



The bridge to possible

White paper
Cisco public

Cisco Application Centric Infrastructure Design Guide

Contents

Introduction	3
Cisco ACI building blocks	4
Physical topology	10
Fabric transport infrastructure design considerations	16
Cisco APIC design considerations	32
Designing the fabric “access”	36
Global configurations	56
Designing the tenant network	60
Default gateway (subnet) design considerations	83
Designing external Layer 3 connectivity	112
Best practices summary	137
For more information	138

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

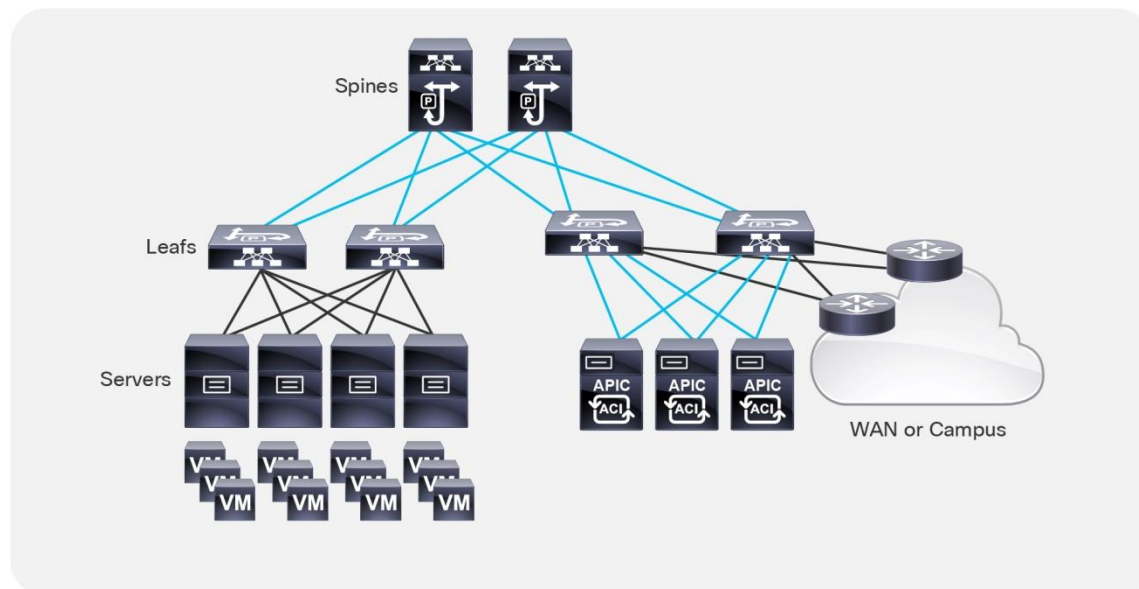


Figure 1.
Cisco ACI fabric

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

At the time of this writing, Cisco ACI Release 4.0 is available, and what is recommended in this design document is applicable to Cisco ACI fabrics running APIC Release 3.2 or newer with or without Virtual Machine Manager integration.

VMware ESXi hosts with VMware vSphere 6.x can be integrated with Cisco ACI either via physical domains (more on this later) or with VMM domains via either VMware vSphere Distributed Switch (vDS), Cisco Application Virtual Switch (AVS) and Cisco ACI Virtual Edge (AVE). This design guide does not include the integration with Cisco ACI Virtual Edge (AVE).

Note: For more information about the support matrix for Virtualization Products with Cisco ACI, please refer to the online documentation:

<https://www.cisco.com/c/dam/en/us/td/docs/Website/datacenter/aci/virtualization/matrix/virtmatrix.html>

For more information about the integration of Virtualization Products with Cisco ACI, please refer to https://www.cisco.com/c/en/us/support/cloud-systems-management/application-policy-infrastructure-controller-apic/tsd-products-support-series-home.html#Virtualization_-_Configuration_Guides

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
- Analytics support
- Multicast routing in the overlay
- Support for link-layer encryption
- Fibre Channel over Ethernet (FCoE) support

You can find the list of available leaf and spine switches at this URL:

<https://www.cisco.com/c/en/us/products/switches/nexus-9000-series-switches/models-comparison.html>

Cisco Nexus 9000 Series hardware

This section provides some clarification about the naming conventions used for the leaf and spines nodes referred to in this document:

- N9K-C93xx refers to the Cisco ACI leafs and the modular chassis
- N9K-X97xx refers to the Cisco ACI spine line cards

The trailing -E and -X signify the following:

- -E: Enhanced. This refers to the ability of the switch to classify traffic into endpoint groups (EPGs) based on the source IP address of the incoming traffic.
- -X: Analytics. This 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 to EPGs.
- -F: support for MAC Security

For port speeds, the naming conventions are as follows:

- G: 100M/1G
- P: 1/10-Gbps Enhanced Small Form-Factor Pluggable (SFP+)
- T: 100-Mbps, 1-Gbps, and 10GBASE-T copper
- Y: 10/25-Gbps SFP+
- Q: 40-Gbps Quad SFP+ (QSFP+)
- L: 50-Gbps QSFP28
- C: 100-Gbps QSFP28
- D: 400-Gbps QSFP-DD

You can find the updated taxonomy at this page:

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus9000/hw/n9k_taxonomy.html

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. A number of leaf-switch choices differ based on function:

- Port speed and medium type
- 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).
- Support for analytics: The newest leaf switches and spine line cards provide flow measurement capabilities for the purposes of analytics and application dependency mappings.

- Support for link-level encryption: The newest leaf switches and spine line cards provide line-rate MAC Security (MACsec) encryption.
- 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.
- Fibre Channel (FC) and Fibre Channel over Ethernet (FCoE): Depending on the leaf model, you can attach FC and/or 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.
- 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.
- Ability to change the allocation of hardware resources to support more Longest Prefix Match entries, or more policy CAM entries, or more IPv4 entries, etc. This concept is called “tile profiles,” and it was introduced in Cisco ACI 3.0. You can find more information at this link:
https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/kb/b_Cisco_APIC_Forwarding_Scale_Profile_Policy.pdf. You may also want to read the verified scalability guide:
https://www.cisco.com/c/en/us/support/cloud-systems-management/application-policy-infrastructure-controller-apic/tsd-products-support-series-home.html#Verified_Scalability_Guides

For more information about the differences between the Cisco Nexus® 9000 Series Switches, please refer to the following:

- <https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/datasheet-c78-738259.html>
- <https://www.cisco.com/c/en/us/products/switches/nexus-9000-series-switches/models-comparison.html>

Spine switches

The spine switches are available in several form factors. You can find the information about the differences between the Cisco Nexus fixed-form-factor spine switches at this link:

<https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/datasheet-c78-739886.html>.

The differences between these spine switches and line cards are as follows:

- Port speeds
- Line-card mode: Newer line cards 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.
- Support for link-level encryption and for CloudSec:

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/aci_multi-site/sw/2x/configuration/Cisco-ACI-Multi-Site-Configuration-Guide-201/Cisco-ACI-Multi-Site-Configuration-Guide-201_chapter_011.html#id_79312

- Support for Cisco ACI Multi-Pod and Multi-Site: Please refer to the specific documentation on Multi-Pod, Multi-Site, and release notes for more details.

Note: You can find information about Multi-Site hardware requirements at this link:

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/aci_multi-site/sw/2x/hardware-requirements/Cisco-ACI-Multi-Site-Hardware-Requirements-Guide-201.html

For more information about the differences between the Cisco Nexus 9500 platform module line cards, please refer to the following link:

<https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/datasheet-c78-732088.html>

The Cisco ACI fabric forwards traffic based on host lookups (when doing routing), 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.

You should use the verified scalability limits for the latest Cisco ACI release and see how many endpoints can be used per fabric:

https://www.cisco.com/c/en/us/support/cloud-systems-management/application-policy-infrastructure-controller-apic/tsd-products-support-series-home.html#Verified_Scalability_Guides

According to the verified scalability limits, the following spine configurations have these endpoint scalabilities:

- Max. 450,000 Proxy Database Entries with four (4) fabric line cards
- Max. 180,000 Proxy Database Entries with the fixed spine switches

The above numbers represent the sum of the number of MAC, IPv4, and IPv6 addresses; for instance, in the case of a Cisco ACI fabric with fixed spines, this translates into:

- 180,000 MAC-only EPs (each EP with one MAC only)
- 90,000 IPv4 EPs (each EP with one MAC and one IPv4)
- 60,000 dual-stack EPs (each EP with one MAC, one IPv4, and one IPv6)

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 (APIC)

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 interfaces for connectivity to the leaf switches. The APIC is also equipped with Gigabit Ethernet interfaces for out-of-band management.

For more information about the APIC models, please refer to this link:

<https://www.cisco.com/c/en/us/products/collateral/cloud-systems-management/application-policy-infrastructure-controller-apic/datasheet-c78-739715.html>

Note: A cluster may contain a mix of different APIC models: however, the scalability will be that of the least 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 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 line cards to connect to the intersite 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 the 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 does not 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 is supported in its software version. If the leaf is running an older version of software and it does not understand a feature, it will reject it; however, it will **not** raise a fault.

For more information about which configurations are allowed with a mixed OS version in the fabric, please refer to the following link: https://www.cisco.com/c/en/us/support/cloud-systems-management/application-policy-infrastructure-controller-apic/tsd-products-support-series-home.html#Software_and_Firmware_Installation_and_Upgrade_Guides

It is important to consider that running a Cisco ACI fabric with different software versions is meant to be just a temporary configuration to facilitate upgrades, and minimal or no configuration changes should be performed while the fabric runs with mixed OS versions.

Fabric Extenders (FEX)

It is possible to connect Fabric Extenders (FEX) to the Cisco ACI leafs; the main purpose of doing so should be to simplify migration from an existing network with Fabric Extenders. If the main requirement for the use of FEX is the Fast Ethernet port speeds, you should consider also the Cisco ACI leaf Cisco Nexus 9348GC-FXP, which was introduced as part of Cisco ACI 3.0.

FEX can be connected to Cisco ACI with what is known as a straight-through topology, and vPCs can be configured between hosts and FEX, but not between FEX and Cisco ACI leafs.

FEX can be connected to leaf front-panel ports as well as converted downlinks (since Cisco ACI Release 3.1).

FEX has many limitations compared to attaching servers and network devices directly to a leaf. The main ones are the following:

- No support for L3Out on FEX
- No Rate limiters support on FEX
- No Traffic Storm Control on FEX
- No Port Security support on FEX
- FEX should not be used to connect routers or L4-L7 devices with service-graph redirect
- The use in conjunction with microsegmentation works, but if microsegmentation is used Quality of Service (QoS) does not work on FEX ports because all microsegmented traffic is tagged with a specific class of service. Microsegmentation and FEX is a feature that at the time of this writing has not been extensively validated.

Support for FCoE on FEX was added in Cisco ACI Release 2.2:

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/1-x/release/notes/apic_m_221.html

When using Cisco ACI with FEX, you want to verify the verified scalability limits; in particular, the ones related to the number of ports multiplied by the number of VLANs configured on the ports (commonly referred to as P, V):

https://www.cisco.com/c/en/us/support/cloud-systems-management/application-policy-infrastructure-controller-apic/tsd-products-support-series-home.html#Verified_Scalability_Guides

With regard to scalability, you should keep in mind the following points:

- The total scale for VRFs, Bridge Domains (BDs), endpoints, and so on is the same whether you are using FEX attached to a leaf or whether you are connecting endpoints directly to a leaf. This means that, when using FEX, the amount of hardware resources that the leaf provides is divided among more ports than just the leaf ports.
- The total number of VLANs that can be used on each FEX port is limited by the maximum number of P,V pairs that are available per leaf for host-facing ports on FEX. As of this writing, this number is ~10,000 per leaf, which means that, with 100 FEX ports, you can have a maximum of 100 VLANs configured on each FEX port.
- At the time of this writing, the maximum number of encapsulations per FEX port is 20, which means that the maximum number of EPGs per FEX port is 20.
- At the time of this writing, the maximum number of FEX per leaf is 20.

Note: For more information about which leaf switch is compatible with which Fabric Extender, please refer to this link:

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus9000/hw/interoperability/fexmatrix/fex_tables.html

For more information about how to connect a Fabric Extender to Cisco ACI, please refer to this link:

<https://www.cisco.com/c/en/us/support/docs/cloud-systems-management/application-policy-infrastructure-controller-apic/200529-Configure-a-Fabric-Extender-with-Applica.html>

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.

As of Cisco ACI 3.2, the ACI fabric allows one tier of leafs only.

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.). 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 Multi-Pod 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 store all the endpoints-to-VTEP mapping entries (spine proxies).

Within a pod, 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

Up until Cisco ACI 3.1, uplink ports on leaf switches were hard-coded as fabric (iVXLAN) ports and could connect only to spine switches. Starting with Cisco ACI 3.1, you can change the default configuration and make ports that would normally be uplinks, be downlinks, or vice-versa. More information can be found at this link:

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_010011.html#id_60593

Note: For information about the optics supported by Cisco ACI leafs and spines, please use this tool: <https://tmgmatrix.cisco.com/home>

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.

VPCs can also be used to connect to existing Layer 2 infrastructure or for L3Out connections (vPC plus a 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; for example, you cannot have a vPC consisting of a 9372TX and -EX or -FX leafs.

Even if two leafs of different hardware generation are not meant to be vPC peers, the Cisco ACI software is designed to make the migration from one leaf to another compatible with vPC. Assume that the fabric has Cisco Nexus 9372PX switch leaf pairs (called, in the following example, 9372PX-1 and 9372PX-2), and they need to be replaced with Cisco Nexus 93180YC-EX leafs (called 93180YC-EX-1 and 93180YC-EX-2).

The insertion of newer leafs works like this:

- When 93180YC-EX-2 replaces 9372PX-2 in a vPC pair, 9372PX-1 can synchronize the endpoints with 93170YC-EX2.
- The vPC member ports on 93180YC-EX-2 stay down.
- If you remove 9372PX-1, the vPC member ports on 93180YC-EX-2 go up after 10 to 20s.
- 93180YC-EX-1 then replaces 9372PX-1, and 93180YC-EX-2 synchronizes the endpoints with 93180YC-EX-1.
- The vPC member ports on both 93180YC-EX-1 and 93180YC-EX-2 go up.

Note: You can find more information at the following link:

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/1-x/Operating_ACI/guide/b_Cisco_Operating_ACI/b_Cisco_Operating_ACI_chapter_01101.html#id_56210

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.

Choosing between VRF-lite and GOLF

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), or Border Gateway Protocol (BGP), as shown in Figure 2. 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 (also known as GOLF): 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 3. 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.

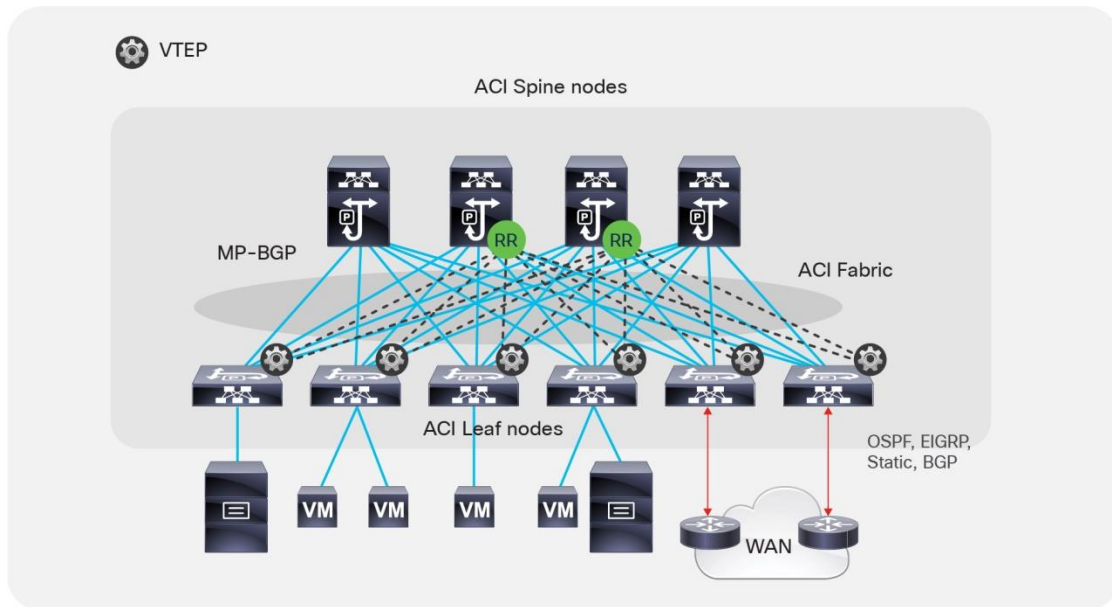


Figure 2.
Connectivity to the outside with VRF-lite (standard L3Out in Cisco ACI)

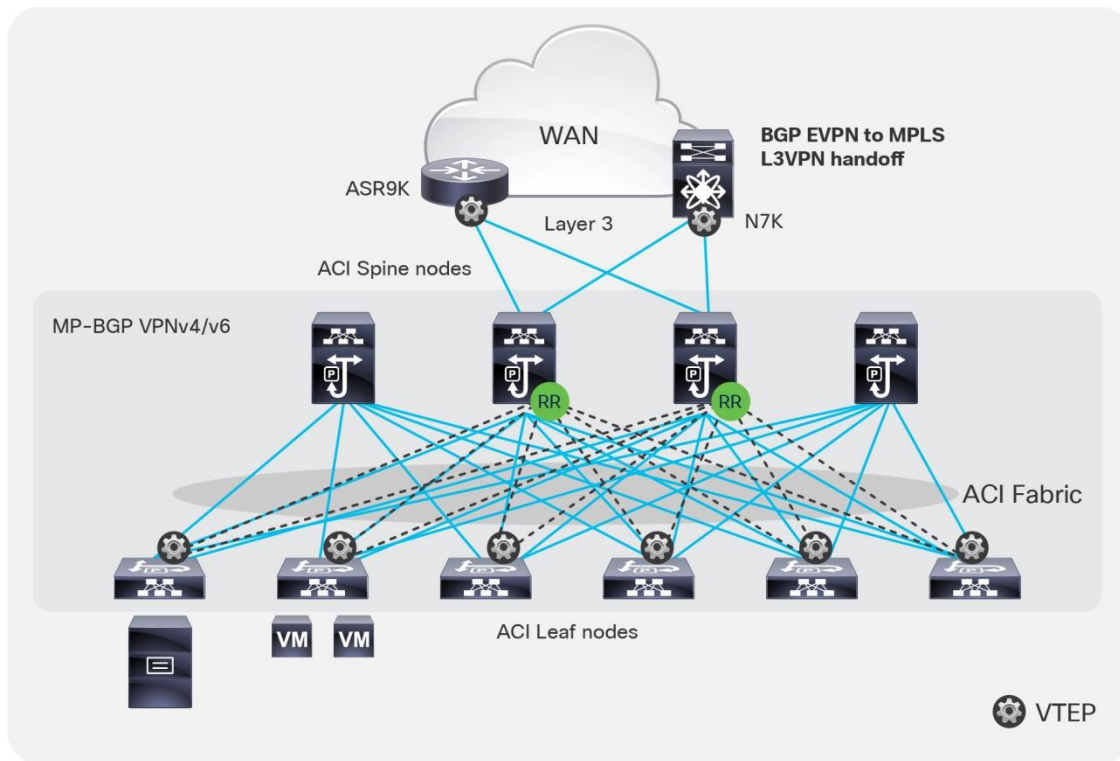


Figure 3.
Connectivity to the outside with Layer 3 EVPN services

The topology in Figure 2 works with any router connected to a leaf node. This topology is based on VRF-lite. The topology in Figure 3 requires that the WAN routers support MP-BGP EVPN, OpFlex protocol, and VXLAN. With the topology in Figure 3, the fabric infrastructure is extended to the WAN router, which effectively becomes the equivalent of a border leaf in the fabric.

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

- There is only one MP-BGP session for all VRFs and tenants.
- 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 the OpFlex protocol; for example, the autoprogramming of VRFs on GOLF routers. This feature is optional.
- VXLAN data plane handoff to GOLF router.

Note: With releases prior to Cisco ACI 4.0 the MP-BGP EVNP solution also offered the advantage of being able to announce host routes to the outside. This capability has also been added to the VRF-lite solution as of Cisco ACI 4.0:

<https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/4-x/L3-configuration/Cisco-APIC-Layer-3-Networking-Configuration-Guide-401.pdf>

The main disadvantages of using the MP-BGP EVPN solution at the time of this writing are:

- No support for multicast
- Limitations of VRF route sharing with “GOLF” VRFs

It is possible to start from the topology in Figure 2 and migrate later, if desired, to the topology in Figure 3.

Note: This design guide does not cover the details of extending the fabric using MP-BGP EVPN, OpFlex protocol, and VXLAN. For more information about GOLF, please refer to the following document:

<https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-736899.html>

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.

Using border leafs for server attachment

Attachment of endpoints to border leaf switches is 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, 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 could also configure the VRF instance for egress policy by selecting the Policy Control Enforcement Direction option Egress under Tenants > Networking > VRFs.

The recommendation at the time of this writing is that if you deploy a topology that connects to the outside through border leaf switches that are also used as computing leaf switches, and if the VRF instance is configured for ingress policy (which is the default), you should disable remote endpoint learning on the border leaf switches.

Depending on the Cisco ACI version, 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
- Or from System > System Settings > Fabric Wide Setting > Disable Remote EP Learning

Note: The Disable Remote EP Learn configuration option disables the learning of the mapping between remote endpoint IP addresses and VTEPs, and only on border leaf switches that have a VRF instance with ingress policy enabled. This configuration option does not change the learning of the MAC addresses of the endpoints on the local leaf. Disabling remote endpoint learning for the border leaves is compatible with Multicast Routing. The section titled “Disable Remote Endpoint Learning” provides additional information.

Do not use the L3Out to connect servers

Border leaf switches can be configured with three types of interfaces to connect to an external router:

- Layer 3 (routed) interface
- Subinterface with IEEE 802.1Q tagging
- Switch Virtual Interface (SVI)

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

The L3out is meant to attach routing devices. It is not meant to attach servers directly on the SVI of an L3Out. Servers should be attached to EPGs and Bridge Domains (BDs).

There are multiple reasons for this:

- The L2 domain created by an L3Out with SVIs is not equivalent to a regular bridge domain.
- The L3ext classification is designed for hosts multiple hops away.

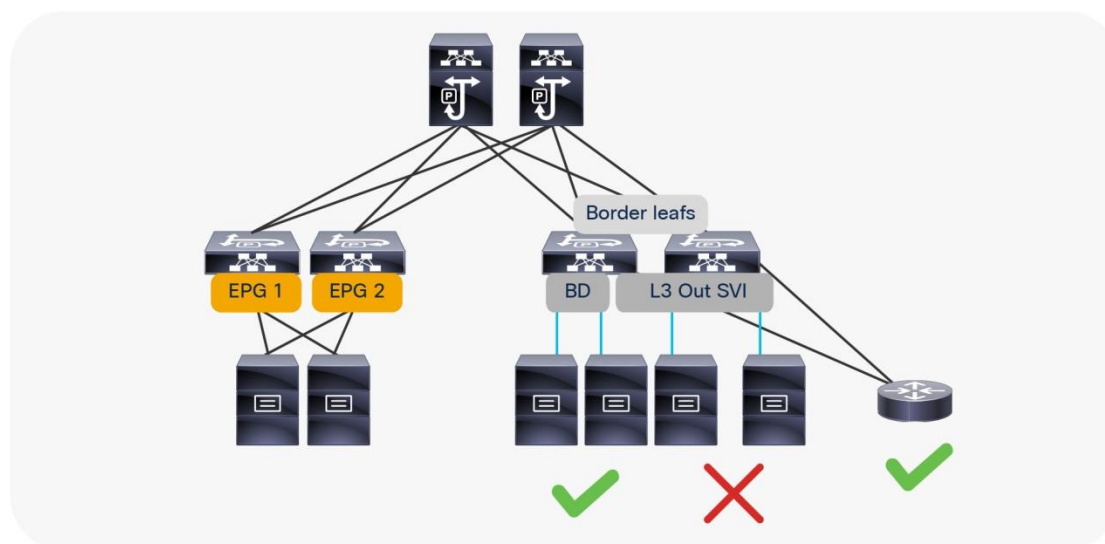


Figure 4.
Do not use the L3Out to connect servers

L3Out and vPC

You can configure static or dynamic routing protocol peering over a vPC for an L3Out without any special design considerations.

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 data centers, it makes sense to have leaf nodes dedicated to the connection of L4–L7 services.

For deployment of service graphs with the service redirect feature, dedicated service leaf nodes must be used if the leafs are first-generation Cisco ACI leaf switches. With Cisco Nexus 9300 EX and newer switches, you do not have to use dedicated leaf switches for the L4–L7 service devices for the service graph redirect feature.

Fabric transport infrastructure design considerations

Cisco ACI forwarding is based on a VXLAN overlay. Leaf nodes are virtual tunnel endpoints (VTEPs), which, in Cisco ACI terminology, are known as PTEPs (physical tunnel endpoints).

Cisco ACI maintains a mapping database containing information about where (that is, on which TEP) 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.

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 the tenant “infra”.

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.

When designing the infrastructure, you need to decide which VLAN to use to carry this traffic, and which IP address pool to use for TEP addressing:

- The infrastructure VLAN should not overlap with existing VLANs that may be in the existing network infrastructure to which you are connecting. The section “Infrastructure VLAN” provides more details.
- The TEP IP address pool should not overlap with existing IP address pools that may be in use by the servers. The section “TEP address pool” provides more details.

Choosing the leaf forwarding profile

When deploying the fabric you may want to define from the very beginning which forwarding profile is more suitable for the requirements of your data center.

The default profile configures the leaf for support of both IPv4 and IPv6 and Layer 3 multicast capacity. But if you plan to use Cisco ACI primarily as a Layer 2 infrastructure, the IPv4 profile with more MAC entries and no IPv6 entries may be more suitable. If, instead, you plan on using IPv6, the high dual-stack profile may be more suitable for you. Some profiles offer more capacity for the Longest Prefix Match table for designs where, for instance, Cisco ACI is a transit routing network, in which case the fabric offers less capacity for IPv4 and IPv6.

The profile configuration is done per leaf, so you can potentially define different scale profiles for leaves that are used for different purposes; for example, you may want to configure a leaf that is used as a dedicated border leaf with a bigger Longest Prefix Match table.

For more information about the configurable forwarding profiles, please refer to this link:

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/kb/b_Cisco_APIC_Forwarding_Scale_Profile_Policy.pdf

Fabric-id

When configuring a Cisco ACI fabric, you need to give a fabric-id to it. The fabric-id should not be confused with the pod-id or the site-id. You should just give “fabric-id 1,” unless there is some specific reason not to (for instance, if you plan to use GOLF with Auto-RT, and all sites belong to the same ASN).

Infrastructure VLAN

The APIC communicates with the Cisco ACI fabric through a VLAN that is associated with the tenant called infrastructure (which appears in the APIC User Interface as tenant “infra”). This VLAN is used for internal control communication between fabric nodes (leaf and spine nodes and APICs).

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.

From the GUI, you can see which infrastructure VLAN is in use, as in Figure 5. From the command-line interface, you can find the infrastructure VLAN; for instance, by using this command on a leaf:

```
leaf1# show system internal epm vlan all | grep Infra
```

System

Tenants

Fabric

Virtual Networking

L4-L7 Services

Admin

Operations

Apps

QuickStart

Dashboard

Controllers

System Settings

Smart Licensing

Faults

Config Zones

Events

Audit Log

Active Sessions

Controllers

Quick Start

Topology

Controllers

apic2 (Node-2)

apic3 (Node-3)

apic-a1 (Node-1)

Cluster as Seen by Node

Interfaces

Storage

NTP Details

Equipment Fans

Power Supply Units

Equipment Sensors

Memory Slots

Processes

Interfaces

eth1-1	1500	58:F3:9C:F7:A2:70	up
eth1-2	1500	58:F3:9C:F7:A2:70	down
eth2-1	1500	F0:7F:06:3E:E4:13	up
eth2-2	1500	F0:7F:06:3E:E4:13	down

Aggregated Interfaces

Name	MTU	MAC	Associated Physical Interfaces	Active Interface
bond0	1500	F0:7F:06:3E:E4:13	eth2/1, eth2/2	eth2/1
bond1	1500	58:F3:9C:F7:A2:70	eth1/1, eth1/2	eth1/1

L3 Management Interfaces

Name	MTU	MAC	Encap
bond0.4093	1496	F0:7F:06:3E:E4:13	vlan-4093
bond1	1500	58:F3:9C:F7:A2:70	unknown

Figure 5.
Bond and infrastructure VLAN on the APIC

The infrastructure VLAN is also used to extend the Cisco ACI fabric to another device. For example, when using Cisco ACI with Virtual Machine Manager (VMM) integration, the infrastructure VLAN can be used by AVE or AVS to send DHCP requests and get an address dynamically from the Cisco ACI fabric TEP pool and to send VXLAN traffic.

