for a given release. See https://www.cisco.com/c/en/us/support/cloud-systems-management/application-policy-infrastructure-controller-apic/tsd-products-support-series-home.html#ReleaseNotes.

At the time of this writing, the verified scalability limits for Release 2.3 are published at this link: https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/2-x/verified_scalability/b_Verified_Scalability_2_3_1x_and_12_3_1x.html.

Resolution and Deployment Immediate On-Demand Options
You can significantly reduce the consumption of hardware resources by setting the EPG Resolution and Deployment Immediate modes as On-Demand. As described in the section “Resolution and Deployment Immediate of VRFs, Bridge Domains, EPGs, and Contracts,” this setting helps ensure that if no virtual machine is attached to a given EPG on a leaf node, the VRF, bridge domain, EPG, and contracts related to this EPG are not installed on the hardware of this leaf node.

**Note:** The On-Demand option works only when the APIC can communicate with the VMM, so if all the APICs in a cluster are absent, a virtual machine that changes the virtualized host and the leaf node may not have connectivity after migration unless other virtual machines connected to the same EPG were already present on the destination leaf node.

Policy CAM Consumption
Contracts and filters consume space in the policy CAM. The size of the policy CAM depends on the Cisco Nexus 9000 Series Switch hardware that you are using, but you can optimize the consumption of the entries in this table with features such as the use of the deployment immediacy option On Demand and of vzAny.
If you are using first-generation leaf switches, you should limit the use of Layer 4 operation ranges because after Layer 4 operators that implement the range are used, the entries have to be expanded. This recommendation doesn’t apply to the Cisco Nexus 9300-EX and 9300-FX platform switches. With these Cisco ACI leaf switches, you can use Layer 4 operation ranges without having to take into account the expansion of Layer 4 rules.

Using Cisco Nexus 9300-EX and 9300-FX platform switches also increases the number of policy CAM rules available, as you can read in the verified scalability limit guide, under the entry “Security TCAM Size.” The number of supported entries for ALE2-based switches is 61,000 per leaf.

You need to distinguish the use of contracts in two different scenarios:

- Contracts between an L3Out connection (Layer 3 external EPG) and an internal EPG: Starting from Cisco ACI Release 1.2, you can reduce the possibility that the border leaf TCAM becomes a bottleneck for policy filtering by distributing the filtering to the computing leaf switches. This configuration is called ingress filtering (or ingress policy enforcement direction) and it is configured at the VRF level.
- Contracts between EPGs: Consider using the vzAny feature when applicable. Consider using vzAny to match the established flag for TCP traffic.

The section “Ingress Versus Egress Filtering Design Recommendations” illustrated how to optimize the use of the policy CAM resources. You should keep ingress filtering configured at the VRF level unless there's a specific need to change this configuration.


**Capacity Dashboard**

The Cisco ACI GUI includes the Capacity Dashboard, which is a tool for monitoring the current use of fabric resources (Figure 122). With this tool, you can see the total number of endpoints, bridge domains, VRF instances, and EPGs in the fabric compared to the maximum as well as the consumption of resources per leaf.


**Figure 122. Capacity Dashboard**
Best Practices Summary

This section summarizes some of the best practices presented in this document and provides a checklist you can use to verify configuration settings before deploying a Cisco ACI fabric:

- Physical design of the fabric: Consider from the beginning how you want to organize leaf nodes in vPC peers, and how you want to provide routed connectivity to the outside: with MP-BGP EVPN or VRF-lite.
- Physical design of the fabric: Consider whether you need to use a dedicated border leaf node or the border leaf should also be a computing leaf node.
- Controller design: Consider how many controllers you need based on scalability and high availability and be aware of how configuration and run-time data are saved and can be recovered.
- Fabric access design: Consider the choice of the infrastructure VLAN and of the TEP pool. Consider the use of per-VLAN MCP to eliminate loops, and be sure that you understand how Spanning Tree Protocol interacts with the Cisco ACI fabric.
- Object configuration for multiple tenants: If you need to configure objects to be used by multiple tenants, you should configure them in the Common tenant, but make sure you understand how object names are resolved and the use of contracts with global scope.
- Tenant design with VRF: Consider whether you want to use VRF in the Common tenant, or whether you want a VRF per tenant. Make sure that you know how to choose between ingress and egress filtering on the VRF.
- Tenant design with bridge domain: When creating a bridge domain, be sure to associate the bridge domain with a VRF instance even if you intend to use the bridge domain only for Layer 2 switching. Make sure you understand how the mapping database functions with or without IP routing and how ARP optimizations work. Then tune the bridge domain accordingly. In most cases, you can optimize flooding by using hardware proxy, by keeping IP routing enabled, a Subnet defined and sometimes by not allowing ARP flooding. In the presence of active-standby NIC teaming or active-active NIC teaming (other than vPC) and in the presence of floating IP addresses, you may need to tune the bridge domain further.
- Bridge domain subnet: Define one subnet as primary. Do not configure more than one MAC address unless you need to do so for Layer 2 extension. The subnet used as the default gateway should always be configured under the bridge domain.
- Tuning Bridge Domains: be careful when changing bridge domain settings because several optimizations options can be disruptive. At the time of this writing changing the bridge domain configuration from hardware-proxy to unknown unicast flooding and vice-versa is disruptive, as it is enabling or disabling Limit IP Learning to Subnet.
- EPG and contracts: When creating an EPG, be sure to associate the EPG with the bridge domain. If the EPG is extended to the VMM, be sure to specify the VMM domain. If it is associated with a port through static binding, be sure to specify the physical domain too.
- EPG and contracts: When associating EPGs with bare-metal servers, use Access (untagged) as the access-port option with Cisco Nexus 9300-EX and 9300-FX platform switches. Otherwise, you should use Access (802.1p).
- EPG and contracts: If you plan to configure shared services using route leaking between tenants and VRF instances, you also need to enter subnets or /32 under the EPG that is the shared services provider, but be sure that each EPG subnet is not overlapping with the other EPGs.
• L3Out design: make sure you understand the interactions between L3Our, vPC, SVI encapsulation and routing in order to define correctly the L3Out configuration.

• Virtual infrastructure: Configure different vDSs for management connectivity and for virtual workload connectivity. Consider the Resolution and Deployment Immediacy On-Demand option for most workloads to limit the consumption of hardware resources. Consider the Resolution Immediacy Pre-Provision option for management connectivity.

• Virtual infrastructure: Consider the use of AVS for remote servers. Be sure to include the infrastructure VLAN in the attach entity profile for the AVS. Consider vPC for the uplink connectivity of the virtualized servers.

For More Information

For more information, please refer to https://www.cisco.com/go/aci.