In a scenario in which the infrastructure VLAN is extended beyond the Cisco ACI fabric (for example, when using AVS, AVE, OpenStack integration with OpFlex protocol, or Hyper-V integration), this VLAN may need to traverse other (that is, not Cisco ACI) devices, as shown in Figure 6.

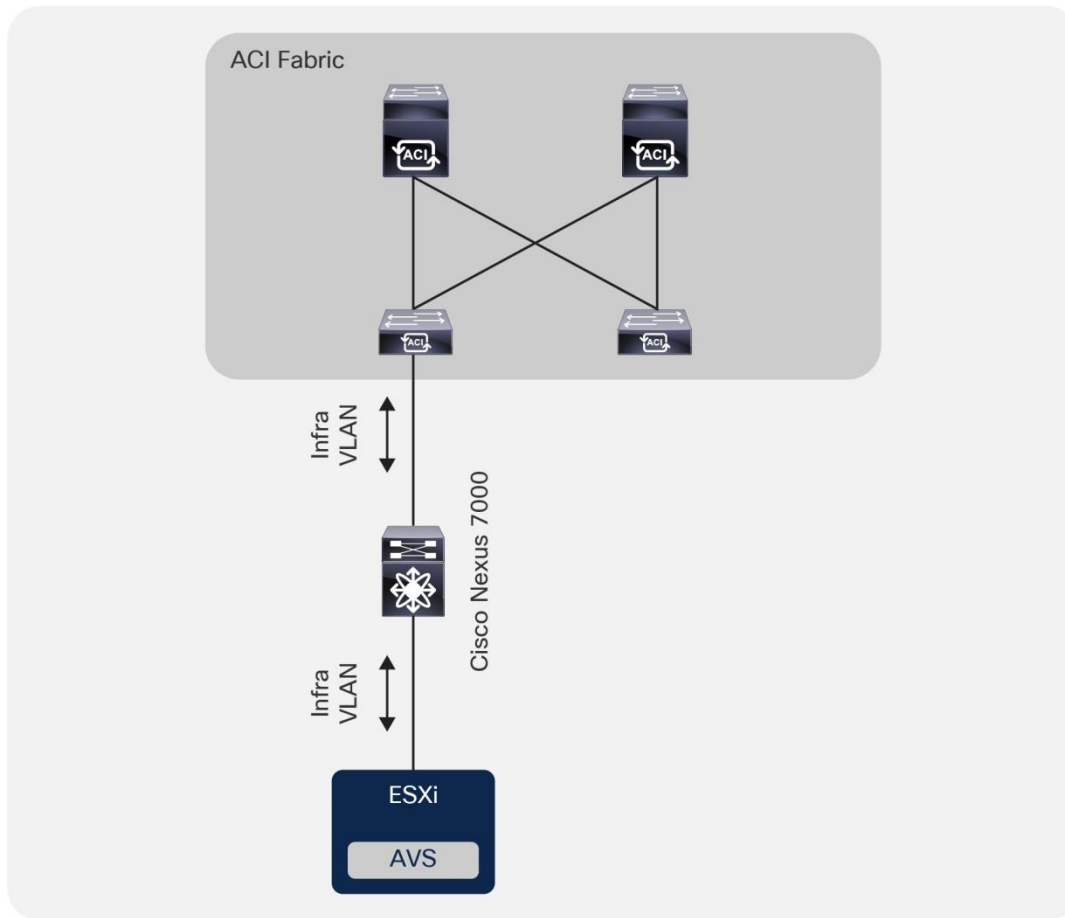


Figure 6.
Infrastructure VLAN considerations

Some platforms (for example, Cisco Nexus 9000, 7000, and 5000 series switches) reserve a range of VLAN IDs: typically 3968 to 4095.

In Cisco UCS, the VLANs that can be reserved are the following:

- FI-6200/FI-6332/FI-6332-16UP/FI-6324: 4030–4047. Note that vlan 4048 is being used by vsan 1.
- FI-6454: 4030–4047 (fixed), 3915–4042 (can be moved to a different 128 contiguous block vlan but requires a reboot).

https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/ucs-manager/GUI-User-Guides/Network-Mgmt/3-1/b_UCSM_Network_Mgmt_Guide_3_1/b_UCSM_Network_Mgmt_Guide_3_1_chapter_0110.html

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 a VLAN < 3915.

Note: In order to enable the transport of the infrastructure VLAN on Cisco ACI leaf ports, you just need to select the checkbox in the Attachable Access Entity Profile (AAEP) that is going to be associated with a given set of ports.

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. Link Layer Discover Protocol (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.

Figure 7 shows how bootup and autoprovisioning 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.

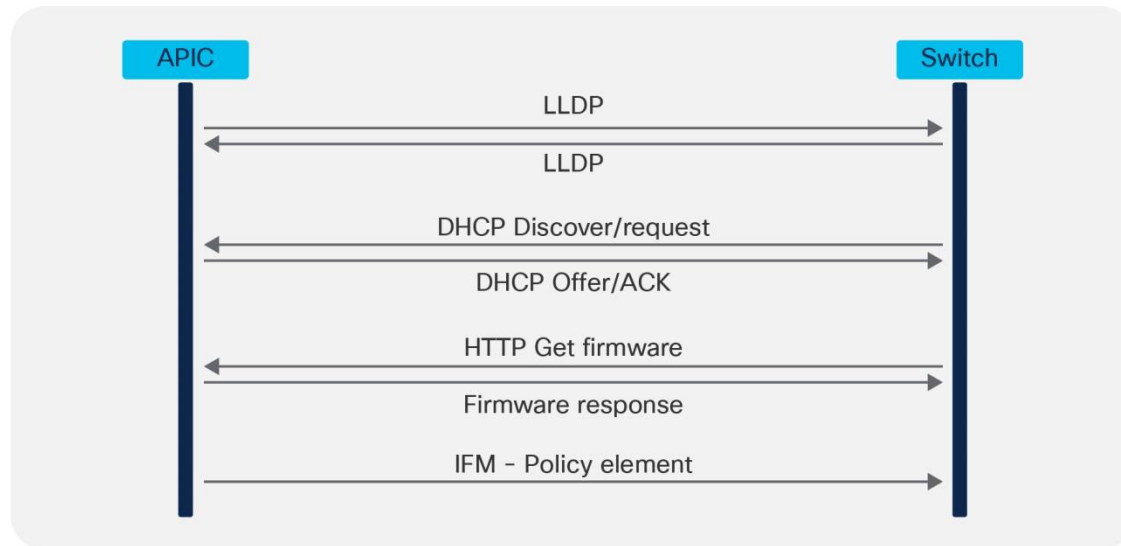


Figure 7.
Leaf or spine bootup sequence

Although TEPs are located inside the fabric, there are some scenarios where the TEP range may be extended beyond the fabric. As an example, when you use AVE, fabric TEP addresses are allocated to the virtual switch. Therefore, it is not advisable to use overlapping addresses between the internal TEP range and the external network in your data center.

The number of addresses required for the TEP address pool depends on a number of factors, including the following:

- Number of APICs
- Number of leaf and spine nodes
- Number of Application Virtual Switches (AVSs), AVE instances, Hyper-V hosts or, more generally, virtualized hosts managed via VMM integration and integrated with OpFlex
- Number of vPCs required

Note: In this calculation, you do not need to include the count of nodes of a different pod because each pod uses its own TEP pool that should not overlap with other pod pools:

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

To avoid issues with address exhaustion in the future, it is strongly recommended to allocate 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, because it cannot be easily modified later.

You can verify the TEP pool after the initial configuration by using the following command:

```
Apic1# moquery -c dhcpPool
```

When planning for the TEP pool, you should also think about potential future use of the following Cisco ACI features, including Multi-Pod, Multi-Site, Remote Leaf, and vPOD.

The following considerations apply:

- Cisco ACI Multi-Pod: In order to count the TEP pool range, you do not need to include the count of nodes of a pod other than the one you are configuring, because each pod uses its own TEP pool that should not overlap with other pod pools:
<https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-737855.html>
- On the other hand, you need to make sure the pool you define is nonoverlapping with other existing or future pods. There is no strict requirement for the TEP pool to be externally routable for Cisco ACI Multi-Pod.
- Cisco ACI Multi-Site: In Multi-Site, the configuration requires the use of specific, publicly routable TEP addresses that are completely independent from the TEP pool: the Control-Plane External Tunnel Endpoint (one per spine connected to the Inter-Site Network), the Data-Plane ETEP (one per Site per pod) and the Head-End Replication ETEP (one per Site). Quoting <https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-739609.pdf>: “The TEP pool prefixes used within each site do not need to be exchanged across sites to allow intersite communication. As a consequence, there are no technical restrictions regarding how those pools should be assigned. However, the strong recommendation is not to assign overlapping TEP pools across separate sites so that your system is prepared for future functions that may require the exchange of TEP pool summary prefixes.”

For Remote Leafs and vPOD information, please refer to:

<https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-740861.html> and https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/aci_vpod/installation-upgrade/4-x/Cisco-ACI-Virtual-Pod-Installation-Guide-401.pdf respectively.

BGP Route Reflector policy

Routing in the infrastructure VRF is based on IS-IS. Routing within each tenant VRF is based on host routing for endpoints that are directly connected to the Cisco ACI fabric.

Cisco ACI uses MP-BGP VPNv4/VPNv6 to propagate external routes within the ACI. 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 of the pods in a Cisco ACI Multi-Pod deployment).

The BGP Route Reflector Policy controls whether MP-BGP runs within the fabric and which spine nodes should operate as BGP reflectors.

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 (Multi-Pod).

To enable and configure MP-BGP within the fabric, modify the **BGPRouteReflectordefault** 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 8.

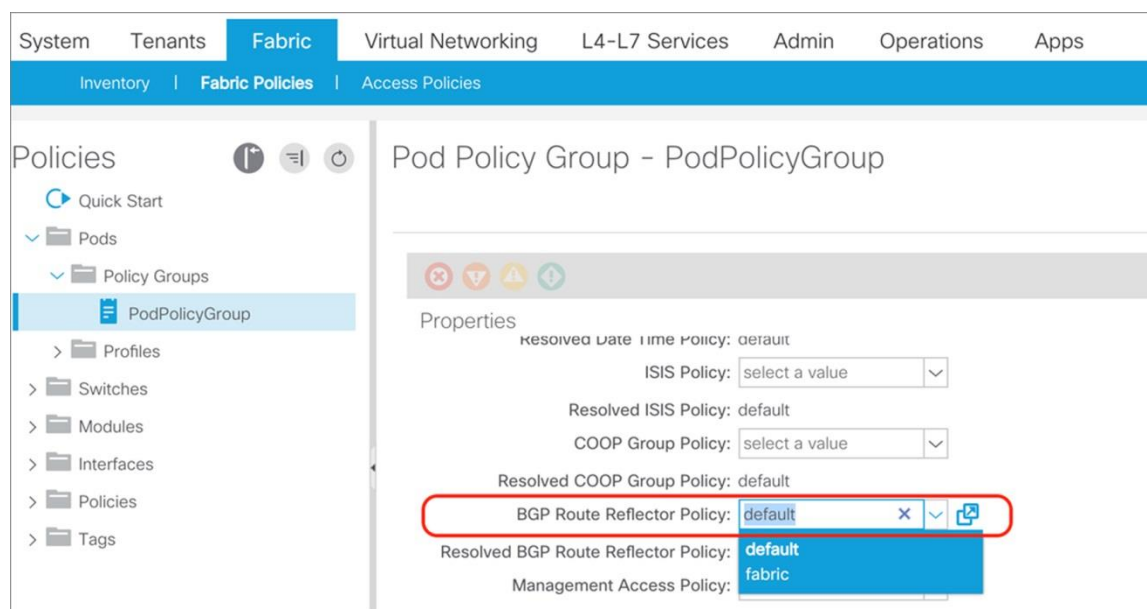


Figure 8.
BGP Route Reflector configuration

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 9 illustrates the routing protocol within the Cisco ACI fabric and the routing protocol between the border leaf and external router using VRF-lite.

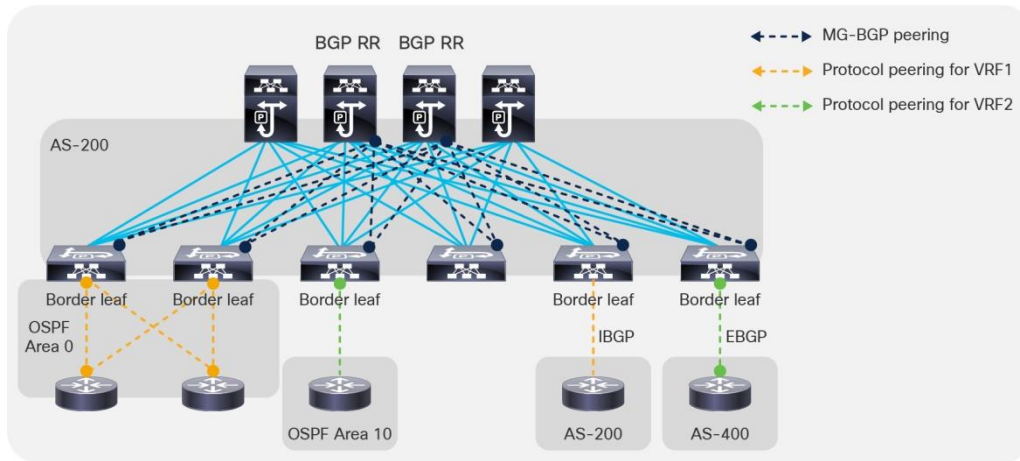


Figure 9.
Routing distribution in the Cisco ACI fabric

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.

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 Multi-Pod 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 10.

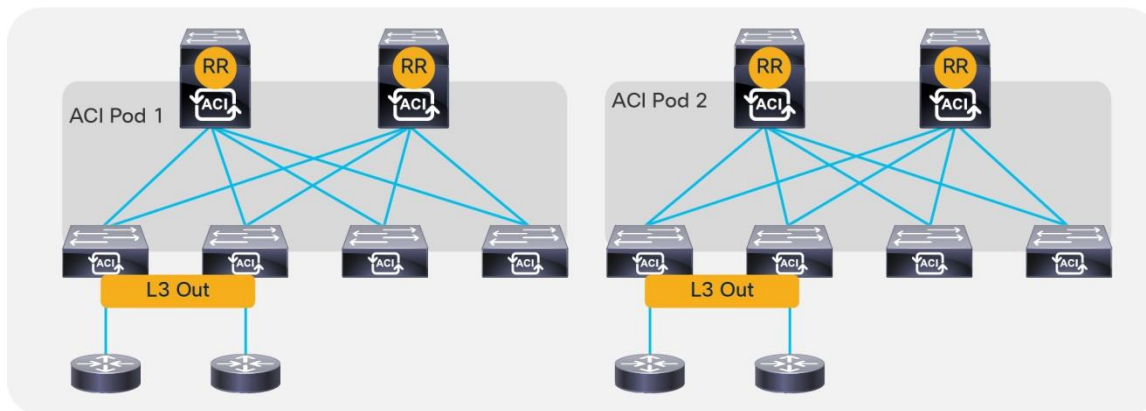


Figure 10.
BGP route-reflector placement

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 per tenant under Tenant > Networking > Protocol Policies > BGP > BGP Timers by setting the Maximum AS Limit value.

Network Time Protocol (NTP) configuration

Make sure to configure the fabric for Network Time Protocol (NTP), Figure 11 illustrates where to configure NTP.

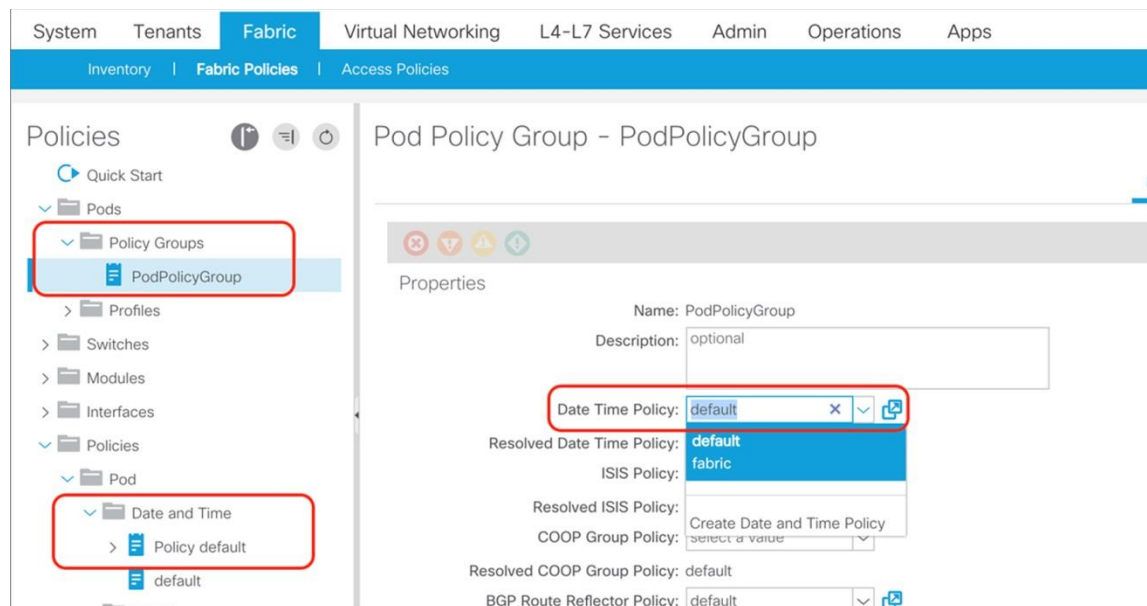


Figure 11.
NTP configuration

You can find more information about the configuration of NTP at this link:

<https://www.cisco.com/c/en/us/support/docs/cloud-systems-management/application-policy-infrastructure-controller-apic/200128-Configuring-NTP-in-ACI-Fabric-Solution.html>

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 (which at the time of this writing can be found under System Settings, COOP group or with older releases under Fabric Policies, Pod Policies) controls the authentication of COOP messages. Two modes are available: Compatible Mode and Strict Mode. Compatible Mode accepts both authenticated and nonauthenticated connections, provides backward compatibility, and is the default option. Strict Mode allows MD5 authentication connections only. The two options are shown in Figure 12.

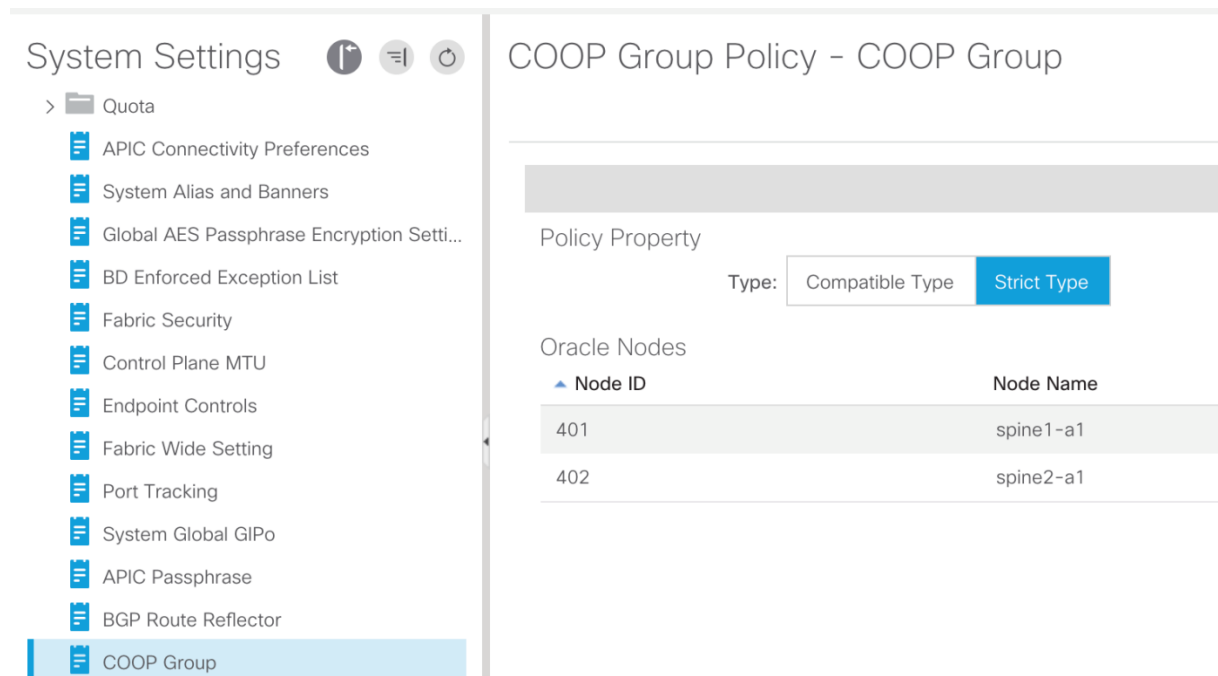


Figure 12.
COOP Group Policy

It is recommended that you enable Strict Mode in production environments to help ensure the most secure deployment.

IS-IS metric for redistributed routes

It is considered a good practice to change the IS-IS metric for redistributed routes to lower than the default value of 63. This is to ensure that, when (for example) a spine is rebooting because of an upgrade, it is not in the path to external destinations until the entire configuration of the spine is completed, at which point the metric is set to the lower metric; for example, 32.

This configuration can be performed from Fabric/Fabric Policies/Policies/Pod/ISIS Policy default.

Maximum transmission unit

Figure 13 shows the format of the VXLAN encapsulated traffic in the Cisco ACI fabric.

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 frame-check sequence [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).

The VXLAN header used in the Cisco ACI fabric is shown in Figure 13.

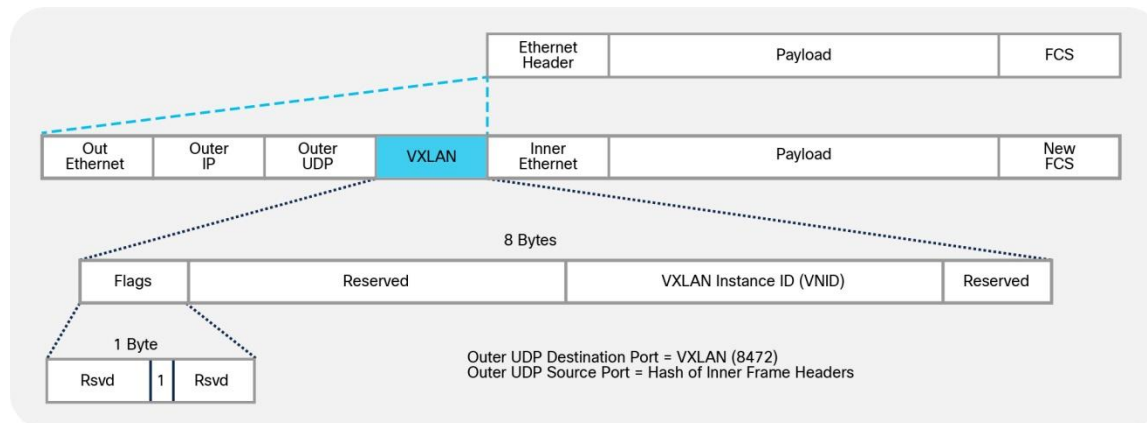


Figure 13.
VXLAN header

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 with the MTU of the incoming packet (which is set by default to 9000 Bytes) + 150 bytes.

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.

You normally do not need to change the MTU defaults of a Cisco ACI fabric, but in case you do, this can be done from: Fabric > Fabric Polices > Policies > Global > Fabric L2 MTU Policy. This MTU refers to the payload of the VXLAN traffic. Starting with Cisco ACI Release 3.1(2), this can be changed to 9216 bytes; the setting takes effect when you configure EPG binding to a port.

Starting with Cisco ACI 3.1(2), the ACI uplinks have an MTU of 9366 bytes (9216 + 150).

If the VXLAN overlay must be carried across an IPN, you need to make sure that the MTU on the L3Out is configured correctly.

For more information about the MTU configuration with Multi-Pod, please refer to this document:

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

Configuring the fabric infrastructure for faster convergence

Fast Link Failover

Starting with Cisco ACI Release 3.1, it is possible to get better failover convergence for fabric link failures by enabling a feature called LBX. LBX uses the ASIC capability of -EX leafs and newer in order to update the hardware forwarding tables (ECMP tables used to forward to the spines or setting the bounce bit when the vPC member ports are down on a leaf) faster and letting the control plane updates happen asynchronously.

This feature can be configured on a per-leaf basis; this is called Fast Link Failover. Enabling this feature on the leaf uplinks prevents the use of these links for the purpose of ERSPAN.

Debounce timer

The debounce timer is a default 100msec timer that is in place between the moment when the link-down event notification is generated and when the hardware tables are updated. This timer can be reduced to 10msec in order to improve convergence.

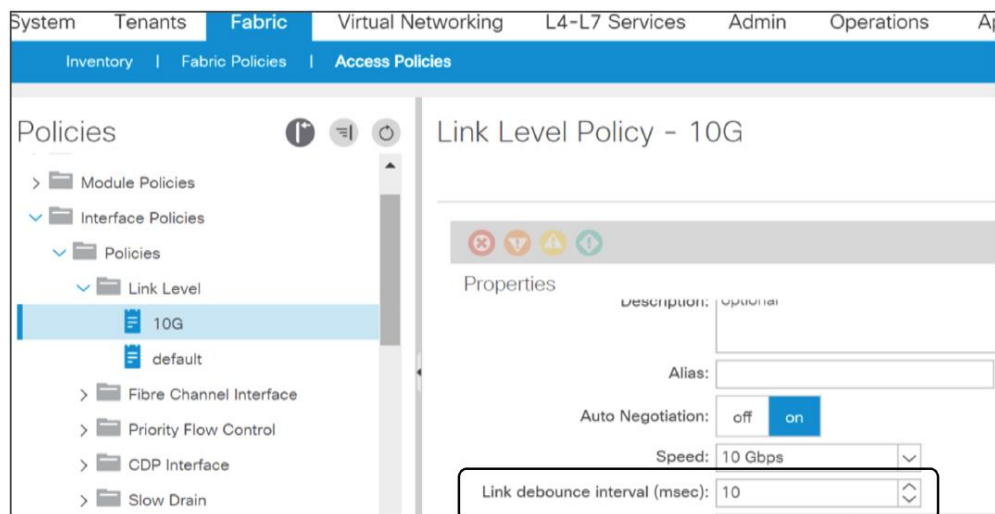


Figure 14.
Debounce timer

Bidirectional Forwarding Detection (BFD) for fabric links

Bidirectional Forwarding Detection (BFD) helps with subsecond convergence times on Layer 3 links. This is useful when peering routers are connected through a Layer 2 device or a Layer 2 cloud where the routers are not directly connected to each other.

From APIC Release 3.1(1), BFD can be configured between leaf and spine switches, and between spines and IPN links for GOLF, Multi-Pod, and Multi-Site connectivity (to be used in conjunction with OSPF or with static routes). BFD for fabric links is implemented for -EX line cards and leafs or newer.

Note: BFD is implemented on spines with -EX or -FX line cards or newer and Cisco Nexus 9364C fixed spines or newer.

Using BFD on leaf-to-spine links can be useful in case of stretched fabrics. This feature would then be used in conjunction with IS-IS. You can configure BFD on IS-IS via Fabric Policies, Interface Policies, Policies, and L3 Interface. If this is a global configuration for all leaf-to-spine links, you can simply modify the default policy; if, instead, this would be a specific configuration for some links, you would define a new L3 Interface policy and apply it to Leaf Fabric Ports Policy Groups.

Figure 15 shows how to enable BFD on fabric links.

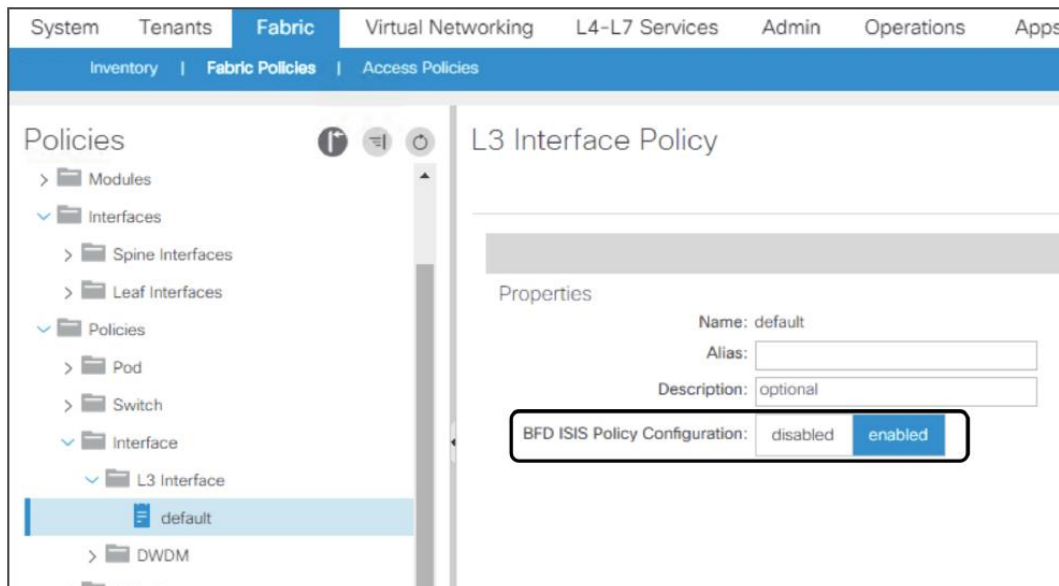


Figure 15.
Enabling Bidirectional Forward Detection on fabric links

Note: You can find more information about BFD at this link:

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/2-x/L3_config/b_Cisco_APIC_Layer_3_Configuration_Guide/b_Cisco_APIC_Layer_3_Configuration_Guide_chapter_0100.html

Quality of Service (QoS)

If the design consists of a single pod without GOLF, the Differentiated Services Code Point (DSCP) in the outer VXLAN header is not something you normally have to care about because that information is handled within the fabric itself in the overlay-1 infrastructure tenant.

In case the fabric is extended via Multi-Pod or GOLF, the VXLAN traffic is traversing a routed infrastructure and proper Quality of Service (QoS) must be in place to ensure the correct functioning of the Multi-Pod architecture. With GOLF or Multi-Pod, if you do not have either dot1p preserve configured or CoS translation configured in the tenant “infra”, the traffic forwarding between pods or between the WAN and the fabric may not work correctly.

Because of the above, it is still recommended, in single pod deployments, also to have either dot1p preserve or tenant “infra” Class of Service (CoS) translation configured, but not both.

Quality of Service for overlay traffic

Cisco ACI Release 4.0 uses six different user-configurable qos-groups to prioritize the traffic and four internally reserved qos-groups.

You can tune the user configurable qos-groups configurations from the Fabric Access Policies > Policies > Global Policies > QOS Class (please see Figure 16). By default, the traffic from a tenant EPG is mapped to the Level 3 class regardless of the CoS of the original packet.

If the dot1p preservation knob is set, the VXLAN DSCP header that is used within the fabric carries both the information about the original Class of Service from the incoming packet and the QoS Class level of the EPG. This ensures that, when the traffic leaves the fabric from an EPG, the CoS of the packet is set to the same value as the original frame (unless you configured a Custom QoS policy to overwrite it). The DSCP value of the original packet (that is, the inner DSCP value) is normally not modified, and is not mapped to the outer VXLAN header either. You can remark the DSCP of the original packet by configuring “Custom QoS.”

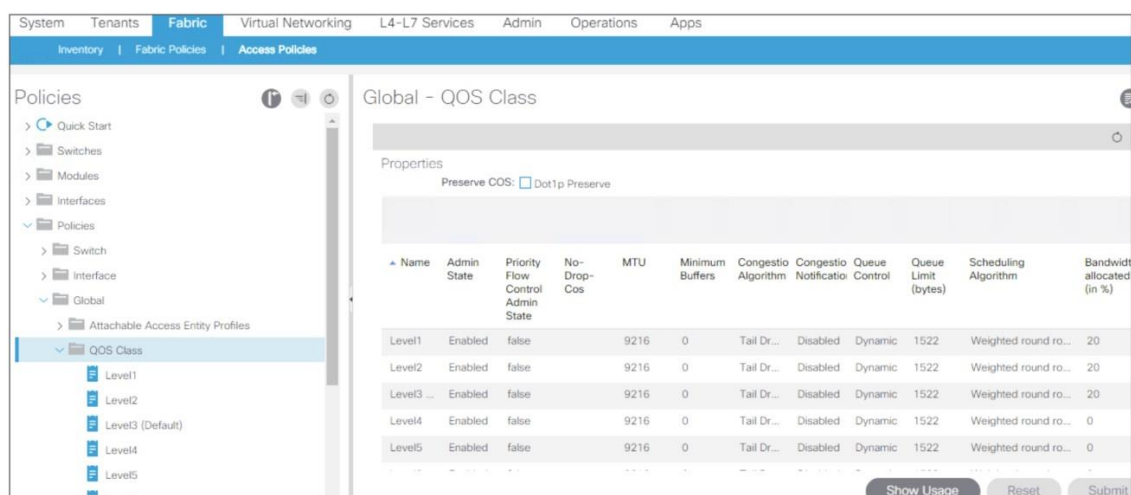


Figure 16.
Cisco ACI fabric QoS groups

Quality of Service for traffic going to an IPN

If you are planning to use Multi-Pod or GOLF, you may have to tune the tenant “infra” Quality of Service configuration. This is because, when doing Multi-Pod or GOLF, the fabric traffic is carried across an IPN network encapsulated in VXLAN, and the traffic must be correctly prioritized.

Often, the IPN network switches set the CoS of the traffic based on the DSCP values of the outer VXLAN header, and the receiving spine uses either the CoS or the DSCP value to associate the traffic with the correct queue in Cisco ACI. With default configurations, the spines receiving traffic from the IPN network assign either DSCP CS6 or CoS 6 to a special QoS class used by Cisco ACI for traceroute; therefore, if regular traffic received on a spine from the IPN is tagged with DSCP CS6 or CoS 6, it may be dropped.

If you configure Fabric Access “dot1p preserve,” the VXLAN DSCP outer header on the IPN carries DSCP values derived from the CoS of the original traffic and information about the qos-group or QoS Class Level that the traffic belonged to within the fabric, both encoded in the outer VXLAN DSCP header.

The main potential disadvantage of “dot1p preserve” is that if you need to configure QoS on the IPN by matching the DSCP values of the VXLAN header, you need to know how CoS and internal Cisco ACI QoS classes are mapped to the DSCP header, and you cannot change which DSCP value is used for what. This can be tricky if you need the flexibility to assign ACI traffic to a DSCP Class Selector that is not already in use.

As an example, if the IPN is used to connect to GOLF for north-to-south traffic and also for pod-to-pod connectivity, there may be north-to-south traffic with an outer VXLAN header of DSCP CS6 (the inner DSCP header may be copied by GOLF devices to the outer VXLAN header).

You may need then to choose DSCP class selectors for pod-to-pod control-plane traffic that does not overlap with the DSCP values used for north-to-south traffic.

If, instead of using dot1p preserve, you configure Cisco ACI tenant “infra” translations, you can map the ACI qos-group traffic to specific DSCP values for the outer VXLAN header. By doing this, you can choose DSCP values that are not already used by other traffic types.

Figure 17 shows how to set the dot1p preserve feature.

Figure 18 shows how to configure qos-group to DSCP translation for tenant “infra”. This is normally done by configuring the tenant “infra” > Policies, Protocol Policies > DSCP class-cos translation policy.

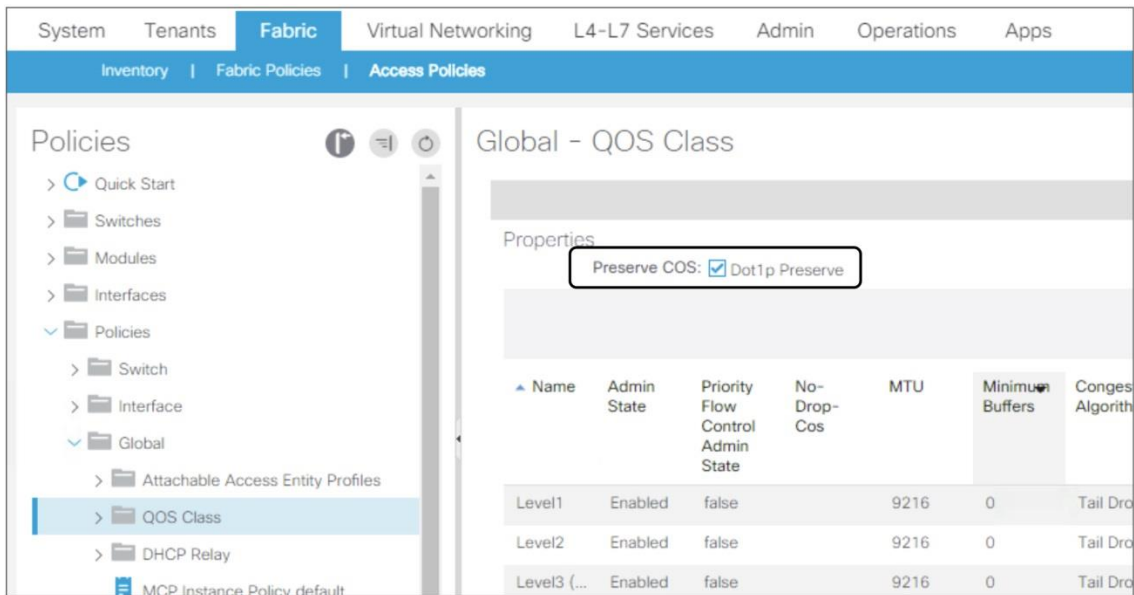


Figure 17.
Enabling dot1p preserve

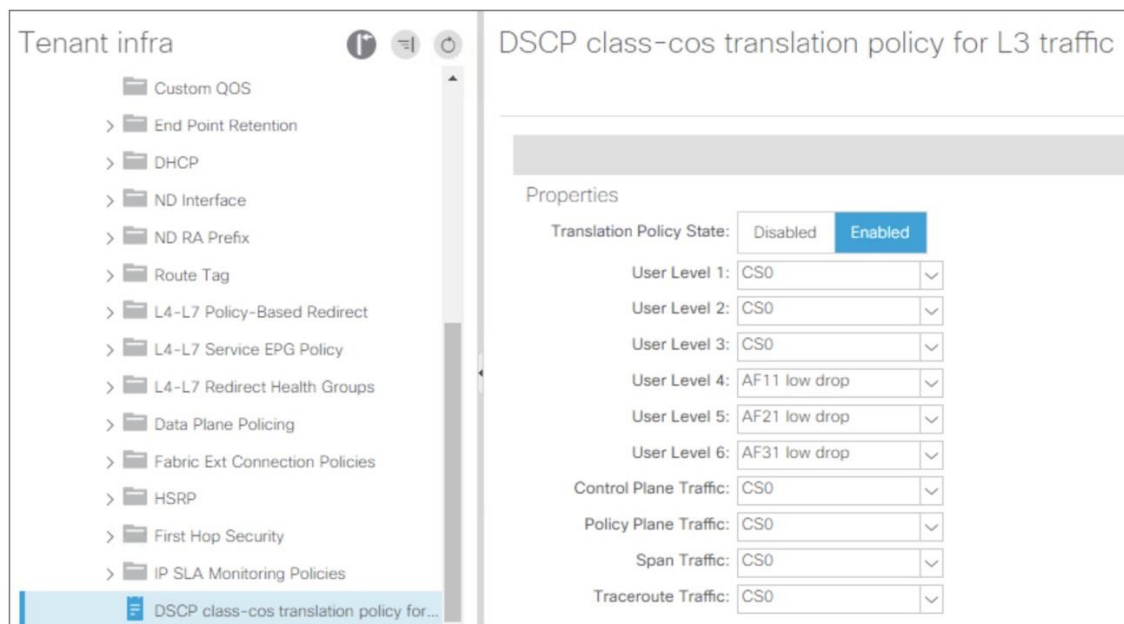


Figure 18.
QoS translation policy in tenant “infra”

You can find more details at this link:

http://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/kb/b_Multipod_QoS.html

The following design guidelines apply:

- You should always configure either dot1p preserve or tenant “infra” translation policies, but not both at the same time.
- CoS 6 and DSCP CS6 values are normally reserved for traceroute traffic, so normally you should ensure that Cisco ACI spines do not receive from the IPN any traffic with CoS 6 or DSCP CS 6.
- If the L3Out connectivity is provided via GOLF, make sure that the routed network between the GOLF device and the Cisco ACI spines is sending VXLAN traffic with outer DSCP header values that are consistent with the way ACI uses them. This means making sure that no VXLAN traffic from GOLF is using the DSCP values that ACI uses for control-plane, policy-plane, span, or traceroute purposes. In some cases this means re-marking VXLAN encapsulated traffic to the DSCP value that Cisco ACI associates with Level 3. There is no need instead to modify the DSCP values of the traffic encapsulated in VXLAN (the inner DSCP header).
- Cisco ACI Release 4.0 introduces more user-configurable qos-groups and the new encoding of these qos-groups into the outer DSCP header. Because of this, when upgrading from Cisco ACI 3.x to Cisco ACI 4.0 in presence of a transient mix-OS fabric, traffic between pods may not always be consistently classified. Having said that, the control-plane traffic classification is kept consistent.

Cisco APIC 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.

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

Cisco APIC teaming

APICs are equipped with two Network Interface Cards (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`.

Figure 19 shows a typical example of the connection of the APIC to the Cisco ACI fabric.

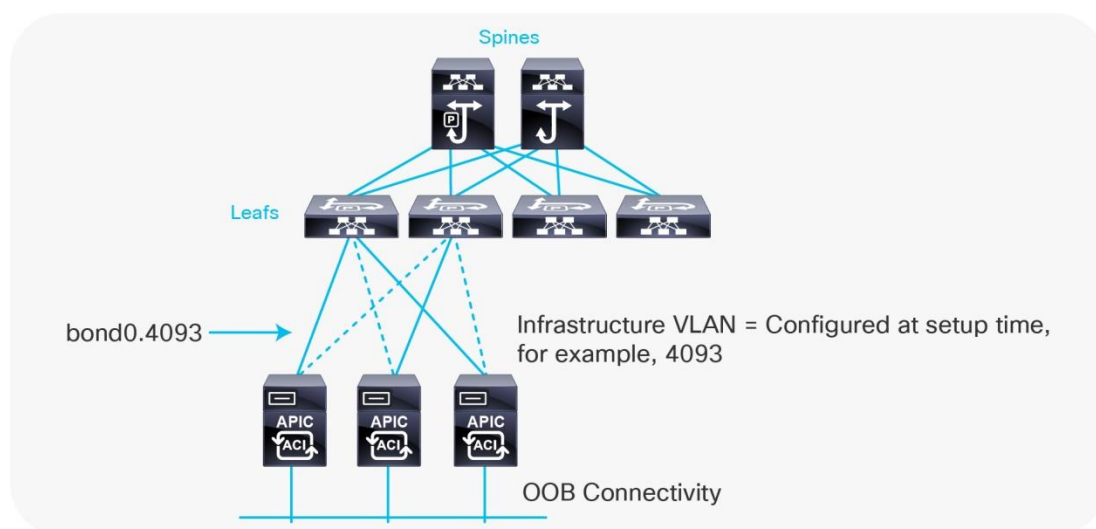


Figure 19.
Cisco APIC connection to the Cisco ACI fabric

APIC software creates bond0 and bond0 infrastructure 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.<infra VLAN>. This subinterface connects to the leaf switch. The infra VLAN ID 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.

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.

Internal IP address used for apps

Cisco ACI Release 2.2 (and newer) has the ability to host applications that run on APIC itself. This is done with a container architecture whose containers are addressed with IP addresses in the 172.17.0.0/16 subnet. At the time of this writing, this subnet range is not configurable, hence when configuring APIC management connectivity, one should make sure that this IP range does not overlap with management IP addresses or with management stations.

APIC clustering

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 mapping from an APIC ID to an APIC IP address and a universally unique identifier (UUID) for 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 of the APICs know the IP addresses of all the other APICs.

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 20.

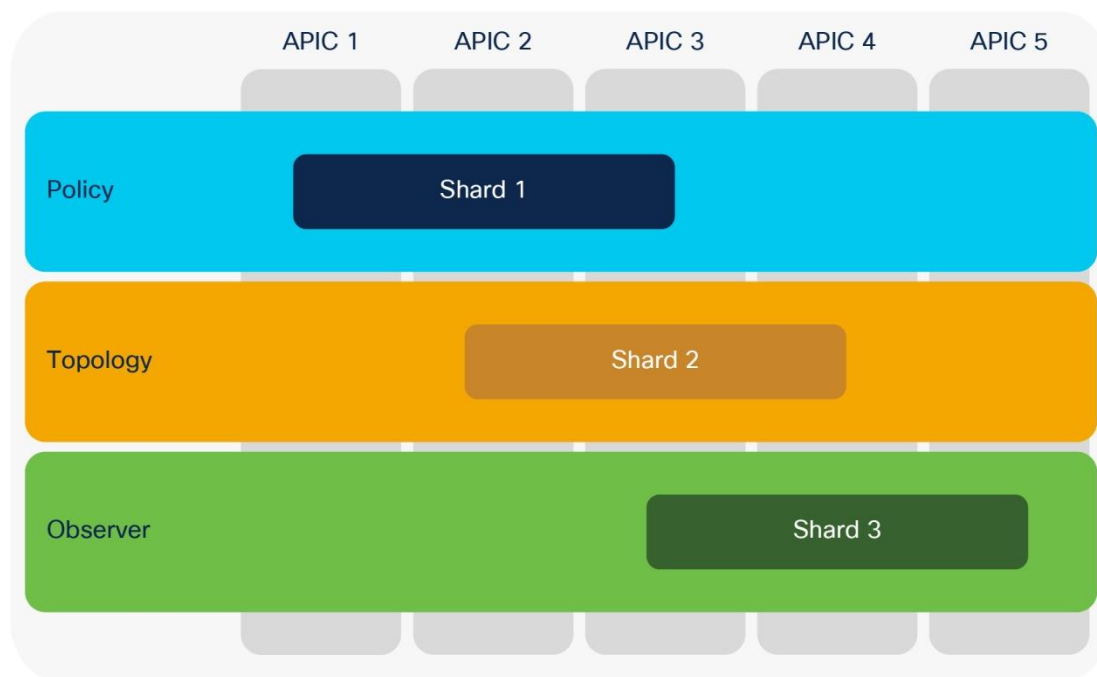


Figure 20.
Cisco APIC data sharding

Figure 20 shows that the policy data, topology data, and observer data are each replicated three times on a cluster of five 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(s) 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 a spare, ready to replace any active APIC in a cluster in one click. This controller does not participate in policy configurations or fabric management. No data is replicated to it, not even administrator credentials (use the `rescue -user login`).

In a cluster of three APICs + 1 standby, the controller that is in standby mode has, for instance, a 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 VXLAN network identifier (VNID) 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 data center 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.)

Designing the fabric “access”

The APIC management model divides the Cisco ACI fabric configuration into these two categories:

- Fabric infrastructure configurations: This is the configuration of the physical fabric in terms of vPCs, VLANs, loop prevention features, underlay BPG protocol, and so on.
- Tenant configurations: These configurations are the definition of the logical constructs such as application profiles, bridge domains, EPGs, and so on.

This section describes the network connectivity preparation steps normally performed as part of the fabric infrastructure configurations.

Naming of Cisco ACI objects

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 1. Sample naming conventions

Policy name	Examples
Tenants	
[Function]	Production Development

Policy name	Examples
VRFs	
[Function]	Trusted Untrusted Production Development
Bridge domains	
[Function]	Web App AppTier1 AppTier2
Endpoint groups	
[Function]	Web App AppTier1 AppTier2
Attachable access entity profiles	
[Function]	VMM BareMetalHosts L3Out_N7K
VLAN pools	
[Function]	VMM BareMetalHosts L3Out_N7K
Domains	
[Function]	BareMetalHosts VMM L2DCI L3DCI
Contracts	
[Prov]_to_[cons]	Web_to_App
Subjects	
[Rulegroup]	WebTraffic

Policy name	Examples
Filters	
[Resource-Name]	HTTP
Application profiles	
[Function]	SAP Exchange
Interface policies	
[Type] [Enable Disable]	CDP_Enable CDP_Disable LLDP_Disable
Interface policy groups	
[Type]_[Functionality]	vPC_ESXi-Host1 PC_ESXi-Host1 PORT_ESXi-Host1
Interface profiles	
[Node1]_[Node2]	101_102

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.

Note: You should not use a name with “N-” (N followed by hyphen) as a substring for the object that defines a L4-L7 device. You should not use a name that includes the substring “C-” for a bridge domain that is used as part of the service graph deployment (meaning a bridge domain that needs to be selected in the device selection policy configuration of a service graph).

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, which states:

“In the case of policy resolution based on named relations, if a target MO [Managed Object] 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 object of the same name in the tenant is selected instead of the object in Tenant common. For instance if you defined a BD, BD-1 in tenant Tenant-1 and if you defined VRF VRF-1 in Tenant common and also in Tenant-1, you can associate BD-1 to Tenant-1/VRF-1 but you cannot associate BD-1 to Common/VRF-1.

Defining 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

VLAN range	Type	Use
1000 – 1100	Static	Bare-metal hosts
1101 – 1200	Static	Firewalls
1201 – 1300	Static	External WAN routers
1301 – 1400	Dynamic	Virtual machines

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.

Note: In order to prevent misconfiguration, it is recommended to enable the domain validation feature globally at System Settings > Fabric Wide Settings

When choosing VLAN pools, keep in mind that, if the servers connect to ACI via an intermediate switch or a Cisco UCS Fabric Interconnect, you need to choose a pool of VLANs that does not overlap with the reserved VLAN ranges of the intermediate devices, which means using VLANs < 3915.

Cisco Nexus 9000, 7000, and 5000 Series Switches reserve the range 3968 to 4095.

Cisco UCS reserves the following VLANs:

- FI-6200/FI-6332/FI-6332-16UP/FI-6324: 4030-4047. Note vlan 4048 is being used by vsan 1
- FI-6454: 4030-4047 (fixed), 3915-4042 (can be moved to a different 128 contiguous block vlan but requires a reboot). See the following link for more information:

https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/ucs-manager/GUI-User-Guides/Network-Mgmt/3-1/b_UCSM_Network_Mgmt_Guide_3_1/b_UCSM_Network_Mgmt_Guide_3_1_chapter_0110.html

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, and interface-related configurations such as LACP, LLDP, Cisco Discovery Protocol, port channels and vPCs.

The access policy configuration generally follows the workflow shown in Figure 21.

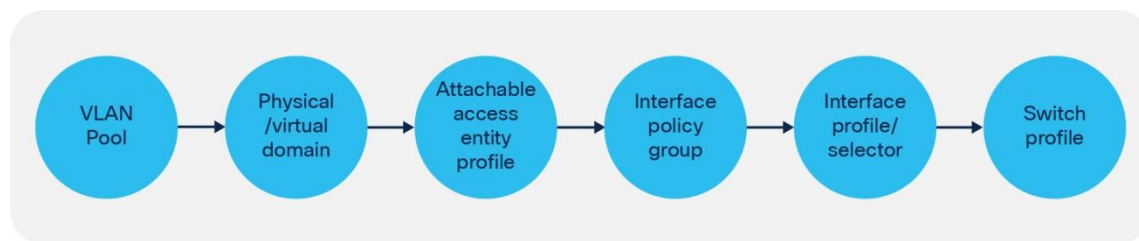


Figure 21.

Access policy configuration workflow

Attachable Access Entity Profiles (AAEPs)

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 Cisco 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 22.

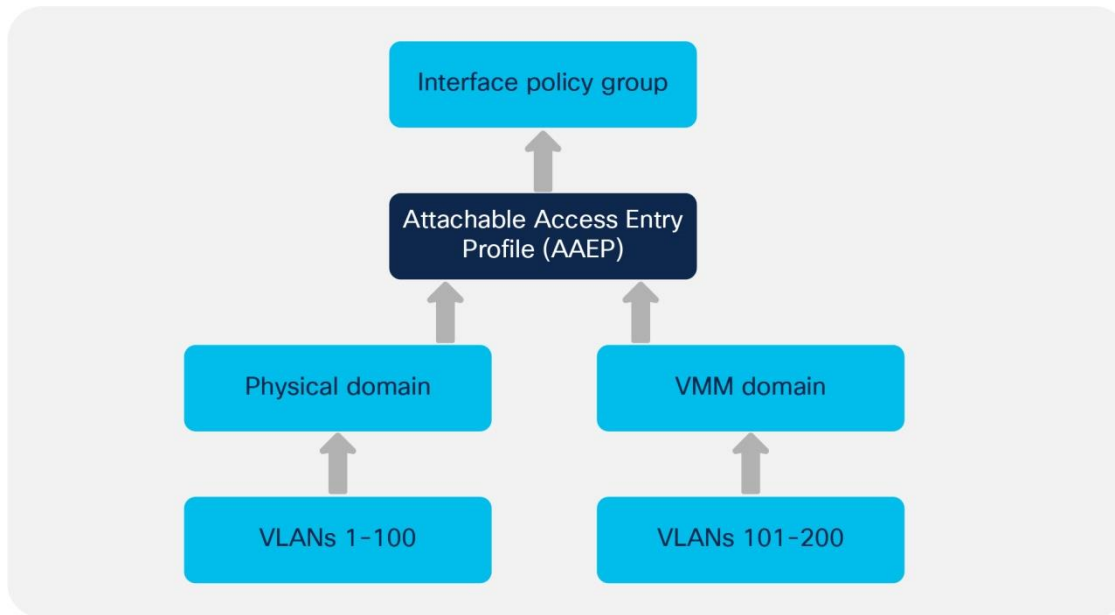


Figure 22.
AAEP relationships

In the example in Figure 22, 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.

Defining the EPG to domain association enables the use of VLANs from the domain range for the static or dynamic bindings defined in this EPG. This same VLAN range must have been enabled on the leaf ports via the AAEP. For instance, imagine that EPG1 from BD1 uses port 1/1, VLAN10, and that VLAN10 is part of physical domain domain1, that same physical domain must have been configured on port 1/1 as part of the fabric access AAEP configuration.

Traffic from a VLAN is mapped to a bridge domain based on tenant configurations, and it is carried in the fabric with a VXLAN VNID that depends on the bridge domain. Each VLAN also has another VNID that is called the “Fabric Encapsulation” or FD_VLAN VNID, which is used for special cases such as when forwarding BPDUs or when using “Flood in Encapsulation”.

In the case of overlapping VLAN ranges used by different domains and different AAEPs, the fabric encapsulation used in the fabric is different for VLAN10 used by EPG1 with domain1 versus VLAN 10 used by EPG2 with domain2. This is a desirable property because the assumption is that EPG1 and EPG2 belong to different Bridge Domains (BDs), and there is no point in forwarding BPDUs (as an example) from one BD to the other.

There are some scenarios where, instead, care should be taken to avoid mapping multiple domains with overlapping VLAN ranges to the same EPG because the fabric encapsulation used can be nondeterministic. As an example, imagine that there are two VLAN pools: pool1 with VLAN range 1 - 10 and pool2 with VLAN range 10 - 15. Imagine that they are associated respectively with physdom1 and physdom2, and that they are both mapped to the same EPG. If the EPG is then mapped to port 1/1 and 1/2 for the same VLAN 10, the fabric encapsulation may not be the same, and, if this BD is used for Layer 2 connectivity with an external switching infrastructure, it may lead to a Layer 2 loop, because the Spanning Tree Protocol may not be able to detect the loop.

In light of this, when defining an AAEP, make sure that the domains mapped to it do not use overlapping VLAN ranges.

Configuring the global setting “Enforce Domain validation” helps ensure that the fabric-access domain configuration and the EPG configurations are correct in terms of VLANs, thus preventing configuration mistakes. This configuration is illustrated in Figure 23.

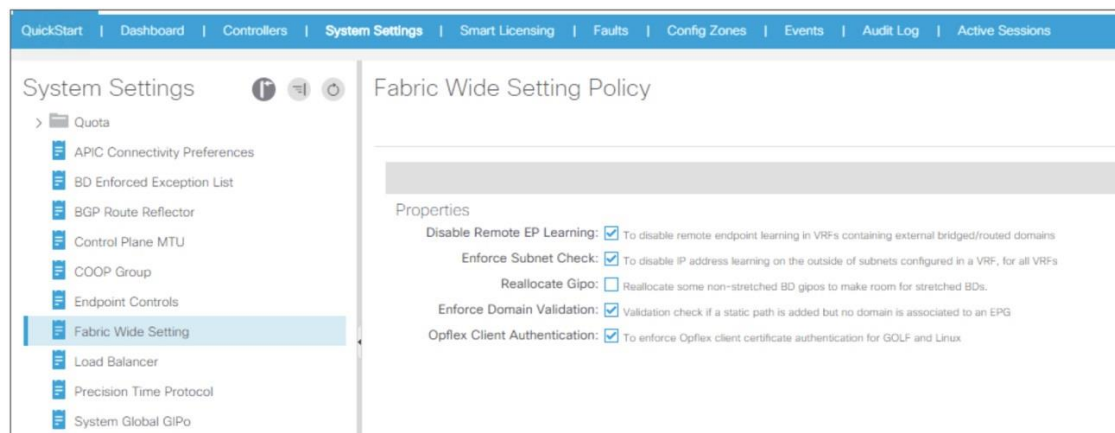


Figure 23.
Configuring domain validation

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 something 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, Switch Virtual Interfaces (SVIs) are not configured on the hardware of the leaf device unless there are endpoints on the leaf that require it. Cisco ACI determines whether these resources are required on a given leaf based on Cisco Discovery Protocol, LLDP, or OpFlex (when the servers support it).

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

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

In Cisco ACI, by default, LLDP is enabled with an interval of 30 seconds and a holdtime of 120 seconds. The configuration is global and can be found in Fabric > Fabric Policies > Global.

CDP uses the usual Cisco CDP timers with an interval of 60s and a holdtime of 120s.

If you do not specify any configuration in the policy-group, LLDP, by default, is running and CDP is not. The two are not mutually exclusive, so if you configure CDP to be enabled on the policy-group, Cisco ACI generates both CDP and LLDP packets.

If you are using fabric extenders (FEX) 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. 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 Cisco 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.

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 24.

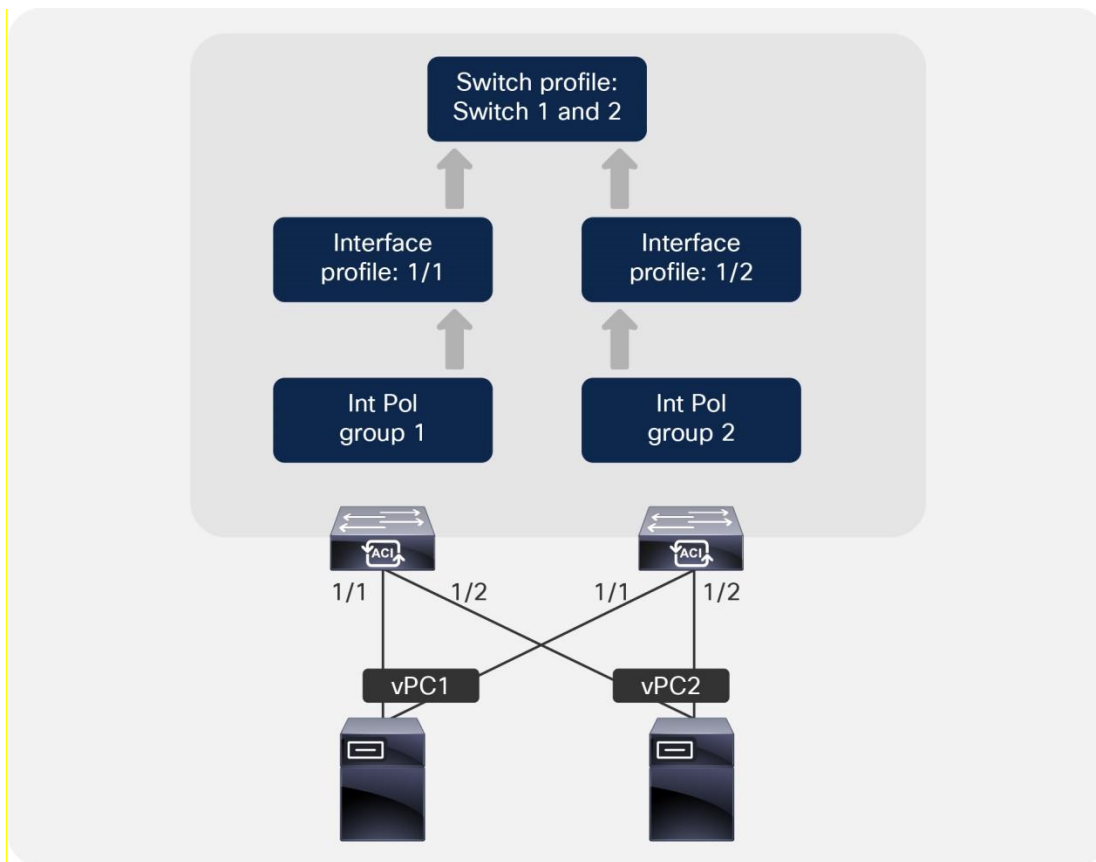


Figure 24.
VPC interface policy groups

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.

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.

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. Re-using leaf access port interface policy groups is fine as long as the person who manages the Cisco ACI infrastructure realizes that a configuration change in the policy group applies potentially to a large number of ports.

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.

Configuration for faster convergence with vPCs

Starting with Cisco ACI Release 3.1, the convergence times for several failure scenarios have been improved. One such failure scenario is the failure of a vPC from a server to the leafs. To further improve the convergence times, you should configure the Link Debounce interval timer under the Link Level Policies for 10ms, instead of the default of 100ms.

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 been configured with Cisco Discovery Protocol, as shown in Figure 25.

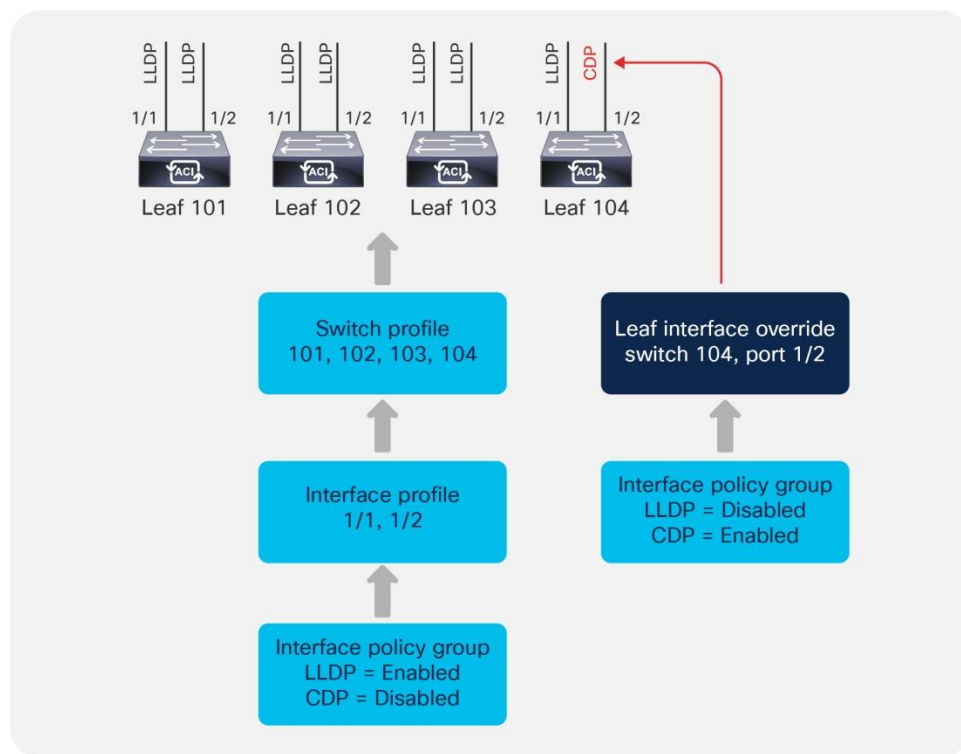


Figure 25.
Interface overrides

Interface overrides are configured in the Interface Policies section under Fabric Access Policies, as shown in Figure 26.

Create Leaf Interface Override

Specify the Policy and Interface

Name: Override-1

Description: optional

Path Type: ☒ Port ☐ Direct Port Channel ☐ Virtual Port Channel

Path: Pod-1/Node-112/eth1/5
Node ID[Fax ID]Card ID/Port ID For example: Node-17/eth1/8, or Node-17/Fax-101/eth1/8

Policy Group: APIC_PolGrp

Figure 26.
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.

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 27).

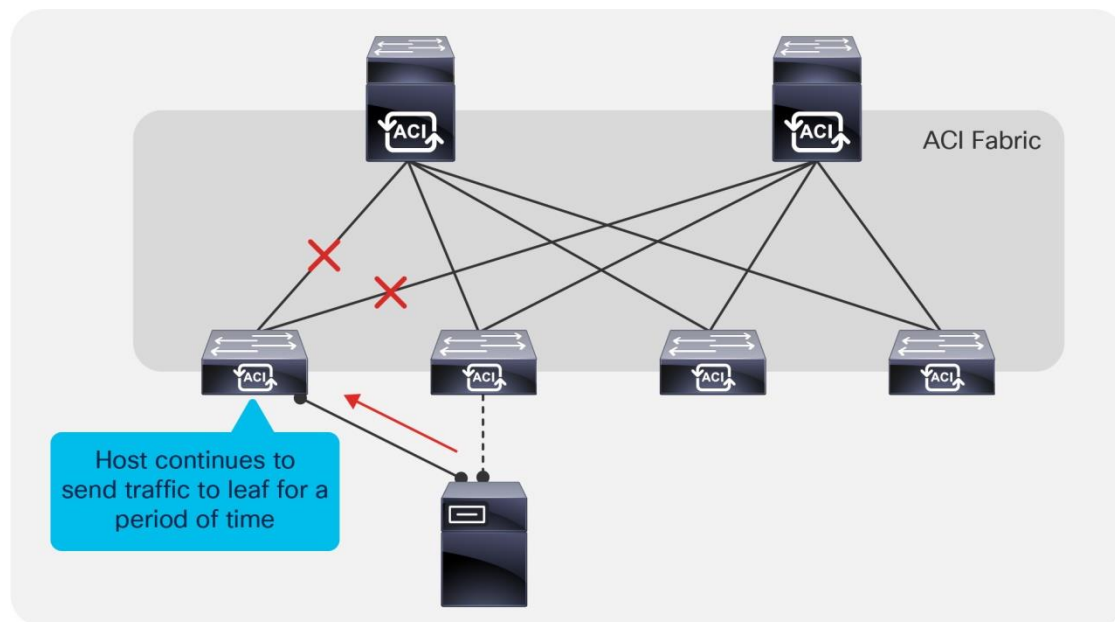


Figure 27.
Loss of leaf connectivity in an active/standby NIC Teaming scenario

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 28.

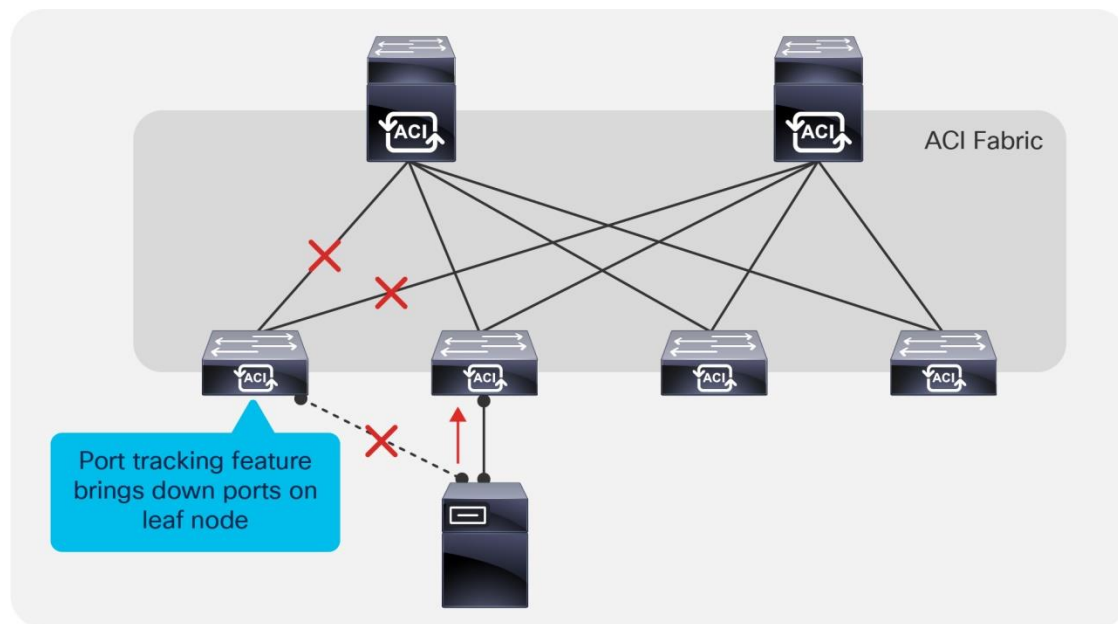


Figure 28.
Active/standby NIC Teaming with port tracking enabled

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

Port tracking is located under System > System Settings > Port Tracking.

Loop mitigation features

Cisco ACI is a routed fabric, hence there is intrinsically no possibility of a loop at the fabric infrastructure level. On the other hand, you can build Bridge Domains (BDs) on top of the routed fabric, and you could potentially introduce loops by merging these domains with external cabling or switching. It is also possible to have a loop on the outside networks connected to the ACI fabric, and these loops could also have an impact on the ACI fabric.

This section illustrates the features that can be configured at the fabric-access policy level in order to reduce the chance for loops and/or reduce the impact of loops on the Cisco ACI fabric.

The following features help prevent loops: the MisCabling Protocol, traffic storm control, and BPDU forwarding on the FD_VLAN, BPDU Guard (this last one only where applicable, because BPDUs may be the right tool to keep the topology loop free).

Other features help minimize the impact of loops on the fabric itself: Control Plane Policing per interface per protocol (CoPP) endpoint move dampening, endpoint loop detection, and rogue endpoint control.

Interface-level Control Plane Policing (CoPP)

Control Plane Policing (CoPP) has been introduced in Cisco ACI 3.1. With this feature, control traffic is rate-limited first by the interface-level policer before it hits the aggregated CoPP policer. This prevents traffic from one interface from “flooding” the aggregate CoPP policer, and as a result ensures that control traffic from other interfaces can reach the CPU in case of loops or Distributed Denial of Service (DDoS) attacks from one or more interfaces. The per-interface-per-protocol policer supports the following protocols: Address Resolution Protocol (ARP), Internet Control Message Protocol (ICMP), Cisco Discovery Protocol (CDP), Link Layer Discovery Protocol (LLDP), Link Aggregation Control Protocol (LACP), Border Gateway Protocol (BGP), Spanning Tree Protocol, Bidirectional Forwarding Detection (BFD), and Open Shortest Path First (OSPF). It requires Cisco Nexus 9300-EX or newer switches.

MisCabling Protocol (MCP)

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 loops. If Spanning Tree Protocol is running on the external switching infrastructure, under normal conditions MCP does not need to disable any link; should Spanning Tree Protocol stop working on the external switches, MCP intervenes to prevent a loop.

Even if MCP detects loops per VLAN, 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.

Spanning Tree Protocol provides better granularity, so, if a looped topology is present, external switches running Spanning Tree Protocol provide more granular loop-prevention. MCP is useful if Spanning Tree Protocol stops working.

It is recommended that you enable MCP on all ports facing external switches or similar devices.

The MCP policy-group level default configuration sets MCP as enabled on the interface, but MCP does not work until and unless MCP is configured as globally enabled.

While the MCP default (enabled) is set on all the interfaces, you need to enable a global MCP configuration for MCP to work.

This can be done via the Global Policies section of the Fabric > Access Policies tab, as shown in Figure 29.

The screenshot shows the Cisco ACI Fabric Policies configuration page. The left sidebar contains a tree view with the following structure:

- System
- Tenants
- Fabric**
 - Inventory
 - Fabric Policies
 - Access Policies**
 - Policies
 - Quick Start
 - Switches
 - Modules
 - Interfaces
 - Policies**
 - Switch
 - Interface
 - Global**
 - Attachable Access Entity Profiles
 - QOS Class
 - DHCP Relay
 - MCP Instance Policy default** (selected)
 - Error Disabled Recovery Policy
 - Monitoring

The main configuration area for 'MCP' is titled 'Properties' and contains the following fields:

- Name: default
- Description: optional
- Admin State: **Disabled** (selected), Enabled
- Controls: ☒ Enable MCP PDU per VLAN
- Key:
- Confirm Key:
- Loop Detect Multiplication Factor: 3
- Loop Protection Action: ☒ Port Disable
- Initial Delay (sec): 180
- Transmission Frequency (sec): 2 (msec): 0

Figure 29.
MCP configuration

The configuration of MCP requires entering a key to uniquely identify the fabric. The initial delay (by default, 180s) is designed to ensure that if the Cisco ACI leaf is connected to an external L2 network, MCP gives enough time to STP to converge.

The minimum time for MCP to detect a loop is $(tx\ freq * loop\ detect\ multiplier + 1)$.

The default transmission frequency is 2 seconds. The loop detect multiplication factor is the number of continuous packets that a Cisco ACI leaf port must receive before declaring a loop; the default is 3. With the default timer, it takes ~7s for MCP to detect a loop.

The loop protection action can be just to generate a syslog; disabling the port upon loop detection can also be enabled here.

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.

Even if MCP can detect loops per-VLAN, 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.

Per-VLAN MCP supports a maximum of 256 VLANs per link, which means that if there are more than 256 VLANs on a given link MCP generates PDUs on the first 256.

Traffic storm control

Traffic 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. Traffic storm control on the Cisco ACI fabric is configured by opening the Fabric > Access Policies menu and choosing Interface Policies.

Traffic 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 traffic storm control begins to drop traffic. This rate can be defined as a percentage or the number of packets per second.

Traffic 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 traffic storm control policy will apply to broadcast and multicast traffic. If, however, the bridge domain is set to flood unknown unicast traffic, traffic storm control will apply to broadcast, multicast, and unknown unicast traffic.

Choosing among endpoint move dampening, endpoint loop protection and rogue endpoint control

Cisco ACI has three features that look similar in that they help when an endpoint is moving too often between ports:

- Endpoint move dampening is configured from the bridge domain under the Endpoint Retention Policy and is configured as “Move Frequency.” The frequency expresses the number of aggregate moves of endpoints in the bridge domain. When the frequency is exceeded, Cisco ACI stops learning on this bridge domain. Notice that a single endpoint moving may not cause this feature to intervene because Cisco ACI has a deduplication feature in hardware. The feature is effective when multiple endpoints move, as it is normally the case with a loop.
- The endpoint loop protection is a feature configured at the global level. The feature is turned on for all bridge domains, and when too many moves are detected you can choose whether Cisco ACI should suspend one of the links that cause the loop (you cannot control which one), or disable learning on the bridge domain.
- Rogue endpoint control is similar to the endpoint loop protection feature in that it is a global setting but when a “loop” is detected, Cisco ACI just quarantines the endpoint; that is, it freezes the endpoint as belonging to a VLAN on a port and disables learning on it.

Endpoint move dampening counts the aggregate moves of endpoints, hence if you have a single link failover with a number of endpoints whose count exceed the configured “move frequency” (the default is 256 “moves”), endpoint move dampening may also disable learning. When the failover is the result of the active link (or path) going down, this is not a problem because the link going down flushes the endpoint table of the previously active path. If instead the new link takes over without the previously active one going down, endpoint dampening will disable the learning after the configurable threshold (256 endpoints) is exceeded. If you use endpoint move dampening you should tune the move frequency to match the highest number of active endpoints associated with a single path (link, port-channel or vPC). This scenario doesn’t require special tuning for endpoint loop protection and rogue endpoint control because these two features count moves in a different way.

Figure 30 illustrates how endpoint loop protection and rogue endpoint control help with either misconfigured servers or with loops. In the figure, an external Layer 2 network is connected to a Cisco ACI fabric, and due to some misconfiguration traffic from H1 (such as an ARP packet, for instance) is looped. If there is only one endpoint moving too often ACI, endpoint move dampening would not disable learning, but endpoint loop protection and rogue endpoint control would. In fact, Cisco ACI leafs have a deduplication feature that lets the ACI count the move of individual endpoints (see the right-hand side of the figure) and detect a loop regardless of whether a single endpoint is moving too often (which most likely is not a loop but maybe an incorrect NIC-teaming configuration) or multiple endpoints are moving too often (as happens with loops).

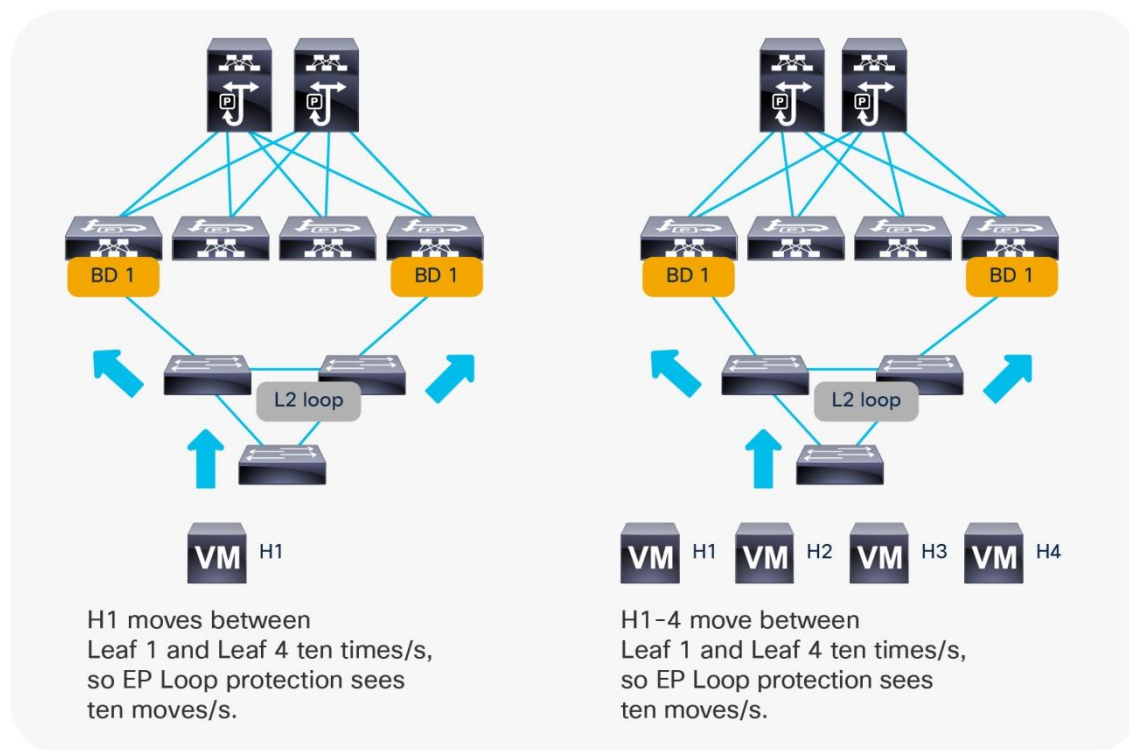


Figure 30.
Cisco ACI leafs count endpoint moves

Endpoint loop protection 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:

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

The default parameters are as follows:

- 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.

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.

Rogue endpoint control is a feature, introduced in Cisco ACI 3.2, that can help in case there are MAC or IP addresses that are moving too often between ports. With rogue endpoint control, only the misbehaving endpoint (MAC/IP) is quarantined, which means that Cisco ACI keeps its TEP and port fixed for a certain amount of time when learning is disabled for this endpoint. The feature also raises a fault to allow easy identification of the problematic endpoint. If rogue endpoint control is enabled, loop detection and bridge domain move frequency will not take effect. The feature works within a site.

Rogue endpoint control does not stop a L2 loop, but it provides mitigation of the impact of a loop on the COOP control plane by quarantining the endpoints.

Rogue endpoint control also helps in case of incorrect configurations on servers, which may cause endpoint flapping. In such a case, Cisco ACI, instead of disabling the server ports, as endpoint loop detection may do, it stops the learning for the endpoint that is moving too often, and it provides a fault with the IP of the endpoint that is moving too often so that the administrator can verify its configuration.

At the time of this writing, it is recommended to enable rogue endpoint control.

Note: If you downgrade from Cisco ACI 3.2 to previous releases, you will need to disable this feature.

Figure 31 illustrates how to enable rogue endpoint control.

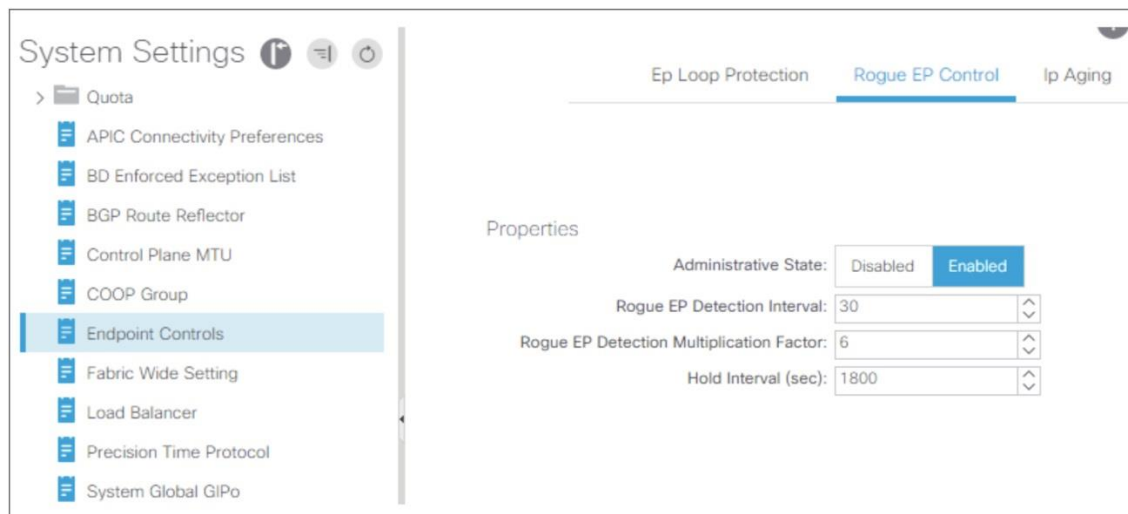


Figure 31.
Rogue endpoint control is enabled in System Settings

Error Disabled Recovery Policy

Together with defining the loop protection configuration, you should also define after how much time ports that were put in an error-disabled state can be brought back up again.

This is done with the Error Disabled Recovery Policy. Figure 32 illustrates how to configure it.

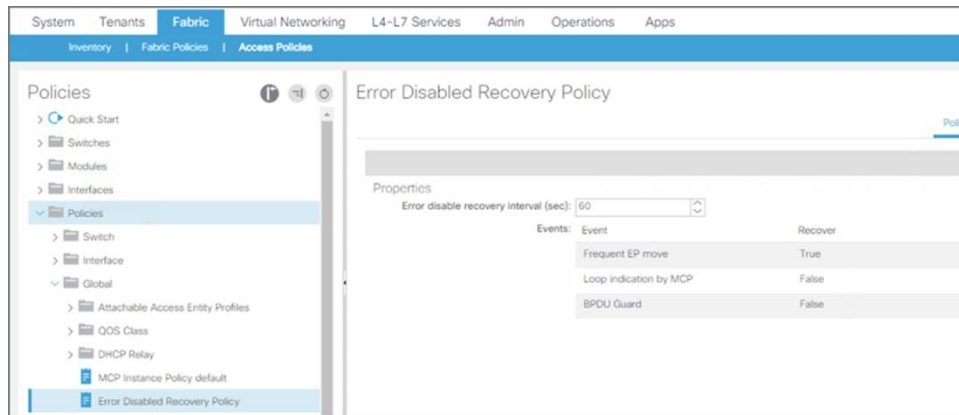


Figure 32.
Error Disabled Recovery Policy

Spanning Tree Protocol considerations

The Cisco ACI fabric does not run Spanning Tree Protocol natively, but it can forward BPDUs within the EPGs.

The flooding scope for BPDUs is different from the flooding scope for data traffic. The unknown unicast traffic and broadcast traffic are flooded within the bridge domain; Spanning-Tree-Protocol BPDUs are flooded within a specific VLAN encapsulation (also known as FD_VLAN), and in many cases, though not necessarily, an EPG corresponds to a VLAN.

Figure 33 shows an example in which external switches connect to the fabric.

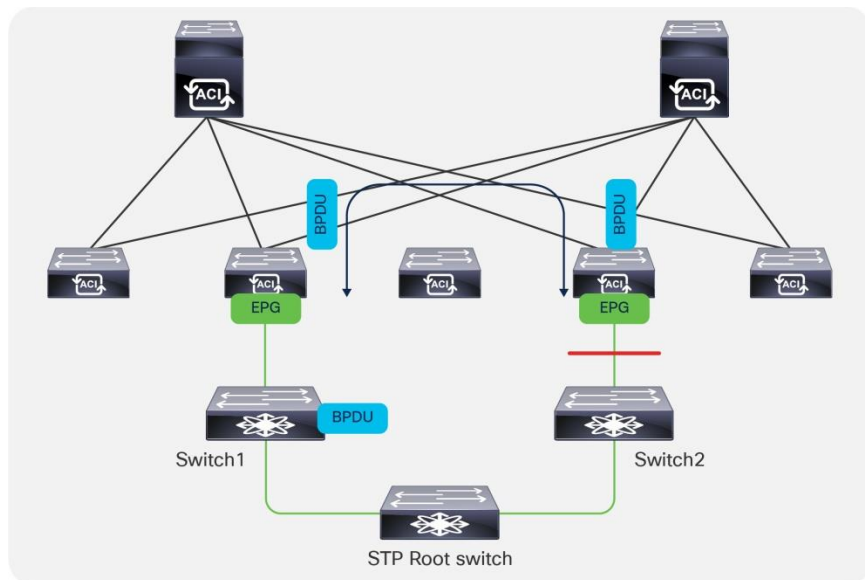


Figure 33.
Fabric BPDU flooding behavior

BPDUs received from a leaf are classified by Cisco ACI as belonging to the control plane qos-group, and this classification is preserved across pods. If forwarding BPDUs across pods, make sure that either dot1p preserve or tenant “infra” CoS translation is configured.

Minimize the scope of Spanning-Tree topology changes

As part of the Spanning-Tree design, you should also make sure that Spanning-Tree topology change notifications (TCNs) due to changes in the forwarding topology of an external Layer 2 network do not unnecessarily flush the bridge domain endpoints in the Cisco ACI fabric.

When Cisco ACI receives a TCN BPDU on a VLAN in a bridge domain, it flushes all the endpoints associated with this VLAN in that bridge domain.

Because of this, within an EPG, the VLANs used for connectivity to a switched network should be different from the VLANs used for endpoints directly attached to the leafs of the fabric. This approach limits the impact of Spanning-Tree TCN events to clearing the endpoints learned on the switched network.

Spanning-Tree BPDU Guard

It is good practice to configure ports that connect to physical servers with BPDU Guard so that if an external switch is connected instead, the port is error-disabled.

It can also be useful to configure BPDU Guard on virtual ports (in the VMM domain).

Configure virtual switches so they forward Layer 2 PDUs

The Cisco ACI fabric can be configured to verify and prevent loops, but if a virtual switch in a virtualized server does not forward Layer 2 PDUs, loops may still be inadvertently introduced. Cisco ACI features such as traffic storm control, rogue endpoint detection, and so on mitigate the effects of loops, but loops can be avoided altogether if the virtual switch is properly configured.

In order to make sure that virtual switches do not drop BPDUs, please refer to this article:

<https://kb.vmware.com/s/article/2047822>, which explains how to change the virtual switch configuration Net.BlockGuestBPDU.

If you configure the above and enable per-VLAN MCP in Cisco ACI, the possibility of a virtualized server introducing loops is greatly reduced.

Best practices for Layer 2 loop mitigation

In summary, in order to reduce the chance of loops and their impact on the fabric, you should do the following:

- Enable global rogue endpoint control to mitigate the impact of loops (and of incorrect NIC Teaming configurations) on the Cisco ACI fabric.
- Enable MCP at Fabric Access Global Policies by entering a key to identify the fabric and by changing the administrative state to enabled. Enable also per-VLAN MCP.

- For Cisco ACI leaf interfaces connected directly to servers:
 - Make sure MCP is enabled on the interface (the default MCP policy normally is with MCP enabled; hence, if you enabled MCP globally, MCP will be enabled on the interface).
 - Make sure Spanning-Tree BPDU Guard is configured.
 - Most virtualized hosts today support both LLDP and CDP, and LLDP is on by default, so just make sure the Cisco ACI leaf interfaces are configured with the protocol that matches the host's capabilities.
- For Cisco ACI interfaces connected to external Layer 2 switches without loops (typically via a single vPC):
 - Configure the external switches to filter BPDUs on the interfaces connected to Cisco ACI so as to limit the impact of TCNs on the ACI fabric.
 - Configure the Cisco ACI interfaces with MCP, BPDU Guard, and, potentially, with traffic storm control. While filtering BPDUs from the external L2 switch prevents Spanning Tree Protocol from protecting against a loop, if MCP and BPDU Guard are configured on this vPC and on all other Cisco ACI interfaces, should another switch be added to the fabric, its ports will be error-disabled.
 - Most networking devices today support both LLDP and CDP, so just make sure the ACI leaf interfaces are configured with the protocol that matches the capabilities of connected network devices.
- For Cisco ACI interfaces connected to external Layer 2 switches via multiple Layer 2 paths:
 - Configure Cisco ACI so that the BPDUs of the external network are forwarded by ACI by configuring EPGs with consistent VLAN mappings to ports connected to the same L2 network.
 - Make sure that the external Layer 2 switches run Spanning Tree Protocol.
 - The Cisco ACI leaf ports connected to the Layer 2 network should have neither BPDU Guard nor BPDU Filter enabled.
 - Spanning Tree Protocol running on the external network puts the redundant links in blocking mode.
 - You can keep MCP running on the same Cisco ACI interfaces: when Spanning Tree Protocol is working, MCP packets will not be looped back to ACI, hence the topology is kept free from loops by Spanning Tree Protocol. Should Spanning Tree Protocol stop working, MCP frames will be forwarded, and as a result, MCP will error-disable the entire link.
 - You can also configure traffic storm control as an additional mitigation in case of loops. Most networking devices today support both LLDP and CDP, so just make sure the Cisco ACI leaf interfaces are configured with the protocol that matches the capabilities of connected network devices.

Global configurations

This section summarizes some of the “Global” settings that are considered best practices.

These settings apply to all tenants:

- Configure two BGP route reflectors from the available spines.
- Disable Remote Endpoint Learning: This is to prevent stale entries in the remote table of the border leafs as a result of conversations between endpoints on fabric leafs and endpoints on the border leaf.
- Enable “Enforce Subnet Check”: This configuration ensures that Cisco ACI learns endpoints whose IP address belongs to the bridge domain subnet; it also ensures that leafs learn the IP of remote endpoints only if the IP address belongs to the VRF that they are associated with.
- Enable IP Aging: This configuration is useful to age individual IP addresses that may be associated with the same MAC address (for instance, in the case of a device that does NAT and is connected to the Cisco ACI).
- Enable “Enforce Domain validation”: This option ensures that the fabric access domain configuration and the EPG configurations are correct in terms of VLANs, thus preventing configuration mistakes.
- At the time of this writing if you plan to deploy Kubernetes (K8s) or Red Hat OpenShift Container Platform, you should deselect Opflex Client Authentication.
- Enable per-VLAN MCP globally: This ensures that, when using MCP on a link, Cisco ACI sends MCP PDUs per VLAN instead of only one per link on the native VLAN.
- Enable either Endpoint Loop Protection or Rogue Endpoint Detection: These features limit the impact of a loop on the fabric by either disabling dataplane learning on a bridge domain where there is a loop or by quarantining the endpoints whose MAC and/or IP moves too often between ports.
- Configure a lower cost for IS-IS redistributed routes than the default value of 63.
- QoS: Even if you are deploying a fabric with a single pod, it is good to configure Quality of Service mappings between a single pod’s qos-groups and the DSCP values to be assigned to the outer header of VXLAN traffic from day-1. This will avoid issues in the future; for example, when extending the fabric via Multi-Pod. The recommended settings can be found at this link:

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/kb/b_Multipod_QoS.html.

Figure 34 shows how to configure the global system settings.

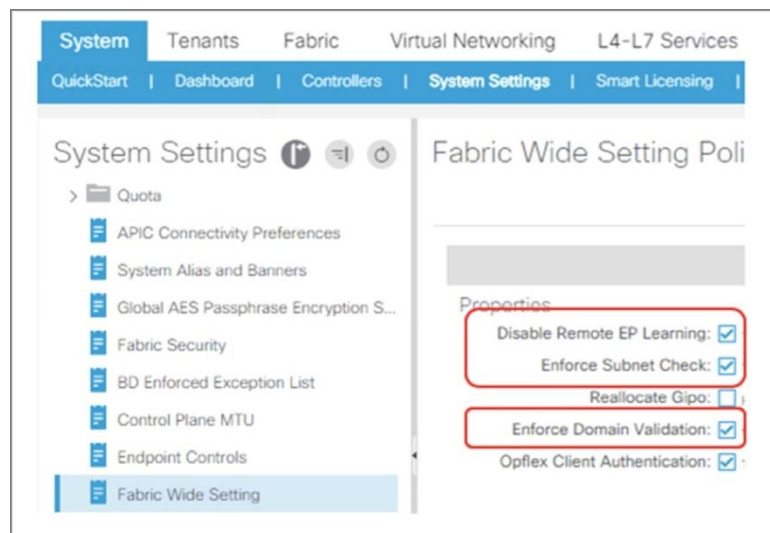


Figure 34.
System settings recommended configuration

Disable Remote Endpoint Learning

A remote endpoint is the IP address of a server on a leaf different from the leaf where the server is located. Cisco ACI leafs learn the remote-endpoint IP addresses in order to optimize policy CAM filtering on the very ingress leaf where traffic is sent from the server to the fabric.

With VRF enforcement direction configured for ingress (which is the default), Cisco ACI optimizes the policy CAM filtering for traffic between the fabric and the L3Out, by making sure that the filtering occurs on the leaf where the endpoint is and not on the border leaf. This is to keep the border leaf from becoming a bottleneck for policy CAM filtering when it has to store all of the policy CAM entries for all of the conversations between endpoints of the fabric leafs and the outside.

With VRF enforcement direction configured for ingress, the border leaf does not learn the source IP address of the server sending traffic to the outside via L3Out. The border leaf still learns the source IP address of the server from the traffic between the server on another leaf and a server connected to the border leaf. With certain traffic patterns, this can generate stale entries on the border leaf, where the border leaf may have the entry for the IP address of the remote server associated with a leaf even if the remote server has moved to another leaf.

This document explains this scenario:

<https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-739989.html>

The section titled “Using border leafs for server attachment” already explained that configuring “Disable Remote Endpoint Learning” solves the above problem.

The “Disable Remote Endpoint Learning” configuration option disables the learning of remote endpoint IP addresses only on border leaf switches that have a VRF instance with ingress policy enabled. This configuration option does not change the learning of the MAC addresses of the endpoints.

The recommendation at the time of this writing is that if you deploy a topology that connects to the outside through border leaf switches that are also used as computing leaf switches; if the VRF instance is configured for ingress policy (which is the default), you should disable remote endpoint learning on the border leaf switches.

Depending on the Cisco ACI version, 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
- Or from: System > System Settings > Fabric Wide Setting > Disable Remote EP Learning

With this option, the IP addresses of the remote multicast sources are still learned, as a result if a server is sending both unicast and multicast traffic and then it moves, unicast traffic won't update the entry in the border leaf, so this could result in stale entries with Cisco ACI versions earlier than ACI 3.2(2).

This option is useful for both first-generation and second-generation leafs.

Note: The section titled “Dataplane learning configurations in Cisco ACI” provides additional information about the configuration options to tune dataplane learning.

Enforce Subnet Check

Cisco ACI offers two similar configurations related to limiting the dataplane learning of endpoints' IP addresses:

- Per-BD limit IP-address learning to subnet
- Global Enforce Subnet Check knob

Enforce Subnet Check ensures that Cisco ACI learns endpoints whose IP addresses belong to the bridge domain subnet; it also ensures that leafs learn remote entries whose IP addresses belong to the VRF that they are associated with. This prevents the learning of IP addresses that are not configured as subnets on the bridge domains of the VRF.

This option is under System Settings > Fabric Wide Settings. For more information please refer to this URL:

<https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-739989.html>

Enabling Enforce Subnet Check clears all of the remote entries and prevents learning remote entries for a short amount of time. The entries in the spine-proxy are not cleared, hence traffic forwarding keeps working even during the configuration change.

Note: While no disruption is expected when enabling Enforce Subnet Check, there is the possibility that a given network is working with traffic from subnets that do not belong to the VRF. If this is the case, enabling this feature will cause interruption of these traffic flows.

Rogue Endpoint Control

The rogue endpoint control feature is described in the section titled “Choosing between endpoint move dampening, endpoint loop protection, and rogue endpoint control.”

Rogue endpoint control is useful in case of server misconfigurations, such as incorrect NIC Teaming configuration on the servers.

This option should be enabled under System Settings > Endpoint Controls. If you configure endpoint loop protection, and, as an action, you choose to disable the ports, you may also want to configure the error-disabled recovery policy, which is in Fabric > External Access Policy > Policies > Global Policies > Error Disabled Recovery Policy

Enable IP Aging

This configuration is useful to age individual IP addresses that may be associated with the same MAC address (for instance, in the case of a device that does NAT and is connected to Cisco ACI). This option is under System Settings > Endpoint Controls. For more information please refer to this URL:

<https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-739989.html>

Enabling domain validations

Cisco ACI has a feature that verifies whether the VLAN used in an EPG matches the AEP configured, that there are no overlaps, and so on.

This feature is enabled at this level: System > Fabric Wide Settings > Enforce Domain Validation.

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 35 illustrates the concept.

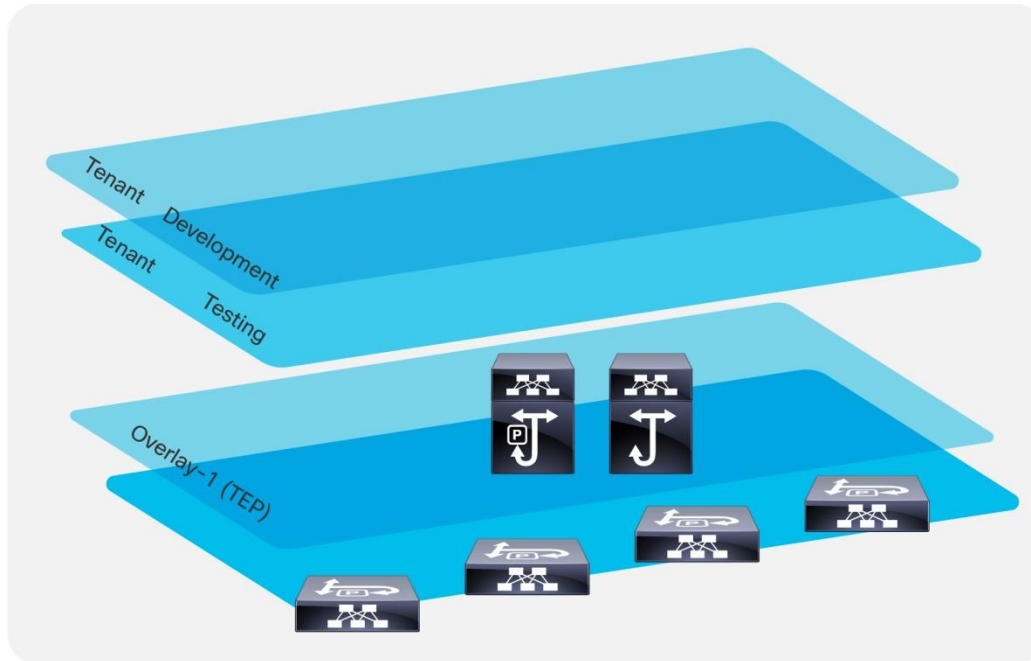


Figure 35.
Tenants are logical divisions of the fabric

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 35, 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 dataplane isolation function. Figure 36 illustrates the relationship among the building blocks of a tenant.

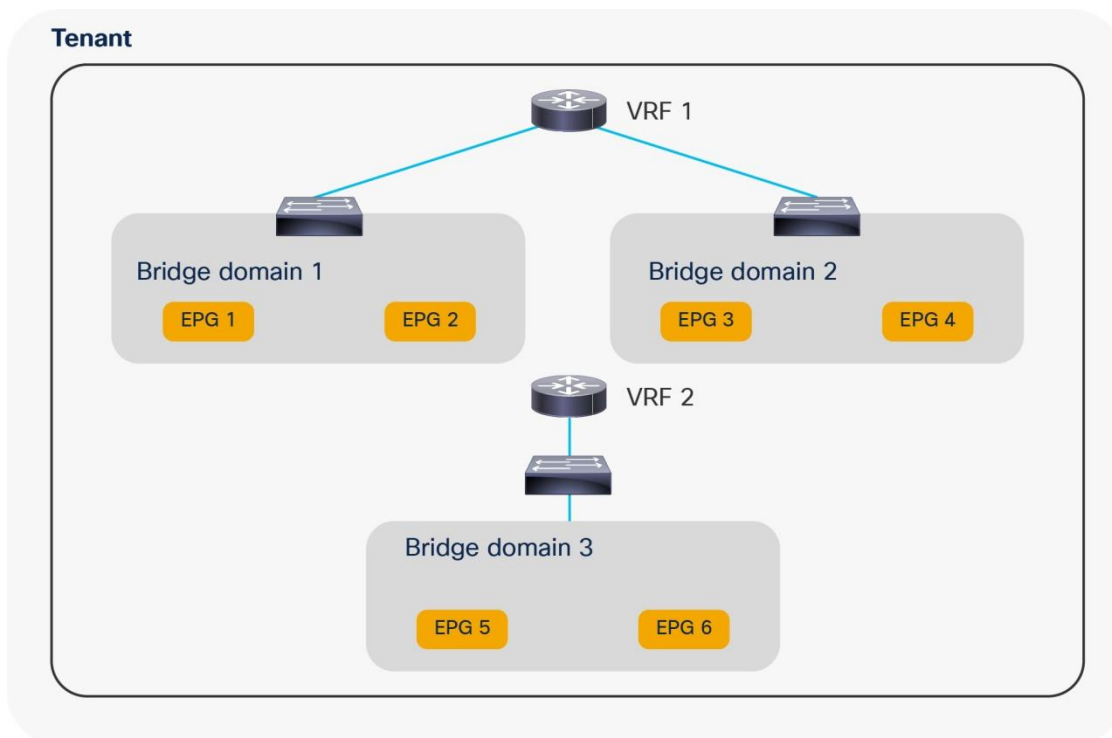


Figure 36.
Hierarchy of tenants, private networks (VRF instances), bridge domains, and EPGs

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 Access Control Lists (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 for 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.

Network-centric and application-centric designs

This section clarifies two commonly used terms to define and categorize how administrators configure Cisco ACI tenants.

If you need to implement a simple topology, you can create one or more bridge domains and EPGs and use the mapping 1 bridge domain = 1 EPG = 1 VLAN. This approach is commonly referred to as a network-centric design.

You can implement a Layer 2 network-centric design where Cisco ACI provides only bridging or a Layer 3 network-centric design where Cisco ACI is used also for routing and to provide the default gateway for the servers.

If you want to create a more complex topology with more security zones per bridge domain, you can divide the bridge domain with more EPGs and use contracts to define ACL filtering between EPGs. This design approach is often referred to as an application-centric design.

These are just commonly used terms to refer to a way of configuring Cisco ACI tenants. There is no restriction about having to use only one approach or the other. A single tenant may have bridge domains configured for a network-centric type of design and other bridge domains and EPGs configured in an application-centric way.

Figure 37 illustrates the two concepts:

- In the network-centric design, there is a 1:1 mapping between bridge domain, EPG, and VLAN.
- In the application-centric design, the bridge domain is divided into EPGs that represent, for instance, application tiers: “web” and “app,” or, more generally, different security zones.

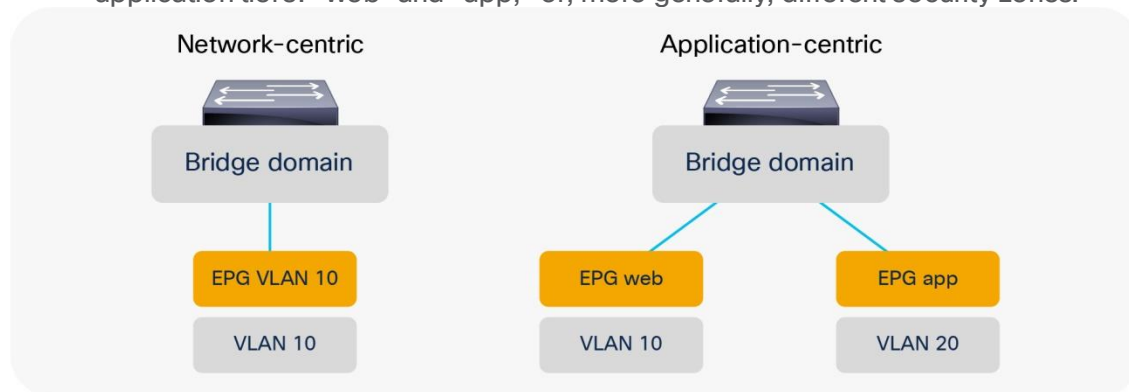


Figure 37.
Network-centric and application-centric designs

Implementing a network-centric topology

If you need to implement a topology with simple segmentation, you can create one or more bridge domains and EPGs and use the mapping 1 bridge domain = 1 EPG = 1 VLAN.

You can then configure the bridge domains for unknown unicast flooding mode (see the section “Bridge domain design considerations” for more details).

In the Cisco ACI object model, the bridge domain has to have a relation with a 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.

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, ACI tries to resolve the relation to objects from Tenant common and with the default settings the object will not be instantiated and APIC will raise a Fault (F0467) on the EPG associated to the BD that has been created without an explicit relation to a VRF. For instance, if you don't associate a BD with a VRF, APIC automatically associates your newly created BD with the VRF from Tenant common (common/default). Whether this association is enough to enable bridging and/or routing from the BD depends on the configuration of the Instrumentation Policy (Tenant common / Policies / Protocol Policies / Connectivity Instrumentation Policy).

Default gateway for the servers

With this design, the default gateway can be outside of the Cisco ACI fabric itself, or Cisco ACI can be the default gateway.

In order to make Cisco ACI the default gateway for the servers you need to configure the bridge domain with a “Subnet” and enable IP routing in the bridge domain.

Making Cisco ACI the default gateway, hence using Cisco ACI for routing traffic, requires a minimum understanding of how Cisco ACI learns the IP addresses of the servers and how it populates the mapping database.

Before moving the default gateway to Cisco ACI, make sure you verify whether the following type of servers are present:

- Servers with active/active transmit load-balancing teaming
- Clustered servers where multiple servers send traffic with the same source IP address
- Microsoft Network Load Balancing servers

If these types of servers are present, you should first understand how to tune dataplane learning in the bridge domain before making Cisco ACI the default gateway for them.

Assigning servers to endpoint groups

In order to connect servers to a bridge domain, you need to define the endpoint group and to define which leaf, port, or VLAN belongs to which EPG. You can do this in two ways:

- From tenant, application profile, EPG, or static port
- From fabric access, AAEP, or application EPG

Layer 2 connectivity to the outside

Connecting Cisco ACI to an external Layer 2 network with a network-centric design is easy because the bridge domain has a 1:1 mapping with a VLAN, thus there is less risk of introducing loops by merging multiple Layer 2 domains via a bridge domain.

The connectivity can consist of a vPC to an external Layer 2 network, with multiple VLANs, each VLAN mapped to a different bridge domain and EPG.

The main design considerations with this topology are:

- Avoiding introducing loops by using only one EPG per BD to connect to the external Layer 2 network. This is normally easy to achieve in a network-centric design because there is only one EPG per BD.
- Avoiding traffic blackholing due to missing Layer 2 entries in the mapping database. This is achieved by configuring the bridge domain for unknown unicast flooding instead of hardware-proxy.
- Limiting the impact of TCN BPDUs on the mapping database

You can limit the impact of TCN BPDUs on the mapping database by doing one of two things:

- If the external network connects to Cisco ACI in an intrinsically loop-free way (for example, via a single vPC), you can consider filtering BPDUs from the external network.
- If the external network connectivity to Cisco ACI is kept loop-free by Spanning Tree Protocol, then you should reduce the impact of TCN BPDUs on the mapping database by making sure that the external Layer 2 network uses a VLAN on the EPG that is different from the VLAN used by servers that belong to the same EPG and are directly attached to Cisco ACI.

For more information, please refer to the sections titled: “Connecting to existing networks and migration designs” and “Connecting EPGs to external switches”.

ACL filtering

For a simple network-centric Cisco ACI implementation, initially you may want to define a permit any-any type of configuration where all EPGs can talk. This can be done in three ways:

- Configuring the VRF for unenforced mode
- Enabling Preferred Groups and putting all the EPGs in the Preferred Group
- Configuring vzAny to provide and consume a permit-any-any contact

Figure 38 illustrates the three options.

The first option configures the entire VRF to allow all EPGs to talk to each other.

Preferred Groups let you specify which EPGs can talk without contracts; you can also put EPGs outside of the Preferred Groups. In order to allow servers in the EPGs outside of the Preferred Group to send traffic to EPGs in the Preferred Group, you need to configure a contract between the EPGs.

The third option consists of making vzAny (also known as EPG Collection for VRF) a provider and consumer of a permit-any-any contract.

The second and third approach are the most flexible because they make it easier to migrate to a configuration with more specific EPG-to-EPG contracts:

- If you used the Preferred Group, you can, in the next phase, move EPGs outside of the Preferred Group and configure contracts.
- If you used vzAny, you can, in the next phase, either add a redirect to a firewall instead of a permit, in order to apply security rules on the firewall, or you can add more specific EPG-to-EPG contracts with an allowed list followed by a deny, in order to gradually add more filtering between EPGs. This is possible because in Cisco ACI, more specific EPG-to-EPG rules have priority over the vzAny-to-vzAny rule.

For more information about contracts, please refer to the section titled “Contract design considerations.”

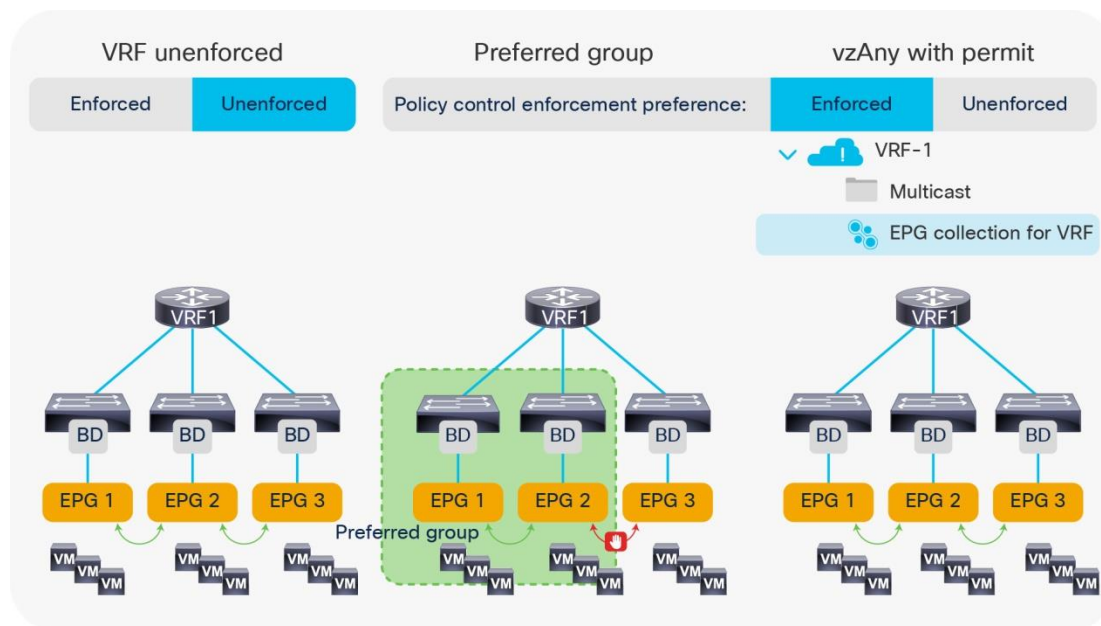


Figure 38.
Contract filtering with a network-centric design

Implementing a tenant design with segmentation(application-centric)

If you implement a Cisco ACI design with segmentation of bridge domain in multiple EPGs, the following design considerations apply:

- Define how many security zones you want to introduce into the topology.
- Use Cisco ACI as the default gateway and configure the bridge domain for hardware-proxy to optimize unknown unicast flooding. If the bridge domain connects to an external Layer 2 network use the unknown unicast flooding option instead.
- When Cisco ACI is the default gateway for the servers, make sure you know how to tune dataplane learning for the special cases of NIC Teaming active/active, for clustered servers, and for MNLB servers.
- Carve as many EPGs per bridge domain based on the number of security zones, keeping in mind the verified scalability limits for EPGs and contracts.
- You have to use a different VLAN (or different VLANs) for each EPG in the same bridge domain on the same leaf. In practice, you should try to use a different VLAN for each EPG in the same bridge domain. VLAN re-use is possible on different bridge domains on the same leaf (there is more on this in the section titled “EPG-to-VLAN mapping”).
- Make sure that the number of EPG+BDs utilized on a single leaf is less than the verified scalability limit. At the time of this writing, the maximum number of EPG+BDs per leaf is 3960.
- Make sure you understand contract rules priorities in order to define correctly the EPG-to-EPG filtering rules by using permit, deny, and optionally service graph redirect.
- You can change the default action for traffic between EPGs in the VRF to be permit or redirect to a firewall by using vzAny with contracts.
- Configure policy CAM compression for contract filters.

EPG and bridge domain considerations

When migrating to an application-centric design, you should start by defining how many security zones you need to define in the tenant network.

Let’s assume, for instance, that you need three security zones: IT, non-IT, and shared services, as in Figure 39.

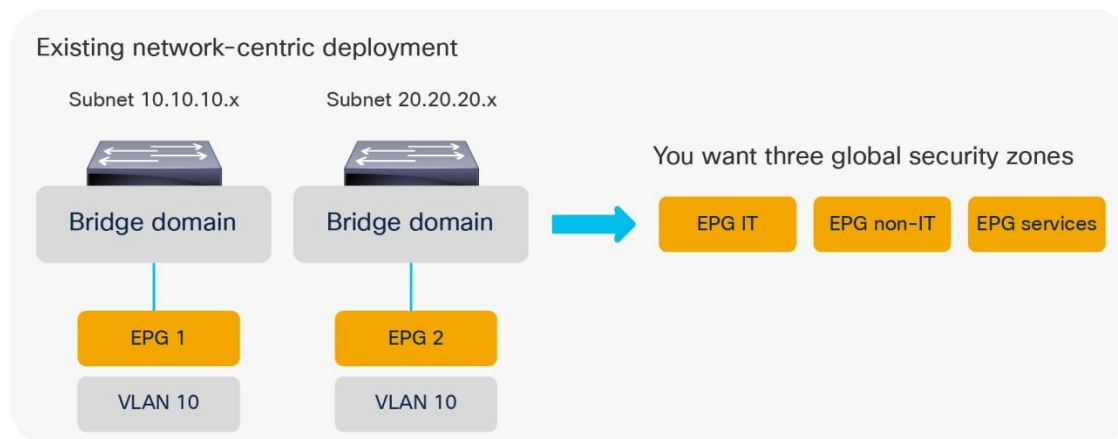


Figure 39.
Migrating from a network-centric design to an application-centric design

You can follow two approaches to introduce these security zones:

- Simply adding an IT-EPG to BD1 and BD2, BD3, etc., which results in a total number of EPGs that is equal to the number of security zones times the number of bridge domains, as in Figure 40.
- Reducing the number of bridge domains by merging them, ideally into one single bridge domain and adding three EPGs to the single bridge domain, as in Figure 41.

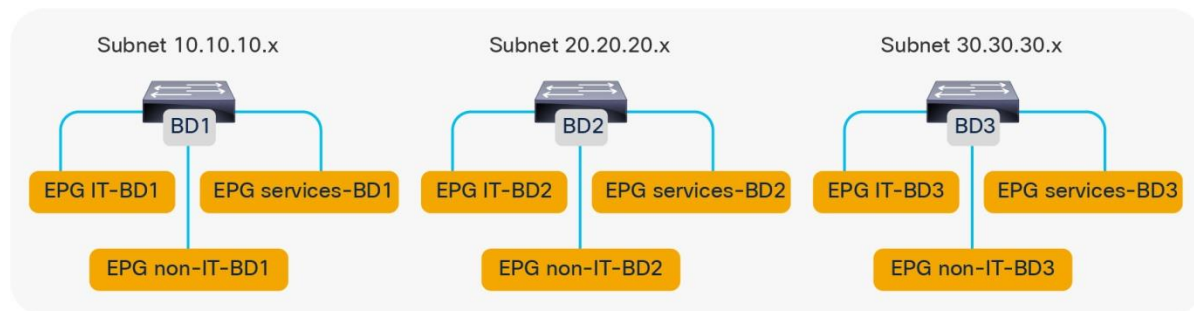


Figure 40.
Creating as many EPGs as security zones in each Bridge Domain (BD)

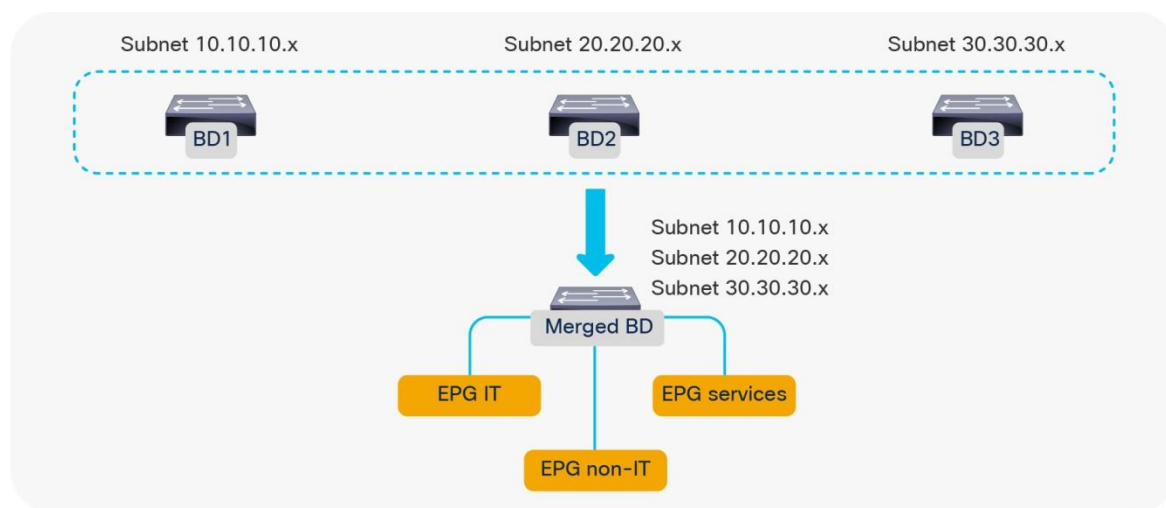


Figure 41.
Reducing the number of bridge domains and creating three EPGs

Adding EPGs to existing bridge domains

The approach of creating additional EPGs in the existing bridge domains has the advantage of maintaining an existing Layer 2 design or bridge domain configuration by just adding security zones.

The disadvantages of adding EPGs to bridge domains are mostly related to scale and manageability:

- At the time of this writing, the validated number of EPG+BDs per leaf is 3960.
- The number of EPGs and contracts can also grow significantly.

With many bridge domains, you are likely going to have many EPGs, and if all EPGs need to talk to all EPGs, the hardware consumption of the policy CAM entry becomes, in the order of magnitude of $\#EPGs * (\#EPG - 1) *$, the number of filters, because of all of the EPG pairs that need to be defined.

The verified scalability guide states that a single EPG providing one contract consumed by 1000 EPGs is a validated design. The verified scalability guide also states that the validated scale for multiple EPGs providing the same contract is a maximum of 100 EPGs, and the maximum number of EPGs consuming the same contract (provided by multiple EPGs) is 100 as well.

Merging bridge domains

With the approach of merging bridge domains into one, the number of EPGs and contracts is more manageable, but, because all EPGs and VLANs are in the same bridge domain, it may be necessary to use the flooding optimization features that Cisco ACI offers.

Flood in Encapsulation is a feature that can be used on -EX leafs and newer that lets you scope the flooding domain to the individual VLANs on which the traffic is received. This is roughly equivalent to scoping the flooding to the EPGs.

Designs based on merged bridge domains with Flood in Encapsulation have the following characteristics:

- Cisco ACI scopes all unknown unicast and multicast flooded traffic, broadcast traffic, and control plane traffic in the same VLAN.
- Cisco ACI performs proxy-ARP in order to forward traffic between servers that are in different VLANs. Because of this, traffic between EPGs (or rather between different VLANs) is routed even if the servers are in the same subnet.
- Flood in Encapsulation also works with VMM domains if the transport is based on VLANs and VXLANs. The support for VXLAN is available starting from Cisco ACI 3.2(5).

For more details, please refer to the section “Bridge domain design considerations”.

When using a single bridge domain with multiple subnets, the following considerations apply:

- The DHCP server configuration may have to be modified to keep into account that all DHCP requests are originated from the primary subnet.
- Cisco ACI works fine with a large number of subnets under the same bridge domain, as described in the Verified Scalability Guide. The number that is validated at the time of this writing is 1000 subnets under the same Bridge Domain (BD) with normal flooding configurations and 400 subnets with Flood in Encapsulation, but when using more than ~200 subnets under the same BD, configuration changes performed to individual BDs in a nonbulk manner (for instance, via GUI or CLI configurations) can take a great deal of time to be applied to the fabric.

Contract design considerations

After dividing the bridge domains in security zones, you need to add contracts between them. The contract configuration can follow two approaches:

- Adding individual contracts between EPGs, with a default implicit deny
- Adding individual contracts between EPGs, with each contract consisting of an allowed list with a deny-any-any entry and with a global vzAny permit
- Configuring vzAny with a contract to redirect all traffic to an external firewall and using specific EPG-to-EPG contracts for specific traffic

The first approach is the allowed list approach, where all traffic is denied unless there is a specific contract to permit EPG-to-EPG traffic. With this approach, it is beneficial to reduce the number of bridge domains—and, as a result, the number of EPGs—for a more scalable and manageable solution.

With the second approach, you can have a default permit that allows all-EPGs-to-all-EPGs traffic, and you can add more specific contracts between EPGs where individual rules permit specific protocols and ports, and there is a deny-any-any rule per EPG pair (that is, per contract). This allows migration to an application-centric approach from a design where there was no ACL filtering in place. With this approach, it is beneficial to reduce the number of bridge domains (and, as a result, the number of EPGs) for a more scalable and manageable solution.

The third approach consists of configuring vzAny as a provider and consumer of a contract with service graph redirect to one or more firewalls. With this approach, any EPG-to-EPG traffic (even within the same bridge domain) is redirected to a firewall for ACL filtering. This approach uses Cisco ACI for segmentation and the firewall for ACL filtering. You can configure EPG-to-EPG specific contracts that have higher priority than the vzAny with redirect to allow, for instance, backup traffic directly via the Cisco ACI fabric without sending it to a firewall.

With this approach, one can still keep many bridge domains and create multiple EPGs in each one of them without too much operational complexity. This is because contract enforcement is performed by the external firewall in a centralized way and only one contract is required in the Cisco ACI fabric for all of the EPGs under the same VRF.

Figure 42 illustrates this approach.

A pair of firewalls or more (you can cluster several firewalls with symmetric policy-based routing [PBR] hashing) is connected to the Cisco ACI fabric. By using vzAny in conjunction with a service graph redirect attached to a contract, all traffic between EPGs is redirected to the firewall pair. For instance, traffic between EPG IT-BD1 and non-IT-BD1 has to go via the firewall first and, similarly, traffic between EPG non-IT-BD1 and services-BD1.

Figure 43 illustrates the configuration of vzAny with the contract to redirect the traffic.

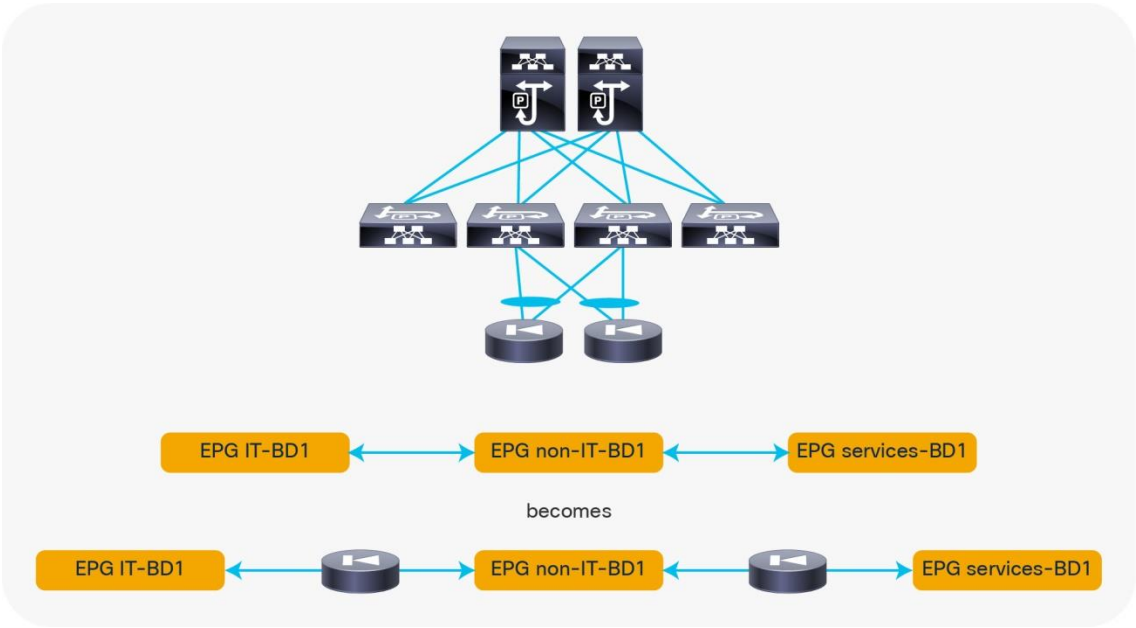


Figure 42.
Using vzAny with redirect to use Cisco ACI with an external firewall

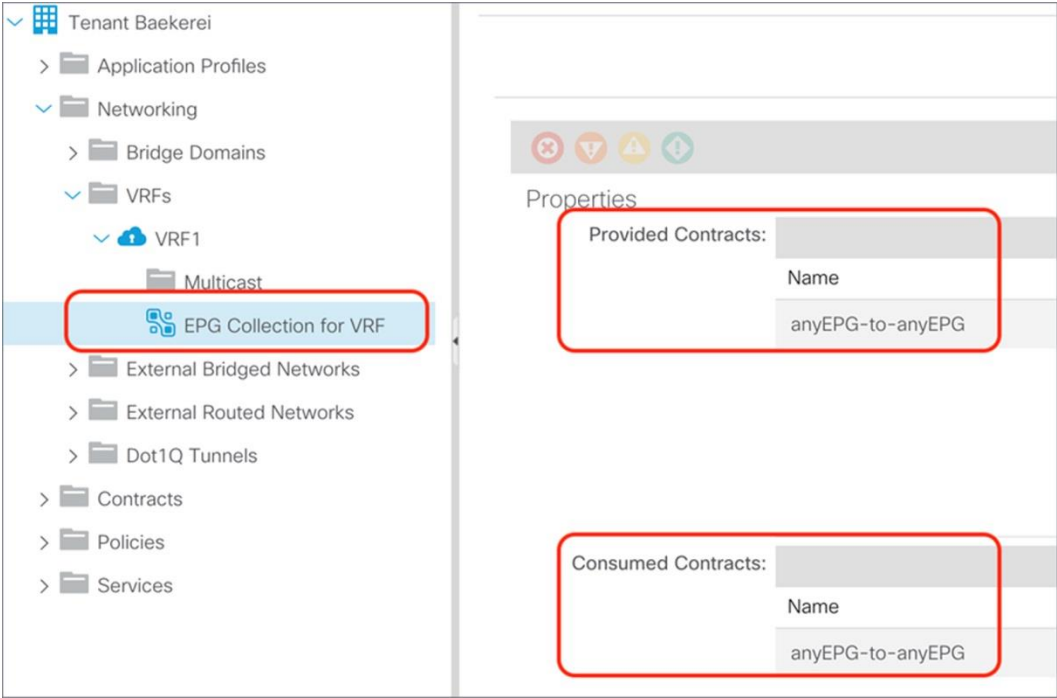


Figure 43.
Configuring vzAny to redirect traffic to an external firewall

With application-centric deployments, the policy CAM is more utilized than with network-centric deployments because of the number of EPGs, contracts, and filters.

Depending on the leaf hardware, Cisco ACI offers many optimizations to either allocate more policy CAM space or to reduce the policy CAM consumption:

- Cisco ACI leafs can be configured for policy-CAM-intensive profiles
- Range operations use one entry only in TCAM
- Bidirectional subjects take one entry
- Filters can be reused with an indirection feature (at the cost of granularity of statistics)

Figure 44 illustrates how to enable policy CAM compression when configuring filters.

The screenshot shows two panels in the Cisco ACI configuration interface. The 'Add Filter' panel on the left has a 'Filter' dropdown set to 'select a value', 'Directives' with checkboxes for 'Log' and 'Enable Policy Compression' (the latter is checked), and an 'Action' section with 'Deny' and 'Permit' buttons. The 'Filters' panel on the right shows a table with columns 'Name' and 'Directives'. The 'Name' dropdown is set to 'select an option', and the 'Directives' dropdown is set to 'Enable Policy Compression'.

Figure 44.
Enabling compression on filters

Note: Policy CAM compression cannot be enabled/disabled on contracts that are already programmed in hardware. If you need to enable/disable policy compression, you should create a new contract and use it to replace the pre-existing one.

Note: For more information about contracts, please refer to the section: “Contract design considerations.”

VRF design considerations

The VRF is the dataplane 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 an 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.

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. The default is “ingress,” which means that the “ingress” leaf filters the traffic from the Cisco ACI fabric to the L3Out, and traffic from the L3Out to servers connected to the Cisco ACI fabric is filtered on the leaf where the server is connected.

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/support/cloud-systems-management/application-policy-infrastructure-controller-apic/tsd-products-support-series-home.html#Verified_Scalability_Guides

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 45).

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.

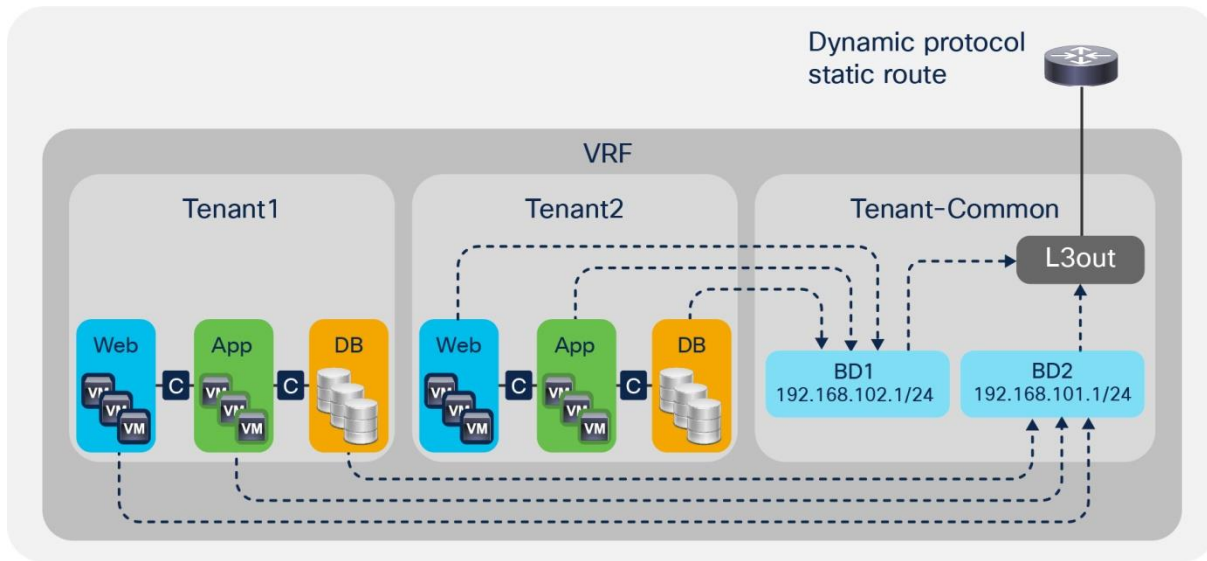


Figure 45.

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 46). 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.

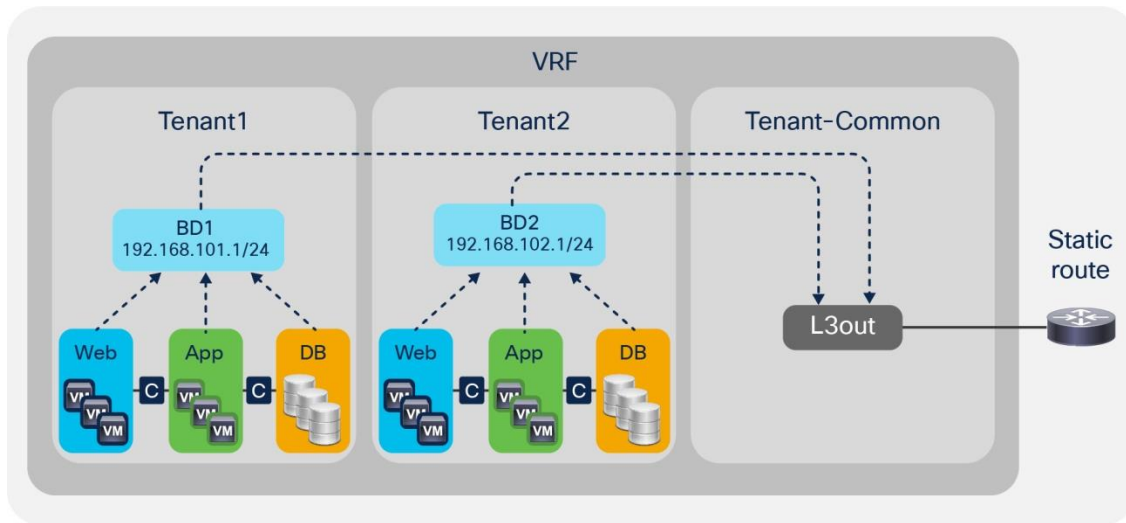


Figure 46.
Shared L3Out connection with the VRF instance in the common tenant

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

The VRF can be configured for ingress policy enforcement or egress policy enforcement. This configuration controls whether the ACL filtering performed by contracts that are configured between L3ext and EPGs is implemented on the leaf where the endpoint is or on the border leaf.

You can configure the VRF instance for ingress or egress policy by selecting the Policy Control Enforcement Direction option Egress under Tenants > Networking > VRFs.

The configuration options do the following:

- Ingress policy enforcement means that the ACL filtering performed by the contract is implemented on the leaf where the endpoint is located. This configuration makes the policy CAM of the border leaf less utilized because the policy CAM filtering rules are configured on the “compute” leafs. With ingress policy enforcement, the filtering happens consistently on the “compute” leaf for both directions of the traffic.
- Egress policy enforcement means that the ACL filtering performed by the contract is also implemented on the border leaf. This makes the policy CAM of the border leaf more utilized. With egress policy enforcement, the border leaf does the filtering for the L3Out-to-EPG direction after the endpoint has been learned as a result of previous traffic; otherwise, if the endpoint to destination class mapping is not yet known on the border leaf, the policy CAM filtering happens on the compute leaf.

The ingress policy enforcement feature improves policy CAM utilization on the border leaf nodes by distributing the filtering function across all regular leaf nodes, but it distributes the programming of the L3ext entries on all the leafs. This is mostly beneficial in case you are using first-generation leafs and in the case where the L3ext table is not heavily utilized. The egress policy enforcement feature optimizes the use of entries for the L3ext by keeping the table configured only on the border leaf.

Some features scale or work better with ingress filtering and other features work better with egress filtering.

At the time of this writing (as of Cisco ACI 4.0), most features work better with, and some require, ingress filtering. The features that at the time of this writing require ingress filtering are:

- IP-EPG
- Direct Server Return
- GOLF
- Multi-Site with L4-L7 service graph based on PBR

Some other features such as Quality of Service (QoS) on the L3Out, require, instead, egress filtering.

As already discussed in the section titled “Placement of outside connectivity / using border leafs for server attachment,” the use of ingress policy for VRF instances and attachment of endpoints to border leaf switches is fully supported when all leaf switches in the Cisco ACI fabric are second-generation, such as the Cisco Nexus 9300-EX and -FX platform switches.

It is also recommended to 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, as described in the section titled “Global configurations / Disable Remote Endpoint Learning.”

Bridge domain design considerations

The main bridge domain configuration options that should be considered when tuning bridge domain behavior are as follows:

- Whether to use hardware proxy or unknown unicast flooding
- Whether to enable or disable Address Resolution Protocol (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
- Whether to use Flood in Bridge Domain with Flood in Encapsulation

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 attempts 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:** If this option is selected, the fabric does not learn IP addresses from a subnet other than the one configured on the bridge domain. If Enforce Subnet Check is enabled globally, this option is not necessary.

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. When changing the bridge domain configuration, be sure to keep in mind that this change can cause traffic disruption.

Bridge domain configuration for 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 in the bridge domain.

Consider the topology of Figure 47. The reason for using unknown unicast flooding instead of hardware proxy in the bridge domain is that 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 leaf 1 and leaf 2 may trigger the learning of the MAC addresses of the servers connected to switch A and B because they would perform an ARP address resolution 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 leaf 2 still have valid ARP entries for the hosts connected to switch A and switch B, so they will not perform an ARP address resolution again 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 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.

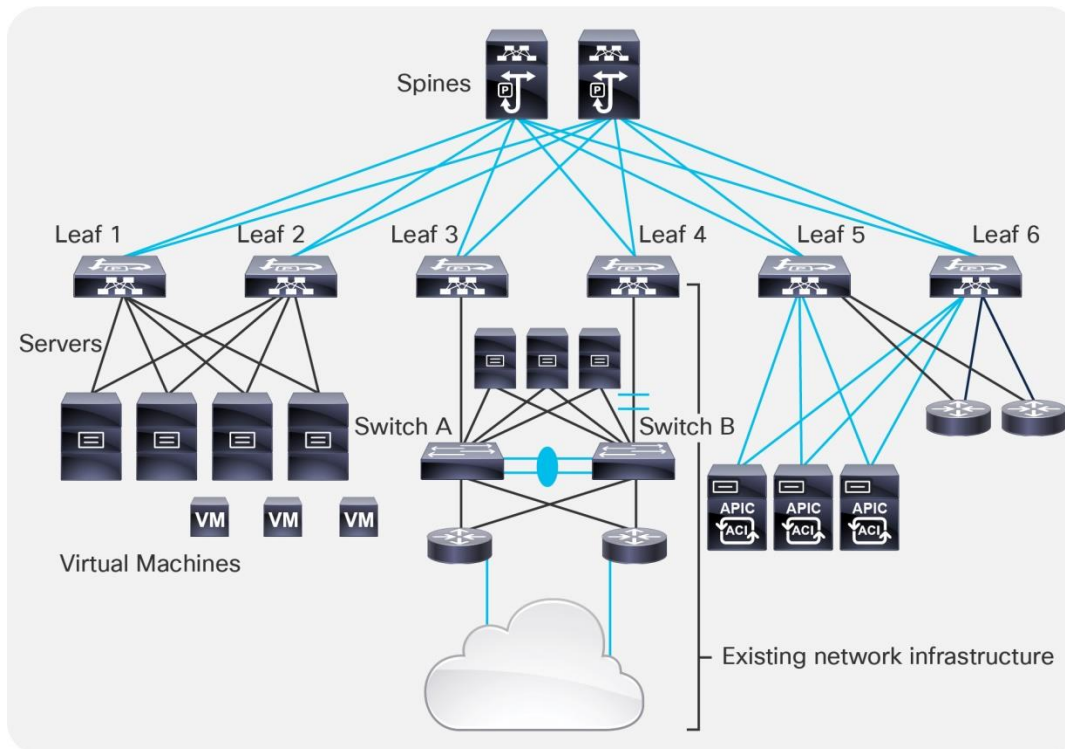


Figure 47.

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. Selecting 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).

Bridge domain flooding

By default, bridge domains are configured for Flood in Bridge Domain. This configuration means that, when a multdestination frame (or an unknown unicast with unknown unicast flooding selected) is received from an EPG on a VLAN, it is flooded in the bridge domain.

Consider the example shown in Figure 48. 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 from all ports that are in the bridge domain across all EPGs, regardless of the VLAN encapsulation.

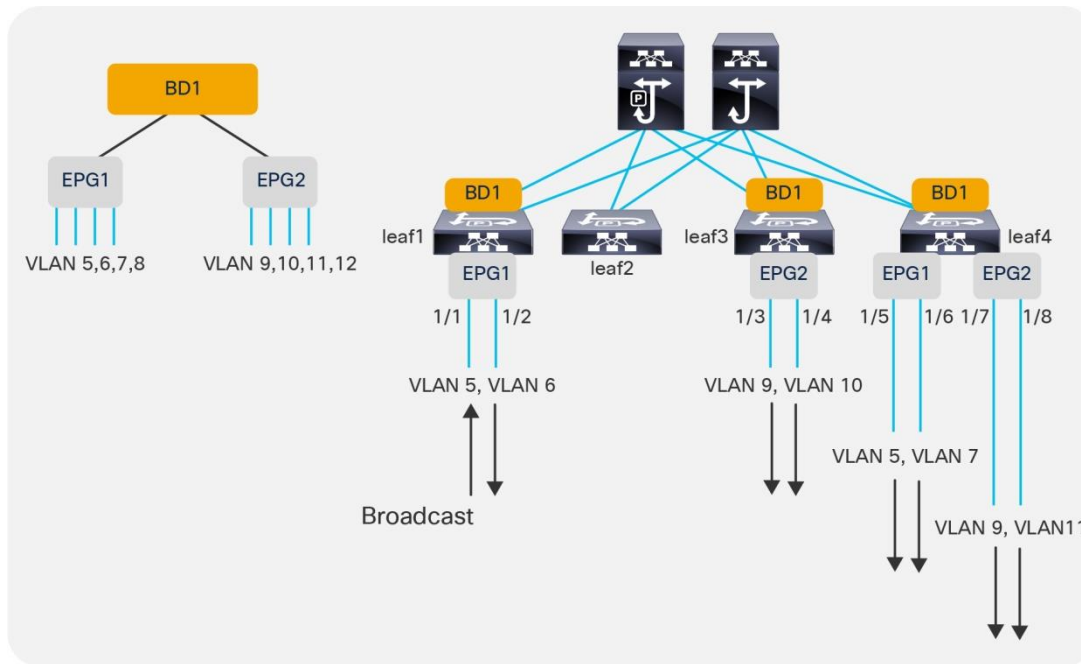


Figure 48.
Bridge domain, EPGs, and VLANs

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 Protocol 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 bridge domain that have the **same encapsulation**.

As Figure 49 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.

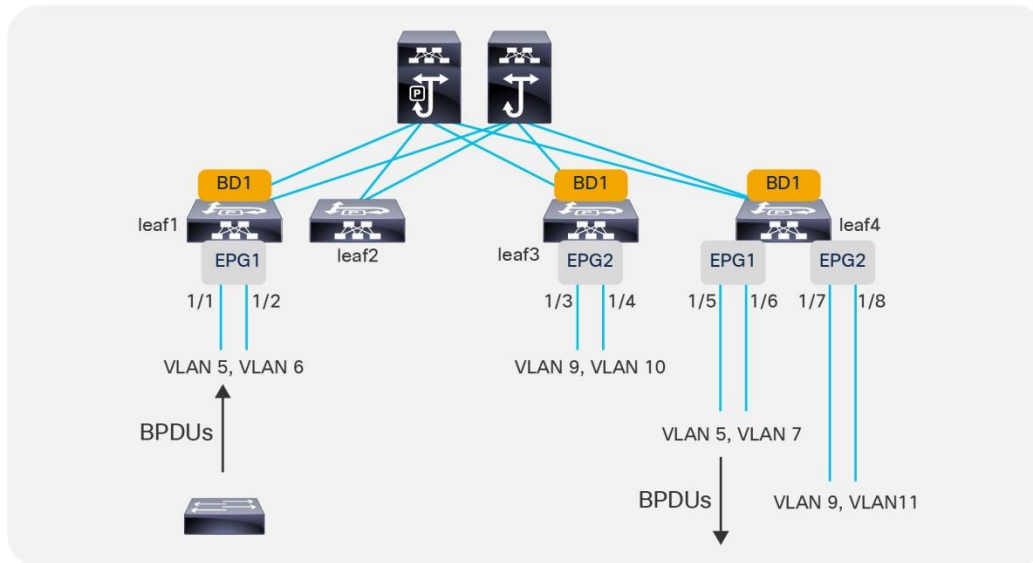


Figure 49.
BPDU flooding in the fabric

Flood in Encapsulation

Flood in Encapsulation is a feature that can be useful when merging multiple existing Layer 2 domains into a single bridge domain and you want to scope the flooding domain to the VLAN from which the traffic came.

MAC addresses in different VLANs that are in the same bridge domain must be unique.

Flood in Encapsulation is a feature that can be used on -EX leafs and newer that lets you scope the flooding domain to the individual VLANs that the traffic is received on.

With Flood in Encapsulation, Cisco ACI floods packets to all of the EPGs having the same VLAN encapsulation with encapsulations coming from same “namespace” (that is, from the same VLAN pool under the same domain). Because normally you would not use the same VLAN on different EPGs in the same bridge domain, this is roughly equivalent to scoping the flooding to the EPGs.

Flood in Encapsulation, starting from Cisco ACI 3.1, is able to scope the flooding by limiting flooding of the following:

- Multicast traffic
- Broadcast traffic
- Link-local traffic
- Unknown unicast traffic
- Protocols: OSPF, EIGRP, etc.

Designs based on merged bridge domains with Flood in Encapsulation have the following characteristics:

- Cisco ACI scopes all unknown unicast and multicast flooded traffic, broadcast traffic, and control plane traffic in the same VLAN
- Cisco ACI performs proxy-ARP in order to forward traffic between servers that are in different VLANs. Because of this, traffic between EPGs (or rather between different VLANs) is routed even if the servers are in the same subnet.
- Flood in Encapsulation also works with VMM domains if the transport is based on VLANs and VXLANs. The support for VXLAN is available starting from Cisco ACI 3.2(5).

The following features either do not work in conjunction with the bridge domain where Flood in Encapsulation is enabled or have not been validated

- IPv6
- Multicast Routing

Flood in Encapsulation has the following requirements for the bridge domain configuration:

- IP routing must be enabled also for L2 communication between EPGs that are in the same subnet.
- The option for optimizing ARP in the bridge domain (no ARP flooding) cannot be used.

With Flood in Encapsulation, multicast is flooded only on the ports that are on the same VLAN as the incoming traffic. Even if Internet Group Management Protocol (IGMP) snooping is on, the multicast is flooded on the ports in the same encapsulation, the scope of the flooding is dependent on IGMP reports received per leaf. If there was an IGMP report on that specific leaf, traffic is sent to that port only if it is in the same encapsulation.

With Flood in Encapsulation, given that ARP packets are sent to the CPU, there is the risk that one link could use all of the aggregate capacity that the global COPP allocated for ARP. Because of this, it is recommended to enable per protocol per interface COPP to ensure fairness among the ports that are part of the EPG / bridge domain.

You can find more information about Flood in Encapsulation at this link:

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/2-x/L2_config/b_Cisco_APIC_Layer_2_Configuration_Guide/b_Cisco_APIC_Layer_2_Configuration_Guide_chapter_010.html#id_59068

Using hardware-proxy to reduce flooding

Cisco ACI offers the following features to limit the amount of flooding in the bridge domain:

- Flood in Encapsulation, which is designed to scope the flooding domains to EPG/VLANs
- Hardware-proxy, which is, instead, focused on optimizing flooding for unknown unicast traffic while keeping the bridge domain as the flooding domain for other multidestination traffic

You should consider using hardware-proxy when the bridge domain has IP routing enabled and a subnet is defined. This is because, with hardware-proxy on, if a MAC has been aged out in the spine-proxy, traffic destined to this MAC is dropped. For Cisco ACI to maintain an up-to-date mapping database, Cisco ACI must perform an ARP address resolution of the IP addresses of the endpoints; this also refreshes the MAC address table.

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.
- 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.

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 differs.

If you change the bridge domain settings to hardware-proxy, and the ARP entry on the hosts does not 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.

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 both ARP flooding and 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 before reading the "Dataplane learning" section of this document.

Layer 2 multicast and IGMP snooping in the bridge domain

Cisco ACI forwards multicast frames on the overlay multicast tree that is built between leafs and spines.

The Cisco ACI forwarding configuration options control how the frames are forwarded on the leafs.

Cisco ACI forwarding for nonrouted multicast traffic works as follows:

- Layer 2 Multicast frames—that is, multicast frames that do not have a multicast IP address—are flooded.
- Layer 3 Multicast frames—that is, multicast frames with a Multicast IP address: the forwarding in the bridge domain depends on the configurations of the bridge domain.

The following two bridge domain configurations allow optimizing the Layer 2 forwarding of IP multicast frames with or without IP routing enabled:

- IGMP snooping
- Optimized Flood

IGMP snooping is on by default on the bridge domain, because the IGMP snooping policy “default” that is associated with the bridge domain, defines IGMP snooping on.

It is better to define your own IGMP snooping policy so that you can change the querier configuration and the querier interval for this configuration alone without automatically changing many other configurations.

In order to have an IGMP querier, you can simply configure a subnet under the bridge domain, and you need to select the “Enable querier” option.

Cisco ACI refers to “unknown Layer 3 Multicast” as a multicast IP for which there was no IGMP report. Unknown Layer 3 multicast is a per-leaf concept, so a multicast IP is an unknown Layer 3 multicast if on a given leaf there has not been an IGMP report. If there was an IGMP report such as an IGMP join on a leaf, then the multicast traffic for that Multicast Group is not an unknown Layer 3 Multicast, and it is not flooded on the leaf if IGMP snooping is on.

If Optimized Flood is configured, and if an “unknown Layer 3 Multicast” frame is received, this traffic is only forwarded to multicast router ports. If Optimized Flood is configured and a leaf receives traffic for a multicast group for which it has received an IGMP report, the traffic is sent only to the ports where the IGMP report was received.

Cisco ACI uses the Multicast IP address to define the ports to which to forward the multicast frame, hence it is more granular than traditional IGMP snooping forwarding.

Summary of bridge domain recommendations

The recommended bridge domain configuration that works in most scenarios consists of the following settings:

- If IP routing is enabled, the subnet configured and 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.
- If you need to merge multiple Layer 2 domains in a single bridge domain, consider the use of Flood in Encapsulation

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 gateway Switch Virtual Interface (SVI) is configured on a leaf switch wherever the bridge domain of the tenant is present.

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.

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.

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, the subnet defined at the bridge domain should be used as the default gateway also with VRF sharing.

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. The subnets configured on the EPGs under the same VRF must be nonoverlapping. The subnet defined under the EPG should have the No Default SVI Gateway option selected.

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 Fabric 1 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.

Endpoint learning considerations

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.

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. Instead, it 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 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.

Limit IP Learning to Subnet

This option was already described in the section “Global configurations / Enforce Subnet Check.”

Using the Limit IP Learning to Subnet option at the BD level helps ensure that only endpoints that belong to the bridge domain subnet are learned. If Enforce Subnet Check is enabled globally this option is not necessary any more.

Before Cisco ACI 3.0, if this option was enabled on a bridge domain that was already configured for IP routing, Cisco ACI would flush all endpoints that IP learned on the bridge domain, and it would pause learning for two minutes. Starting from Cisco ACI 3.0, endpoint IPs that belong to the subnet are not flushed and learning is not paused.

Endpoint aging

If no activity occurs on an endpoint, the endpoint information is aged out dynamically based on the setting on an idle timer.

The default timer for the table that holds the host information on the leaf switches is 900 seconds.

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.

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.

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.

Endpoint retention policy at the bridge domain and VRF level

The endpoint retention policy is configurable as part of the bridge domain configuration and as part of the VRF configuration.

The same options appear:

- **Bounce Entry Aging Interval**
- **Local Endpoint Aging Interval**
- **Remote Endpoint Aging Interval**
- **Hold Interval:** This entry refers to the Endpoint Move Dampening feature.
- **Move Frequency:** This option refers to the Endpoint Move Dampening feature.

The endpoint retention policy configured at the bridge domain level controls the aging of the MAC addresses.

The endpoint retention policy configured at the VRF level controls the aging of the IP addresses.

If you do not enter any endpoint retention policy, Cisco ACI uses the one from the common tenant:

- Bounce Entry Aging Interval: 630 seconds
- Local Endpoint Aging Interval: 900 seconds
- Remote Endpoint Aging Interval: 300 seconds

The following table illustrates where to configure which option and the effect of these configurations:

Table 3. Endpoint retention policy configuration

	BD level	VRF level
Local IP	Local IP Aging	
Local MAC	Local MAC Aging	
Remote IP		Remote Endpoint Aging
Remote MAC	Remote Endpoint Aging	
Bounce IP entries		Bounce Entry Aging
Bounce MAC entries	Bounce Entry Aging	

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 does not 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.

Dataplane learning

Endpoint learning in the mapping database is used by Cisco ACI to optimize traffic forwarding (in the case of Layer 2 entries) and to implement routing of the traffic (for Layer 3 entries).

Some NIC Teaming configurations and some clustered server implementations require tuning of the dataplane learning feature in Cisco ACI.

Cisco ACI mapping database and the spine-proxy

Cisco ACI implements a mapping database, which holds the information about the MAC, IPv4 (/32), and IPv6 (/128) addresses of all endpoints and the leaf/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)

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 but not recommended 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

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 Cisco 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 VTEP 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.

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.

Bridge domain and IP routing

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

- Learning of the endpoint IPv4 or IPv6 address can occur through Address Resolution Protocol (ARP), Gratuitous ARP (GARP), and Neighbor Discovery.
- Learning of the endpoint IPv4 or IPv6 address can occur through dataplane routing of traffic from the endpoint. This is called **dataplane learning**.
- Dynamic Host Configuration Protocol (DHCP) packets can be used to learn the identity-to-location mapping.

The learning of the IP address, VRF, and VTEP of the endpoint occurs on the leaf on which the endpoint generates traffic. This IP address is then installed on the spine switches through COOP.

“Remote” entries

When traffic is sent from the leaf (leaf1) where the source endpoint is to the leaf (leaf2) where the destination endpoint is, the destination leaf also learns the IP of the source endpoint and which leaf it is on.

The learning happens as follows:

- Leaf1 forwards the traffic to the spine.
- 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 (leaf2) uses information in the VXLAN packet to update its forwarding tables 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 leaf where it is coming from.

Dataplane learning from ARP packets

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.

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.

Dataplane learning configurations in Cisco ACI

In Cisco ACI, by default, the server MAC and IP addresses are learned with a combination of control plane (ARP and DHCP) and dataplane (Layer 2 forwarding for the MAC address and routing for the IP address) learning.

It is possible to disable dataplane learning in different ways:

- **Disable Remote Endpoint Learning:** This is a global knob to disable the learning of the IP addresses for remote entries in the border leaf.
- **Disable Dataplane Learning at the VRF level:** This knob disables dataplane learning for all the IP addresses in the VRF. This disables the learning of IP addresses on the local leaf from routed traffic. This configuration also disables learning of remote IP addresses. This option can be useful when designing for the scenarios described in the section “Floating IP address considerations.”

While the preferred option is to use the knob at the VRF level, there is also a bridge domain level dataplane learning knob, which was initially introduced for use with service graph redirect on the service bridge domain. This option can be used too after consulting with Cisco and only with code starting with Cisco ACI 3.1 and only with -EX leafs.

This knob 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. This knob disables dataplane learning for a specific bridge domain only. This disables the learning of IP addresses on the local leaf from routed traffic and the learning of the MAC from the ARP traffic unless destined to the subnet IP. This configuration also disables learning of remote MAC and IP addresses

The bridge domain must be configured in hardware-proxy mode to avoid unknown unicast flooding due to the fact that MAC addresses are not learned remotely. IP multicast routing does not work on a bridge domain where dataplane learning is disabled.

If the bridge domain was previously configured with dataplane 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.

At the time of this writing, using this configuration option requires the involvement of Cisco Advanced Services to make sure that stale remote entries are cleared correctly.

Remote entries can be cleared via the CLI or via the GUI. Figure 50 illustrates how to clear remote entries from the GUI.

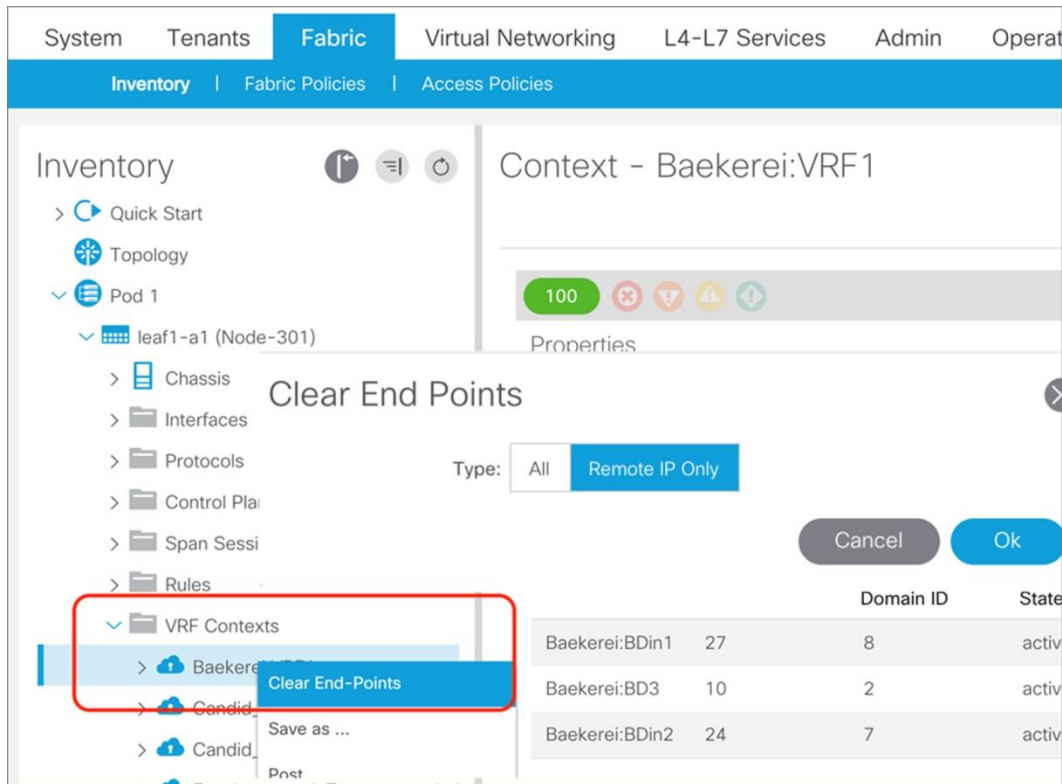


Figure 50.

State remote entries can be cleared via the GUI from the Fabric Inventory view

The dataplane can be disabled per-VRF starting with Cisco ACI 4.0. This configuration is less granular than the per-BD configuration, but it does not require manual clearing of stale remote entries. MAC addresses are still learned remotely, so there is no need to configure the bridge domain for hardware-proxy, and L3 multicast routing works.

The following table compares the option to disable Remote Endpoint (EP) Learning globally with the per-BD configuration and the per-VRF configuration.

Table 4. Dataplane learning configuration options in Cisco ACI

VRF-level Dataplane Learning	BD-level Dataplane Learning	Remote EP Learning (global)	Local MAC	Local IP	Remote MAC	Remote IP	Remote IP (Multicast)
Enabled	Enabled	Enabled	Learned	Learned	Learned	Learned	Learned
Enabled	Enabled	Disabled	Learned	Learned	Learned	Not learned on the border leaf	Learned
Disabled	N/A	N/A	Learned	Learned from ARP	Learned	Not learned	Learned
Enabled	Disabled	N/A	Learned	Learned from ARP	Not learned	Not learned	Not learned

Disabling dataplane learning is useful to integrate certain types of servers, but it has some caveats:

- With dataplane learning disabled, the mapping database is not updated continuously by the traffic; as a result, the control plane has to perform ARP address resolution for the server IP address more frequently.
- With dataplane learning enabled, the ACL filtering happens primarily on the ingress leaf by looking up the destination IP address and finding the IP-to-EPG mapping; with dataplane learning disabled, the ACL filtering happens on the egress leaf, hence there is more traffic traversing the fabric.
- Certain features, such as Rogue Endpoint Control or Anycast, which rely primarily on dataplane learning, become less effective.

For more information, please refer to this document:

<https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-739989.html>

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.

The following are examples of designs where the same IP address is associated with multiple MAC addresses:

- NIC Teaming active/active, such as Transmit Load Balancing
- Microsoft Hyper-V Switch Independent Teaming with Address Hash or Dynamic Distribution
- Designs where, in the same bridge domain, there is a firewall or load balancer with some servers using the firewall or load balancer, and other servers using the Cisco ACI bridge domain, as the default gateway.
- 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.
- Microsoft Network Load Balancing

For these scenarios you may need to consider disabling dataplane learning at the VRF or at the bridge domain level.

NIC Teaming design considerations

Nonvirtualized servers can be connected to the Cisco ACI fabric through NIC Teaming in several ways. The most commonly used NIC teaming options are:

- vPC
- Active/standby NIC Teaming
- Active/Active NIC Teaming

vPC

When you use vPC, no special tuning is required on the bridge domain because the vPC interface is logically equivalent to a single interface.

NIC Teaming active/standby

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.

NIC Teaming active/active

Servers configured with NIC Teaming active/active, such as Transmit Load Balancing, send the same source IP from multiple NIC cards with different MAC addresses.

Virtualized Servers, such as Microsoft Hyper-V with Switch Independent Teaming with Address Hash or Dynamic Load Distribution, can also send the same source IP address from multiple NICs with the MAC address of the NIC.

With these types of servers, the best connectivity option is to change the teaming to one of the following options:

Use other NIC teaming configurations for physical hosts:

- Port-Channeling with LACP in conjunction with vPC on the Cisco ACI leafs
- Active/standby NIC Teaming
- MAC pinning or equivalent, like Microsoft Hyper-V with Switch Independent Teaming Hyper-V Port

If the teaming configuration cannot be changed, you can then disable dataplane learning preferably by changing the VRF configuration, or, if absolutely necessary (for instance you cannot upgrade ACI), by changing the bridge domain configuration after having understood the caveats and asked Cisco for guidance.

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)

Assigning hosts to EPGs from the application profile EPG

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

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

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

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

- Based on VLAN encapsulation
- Based on port and VLAN
- Based on the network and mask or IP address for traffic originating outside the fabric: that is, traffic considered to be part of 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 the source IP address or subnet: 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 the source MAC address: For physical machines, this requires the hardware to support MAC-based classification and Cisco ACI 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 MAC addresses.

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 port group).
- 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 configuration), 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.

Assigning hosts to EPGs from the Attachable Access Entity Profile (AAEP)

You can configure which EPG the traffic from a port belongs to based on the VLAN that it is tagged with. This type of configuration is normally performed from the tenant configuration, but it can be tedious and error prone.

An alternative and potentially more efficient way to configure this is to configure the EPG mappings directly from the Attachable Access Entity Profile (AAEP), as described in Figure 5.

You can find more information about the configuration at this link:

[https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/2-x/basic_config/b APIC Basic Config Guide 2 x/b APIC Basic Config Guide 2 x chapter 0101.html#id 3 0752](https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/2-x/basic_config/b_APIC_Basic_Config_Guide_2_x/b_APIC_Basic_Config_Guide_2_x_chapter_0101.html#id_3_0752)

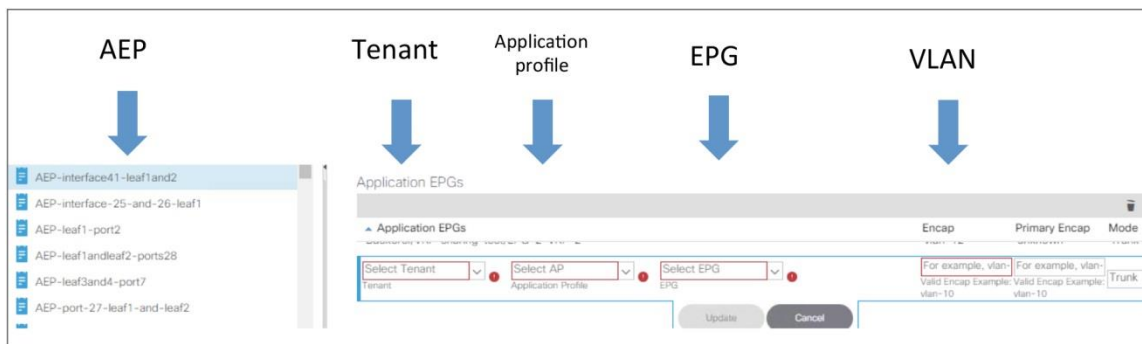


Figure 51.
Configuring EPGs from the AAEP

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 preboot 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 first-generation Cisco Nexus 9300 platform switches, traffic from the EPG in IEEE 802.1p mode will exit the port tagged as VLAN 0.
- If using Cisco Nexus 9300-EX or newer switches, traffic from the EPG in IEEE 802.1p mode will exit the port untagged.

In summary, 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).

If you are using a Cisco Nexus 9300-EX or newer 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.

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 52 illustrates these points.

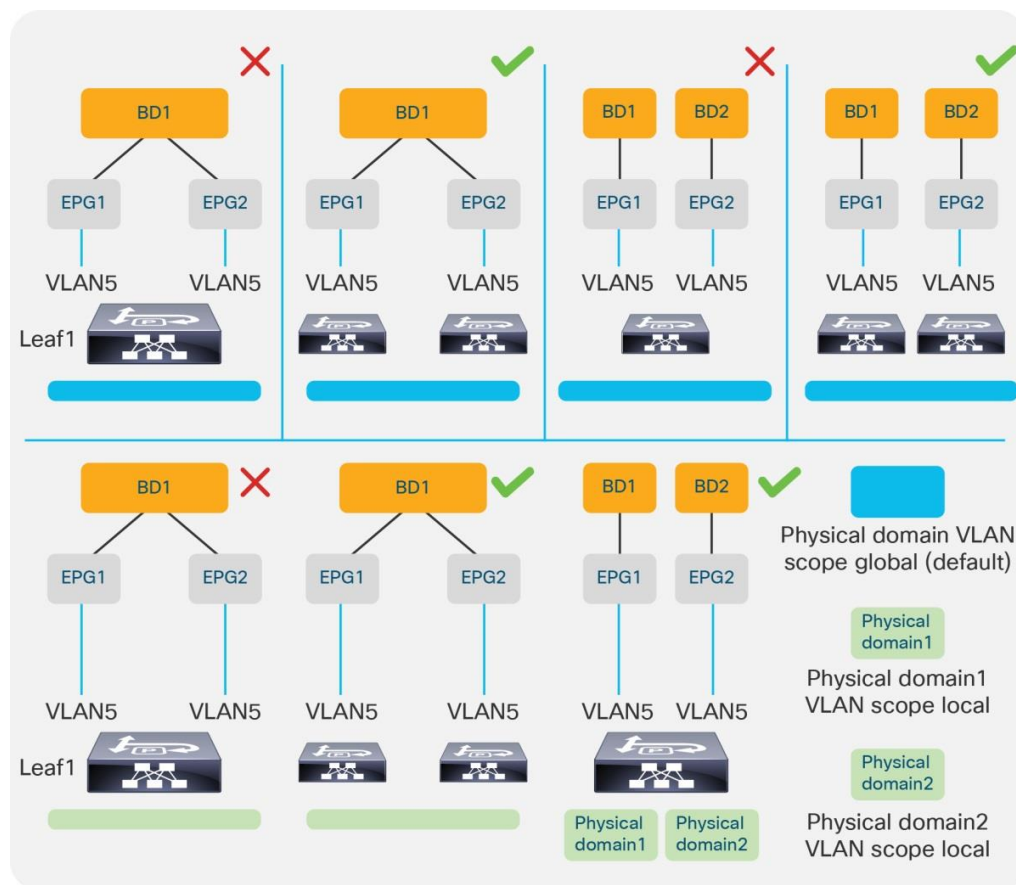


Figure 52.
Bridge domains, EPGs, and physical domains

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 Topology Change Notification (TCN) events to only clearing the endpoints learned on the switched network.

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 53 as an example.

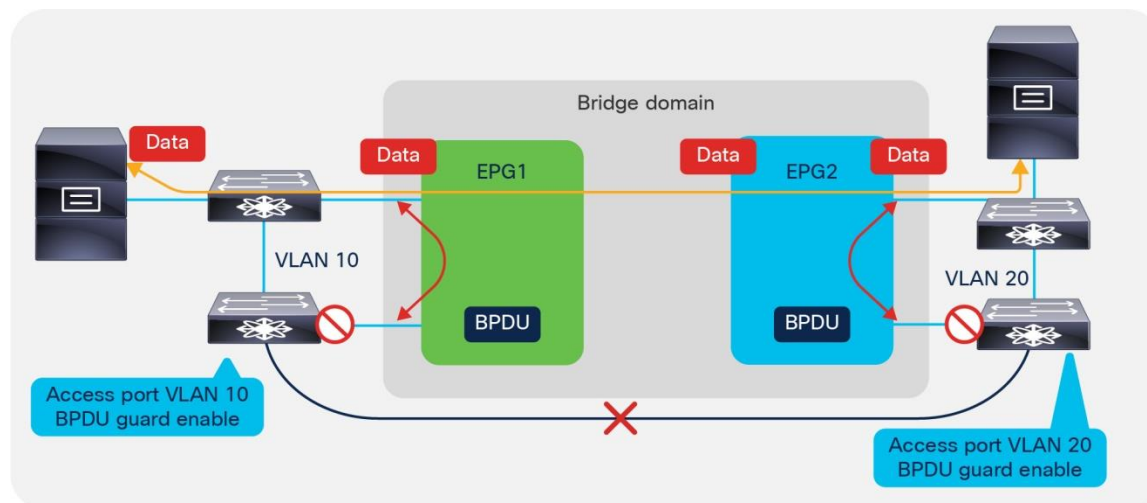


Figure 53.

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 BPDUs within the EPG (within the same VLAN ID). When the Cisco ACI leaf receives the BPDUs on EPG 1, it floods them to all leaf ports in EPG 1, and it does not send the BPDUs 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 BPDUs 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, BPDUs 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) BPDUs flood properly. BPDUs 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 BPDUs need to be flooded based on the VLAN tag in the frame.

However, for MST (IEEE 802.1s), BPDUs do not carry a VLAN tag, and the BPDUs 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 BPDUs are flooded to the desired ports, the user must create an EPG (an MST EPG) for VLAN 1 as native VLAN to carry the BPDUs. 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 to which it connects via the MST EPG, and flushes the local endpoint information associated with these VLANs on these leafs; as result, these entries are removed from the spine-proxy mapping database.

Internal VLANs on the leafs: EPGs and bridge domains scale

By mapping external VLANs to EPGs and bridge domains, the Cisco ACI fabric globally allows you to scale up to ~15,000 EPGs:

https://www.cisco.com/c/en/us/support/cloud-systems-management/application-policy-infrastructure-controller-apic/tsd-products-support-series-home.html#Verified_Scalability_Guides

While the Cisco ACI fabric offers an aggregate capacity of ~15,000 EPGs, on a per-leaf basis you need to keep into account the fact that VLAN tags are used locally to divide the traffic in different EPGs and different bridge domains. The total number of VLANs used on the switch depends on the number of EPGs and bridge domains; the total count must be under 3960. You can monitor the utilization of these hardware resources from the Operations > Capacity Dashboard > Leaf Capacity.

At the time of this writing, individual leafs this scale number also applies when using AVS or AVE with VXLANs, because leafs internally have hardware tables that use VLAN numbers (locally) to keep EPG traffic separate, to map EPGs to bridge domains, and to maintain information about bridge domains.

Cisco ACI uses VXLAN forwarding so the scale of bridge domains fabric-wide is bigger than 3960 (today's supported number is ~15,000 fabric-wide), and VLANs have local significance; hence, the same VLAN number can be reused on other leafs and can be mapped to the same or to a different bridge domain.

In light of this, at the time of this writing you should connect hosts using an aggregate of EPGs and bridge domains higher than 3960 to multiple leafs.

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 or with intra-EPG contracts). Figure 54 shows the relationship between EPGs and contracts.



Figure 54.
EPGs and contracts

An EPG provides or consumes a contract (or provides and consumes a contract). For instance, the App EPG in the example in Figure 54 provides a contract that the App Web consumes, and consumes a contract that the DB EPG provides.

Defining which side is the provider and which one is the consumer of a given contract allows establishing a direction of the contract where to apply ACL filtering; for instance, if the EPG Web is a consumer of the contract provided by the EPG App, you may want to define a filter that allows HTTP port 80 as a destination in the consumer-to-provider direction and as a source in the provider-to-consumer direction.

If, instead, you had defined the Web EPG as the provider and the App EPG as the consumer of the contract, you would define the same filters in the opposite direction; that is, you would allow HTTP port 80 as the destination in the provider-to-consumer direction and as the source in the consumer-to-provider direction.

In normal designs, you do not need to define more than one contract between any EPG pair. If there is a need to add more filtering rules to the same EPG pair, this can be achieved by adding more subjects to the same contract.

Security contracts are ACLs without IP addresses

You can think of security contracts as ACLs between EPGs. As Figure 55 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 the contracts.

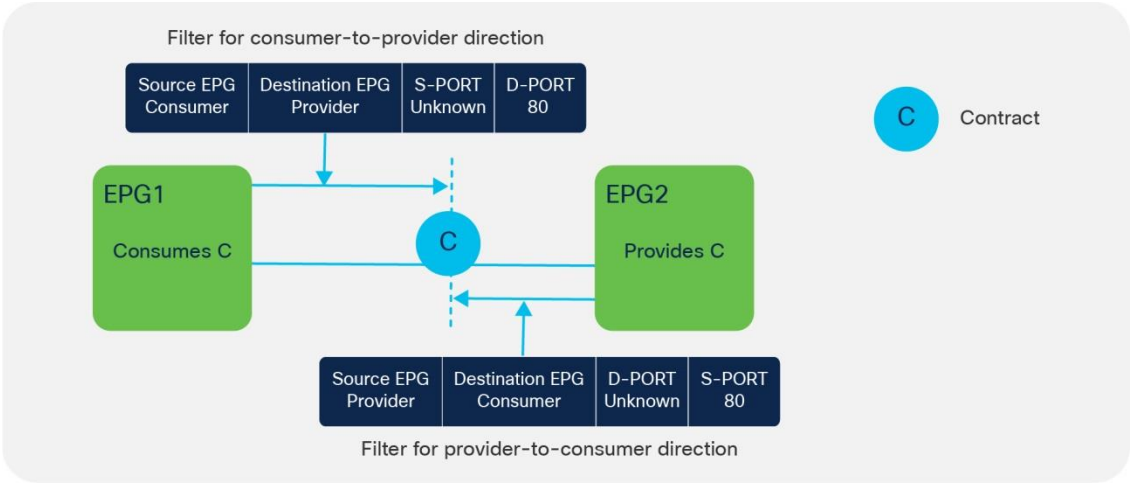


Figure 55.
Contracts are ACLs

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 56 shows how filters and filter entries are configured in the APIC GUI.

CREATE FILTER

Specify the Filter Identity

Name: web-filter

Description: optional

Entries: 1

Name	EtherType	ARP Flag	IP Protocol	Allow Fragment	Source Port / Range	Destination Port / Range	TCP Session Rules
web	IP		tcp	False	unspecified	unspecified	http

Figure 56.
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.

Permit, deny, redirect

The action associated with each filter is either permit or deny.

The subject can also be associated with a service graph configured for PBR.

These options give the flexibility to define contracts where traffic can be permitted, dropped, or redirected.

Please refer to the section “Contracts and filtering rule priorities” to understand which rule wins in case of multiple matching rules.

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, contracts 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.
- Contracts do not include IP addresses because traffic is filtered based on EPGs (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 57).



Figure 57.

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 58).



Figure 58.

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 59).

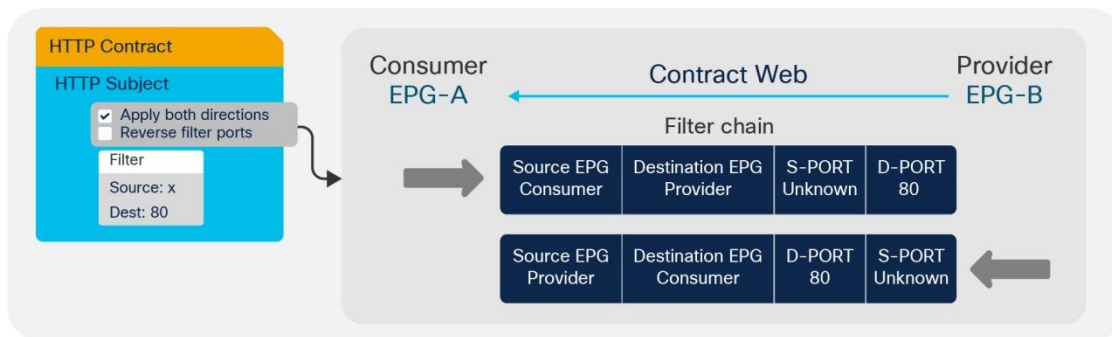


Figure 59.

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 60).

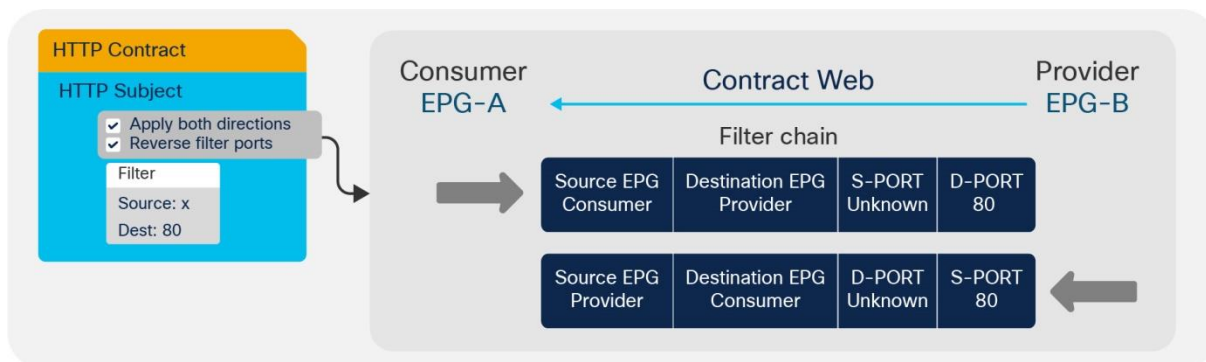


Figure 60.

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 61.

The screenshot displays two overlapping configuration windows in the Cisco ACI GUI. The background window is titled 'Create Contract' and shows the 'Specify Identity Of Contract' section with fields for Name (consumer-to-provider), Scope (VRF), QoS Class (Unspecified), Target DSCP (unspecified), and Description (optional). Below these is a table for 'Subjects' with columns for Name and Description. The foreground window is titled 'Create Contract Subject' and shows the 'Specify Identity Of Subject' section with fields for Name, Description (optional), and Target DSCP (unspecified). It also has checkboxes for 'Apply Both Directions' and 'Reverse Filter Ports', both of which are unchecked. Below this, there are two sections for filter chains: 'Filter Chain For Consumer to Provider' and 'Filter Chain For Provider to Consumer'. Each section contains a table with a 'Filters' header and a 'Name' column, with expand/collapse and add buttons.

Figure 61.
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.

As you can see from this example, more than one contract between any two EPGs is not generally required. If you need to add filtering rules between EPGs, you can simply add more subjects to the contract, and you can choose whether the subject is bidirectional or unidirectional.

If you configure bidirectional subject Cisco ACI programs automatically, the reverse filter port rule and with Cisco Nexus 9300-EX or newer, this can be optimized to consume only one policy CAM entry by using compression.

If you configure unidirectional subject rules, you can define filter ports for the consumer-to-provider direction and the provider-to-consumer direction independently.

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. For instance, instead of configuring multiple times the same filter in every tenant, you can define the filter once in the common tenant and use it from all the other tenants.

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:

- One reason is that the name used for contracts in the common tenant should be unique across all tenants. If a tenant is using a contract called for instance "fromAtoB" from the common tenant (common/fromAtoB), and you define a new contract with the same name inside of the tenant itself (mytenant/fromAtoB), Cisco ACI will change the EPG relations that were previously associated with common/fromAtoB to be associated to the locally defined contract mytenant/fromAtoB.
- Another reason is that, if multiple tenants provide and consume the same contract from the common tenant, you are effectively allowing communication across the EPGs of different 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 to allow the EPG-web of tenant A to talk to the EPG-app of tenant A. Imagine that you also want to allow the EPG-web of tenant B to talk to EPG-app of tenant B. If you configure EPG-app in both tenants to provide the contract web-to-app and you configure EPG-web of both tenants to consume the contract you are also enabling 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 app 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.

Unenforced VRF instances, Preferred Groups, vzAny

As already described in the section titled "Implementing network-centric designs / ACL filtering," in certain deployments, all EPGs associated with a VRF instance may need to be able to communicate freely. In this case, you could configure the VRF instance with which they are associated as "unenforced." This approach works, but then it will be more difficult, later on, to add contracts.

You can also use a VRF instance as "enforced," and use a feature called Preferred Groups. In this case 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.

Another approach consists in configuring a contract that permits all traffic that is applied to all the EPGs in the same VRF, via vzAny.

Using vzAny

vzAny is a special object that represents all EPGs associated with a given VRF instance, including the Layer 3 external EPG. This configuration object can be found in the Cisco ACI GUI in Networking > VRFs > VRF-name > EPG Collection for VRF.

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. Note that vzAny can only be a consumer of shared services, not a provider.

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

http://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/kb/b_KB_Use_vzAny_to_AutomaticallyApplyCommunicationRules_toEPGs.html

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.

A common use of vzAny is to allow all traffic between all EPG pairs and then to add more specific EPG-to-EPG contracts that allow only a set of ports and where all the other traffic between a specific EPG pair is dropped. Having a “deny” for the EPG-to-EPG traffic that does not match the allowed list configuration can be implemented by configuring, for example, a filter entry for “IP any” traffic with action “deny” and priority “lowest.”

Contracts and filtering rule priorities

When using contracts that include a combination of EPG-to-EPG contracts, with EPGs that may be part of Preferred Groups, or vzAny contracts, it is necessary to understand the relative priority of the filtering rules that are programmed in the policy CAM in order to understand the filtering behavior.

The relative priority of the rules that are programmed in the policy CAM are as follows:

- Filtering rules for contracts between specific EPGs have priority 7.
- Filtering rules for contracts defined for vzAny have priority 17 if configured with a filter with an EtherType such as IP and Protocol and source and destination ports that can be any.
- Preferred Group entries that disallow nonpreferred-group EPGs to any, have priorities 18 and 19.
- The implicit permit for Preferred Group members is implemented as any-to-any permit, with priority 20.
- vzAny configured to provide and consume a contract with a filter such as common/default (also referred to as an any-any-default-permit) is programmed with priority 21.
- The implicit deny has priority 21.

Rules with a lower priority number win over rules with a higher numerical value.

Specific EPG-to-EPG contracts have priority 7, hence they win over contracts defined, for instance, with vzAny because it is considered less specific.

Among filtering rules with the same priority, the following applies:

- Within the same priority, deny wins over permit and redirect.
- Between redirect and permit, the more specific filter rule (in terms of protocol and port) wins over the less specific.
- Between redirect and permit, if the filter rules have the same priority, redirect wins.

When entering a filter with a deny action, you can specify the priority of the filter rule:

- **Default value:** the same as the priority would be, in case there is permit for the same EPG pair
- **Lowest priority:** corresponding to vzAny-to-vzAny rules (priority 17)
- **Medium priority:** corresponding to vzAny-to-EPG rules (priority 13)
- **Highest priority:** same priority as EPG-to-EPG rules (priority 7)

Policy CAM compression

Depending on the leaf hardware, Cisco ACI offers many optimizations to either allocate more policy CAM space or to reduce the policy CAM consumption:

- Cisco ACI leafs can be configured for policy-CAM-intensive profiles.
- Range operations use one entry only in TCAM.
- Bidirectional subjects take one entry.
- Filters can be reused with an indirection feature (at the cost of granularity of statistics).

The compression feature can be divided into two main optimizations:

- Ability to look up the same filter entry from each direction of the traffic, hence making bidirectional contracts use half of the entries in the policy CAM. This optimization is available on Cisco Nexus 9300-EX or newer
- Ability to reuse the same filter across multiple EPG pairs/contracts. This optimization is available on Cisco Nexus 9300-FX or newer.

The two features are enabled as a result of choosing the “Enable Policy Compression” option in the filter configuration.

Note: Policy CAM compression cannot be enabled/disabled on contracts that are already programmed in the hardware. If you need to enable/disable policy compression, you should create a new contract and use it to replace the pre-existing one.

The ability to reuse the same filter is a policy CAM indirection feature where a portion of the TCAM (first-stage TCAM) is used to program the EPG pairs and the link to the entry in the second-stage TCAM that is programmed with the filter entries. If more than one EPG pairs requires the same filter, it can be programmed in the first-stage TCAM and point to the same filter entry in the second-stage TCAM.

With Cisco Nexus 9300-FX and Cisco Nexus 9300-FX2 hardware, when you can enable “compression,” this enables both the bidirectional optimization and, if the scale profile you chose allows it, policy CAM indirection.

Whether a leaf does policy CAM indirection depends on the profile you chose:

- Cisco Nexus 9300-FX can do policy CAM indirection with the default profile, IPv4 scale profile, and High Dual Stack profile.
- Cisco Nexus 9300-FX2 can do policy CAM indirection with the default profile and IPv4 scale profile, but not with the High Dual Stack profile.

You can find more information about policy CAM compression at this link:

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/4-x/basic-configuration/Cisco-APIC-Basic-Configuration-Guide-401/Cisco-APIC-Basic-Configuration-Guide-401_chapter_0110.html#id_76471

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 data center 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 domain 1, 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) 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 domain 1, 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 port group and EPG that is associated with this physical NIC and leaf. If a virtual machine vNIC is associated with an EPG whose physical NIC 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, above) regardless of whether the virtual machine on the virtualized host has sent traffic.
- **On-Demand:** The policy CAM is programmed as soon as first dataplane packet reaches the switch.

Table 5 illustrates the configuration options.

Table 5. Resolution and Deployment Immediacy

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

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 leaf2. An EPG, EPG1, is associated with BD1, which in turn is associated with VRF1. A virtual machine, VM1, is attached to EPG1 through a port group. 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 contract 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 leaf2.

Note: The On-Demand option is compatible with vMotion migration of Virtual Machines and requires coordination between APIC and the VMM. If all the APICs in a cluster are down, vMotion movement of a virtual machine from one virtual host connected to one leaf node to another virtual host connected to a different leaf node may 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. The 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.

According to the Cisco ACI Fundamentals document

(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_01011.html):

“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.

Quality of Service (QoS)

In releases of Cisco ACI up to and including 3.2, there are three user-configurable classes: Level1, Level2, and Level3. Starting with Cisco ACI Release 4.0, there are six user-configurable classes.

The queuing configuration for each Level is configured from Fabric > Access policies > Global policies > QoS class policies > Level <x>.

The classification of the traffic to the QoS group or level is based either on the DSCP or dot1p values of the traffic received from the leaf front panel ports (**Custom QoS** policy under the EPG), or on the contract between EPGs (**QoS Class** under the contract), or on the source EPG (**QoS Class** under the EPG).

If in the Custom QoS configuration there is a match of both the DSCP and CoS values, the classification based on the DSCP value takes precedence.

If the EPG does not have a specific QoS policy configured, the traffic is assigned to the Level 3 class (the default **QoS Class**).

If dot1p preserve is configured, the incoming traffic is assigned to the QoS group or level based on the EPG configuration, but the original CoS is maintained across the fabric.

If dot1p preserve is configured and Custom QoS is configured without a target CoS value, the original CoS is preserved, if instead the configuration specifies a target CoS, then the CoS is rewritten to the target CoS.

You can remark the traffic DSCP values by configuring the target CoS or the target DSCP values either as part of the Custom QoS configuration under the EPG or as part of the contract configuration.

Figure 62 illustrates the various QoS options.

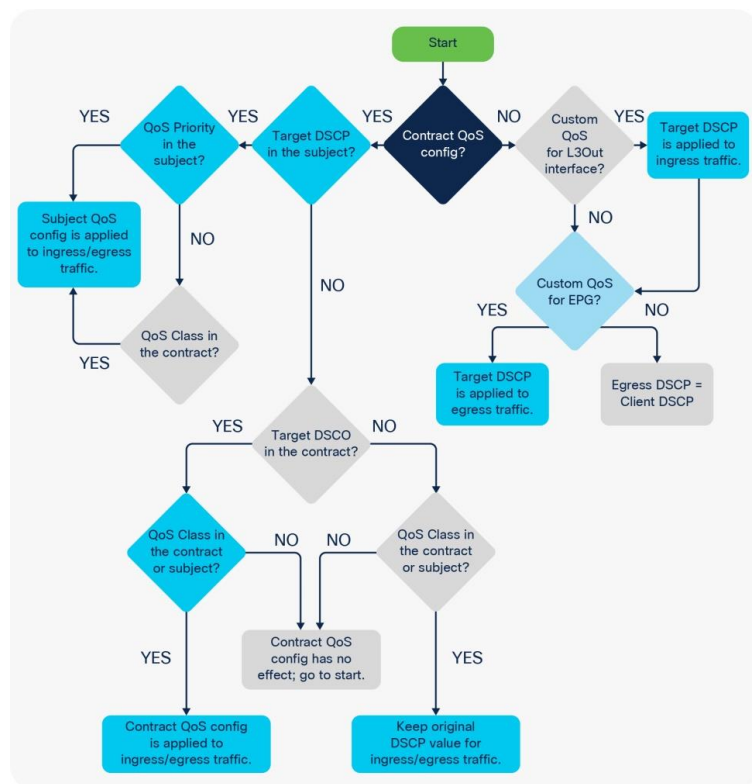


Figure 62.

QoS configuration options in Cisco ACI

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 an L3Out connection.

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 into BGP and, as a result, installed on the leaf nodes.

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). Using MP-BGP EVPN has the advantage of not requiring separate L3Out policies for each individual tenant and VRF.

Layer 3 outside (L3Out) 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.

This section describes the building blocks and the main configuration options of the L3Out. For more details, you can refer to the Layer 3 Configuration guide:

- For Cisco ACI Releases 3.x and earlier:
https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/2-x/L3_config/b_Cisco_APIC_Layer_3_Configuration_Guide.html
- For Cisco ACI Release 4.0:
<https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/4-x/L3-configuration/Cisco-APIC-Layer-3-Networking-Configuration-Guide-401.html>

A 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 63, 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, but in case of multiple L3Outs, you should use more specific subnets in the external network definition (please refer to the section titled “External network configuration options” for more information).

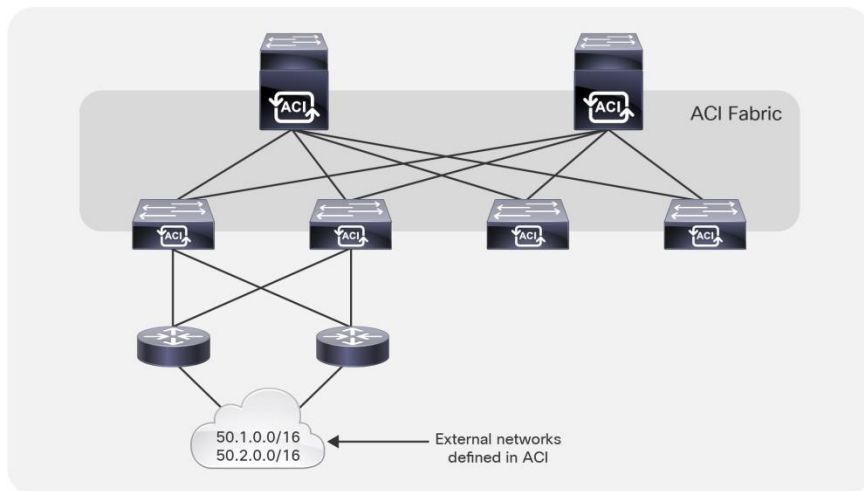


Figure 63.
External network

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 64. The other checkboxes are relevant for transit and shared-services scenarios and are described later in this section.

Subnet - 50.1.0.0/16

↺ ↻
⚠ ⚠ ⚠ ⚠

Properties

IP Address: 50.1.0.0/16
address/mask

Scope:

- ☐ Export Route Control Subnet
- ☐ Import Route Control Subnet
- ☒ External Subnets for the External EPG
- ☐ Shared Route Control Subnet
- ☐ Shared Security Import Subnet

Aggregate:

- ☐ Aggregate Export
- ☐ Aggregate Import
- ☐ Aggregate Shared Routes

Figure 64.
Defining traffic filtering for outside traffic

L3Out simplified object model

Figure 65 shows the object model for L3Out. This helps in understanding the main building blocks of the L3Out model.

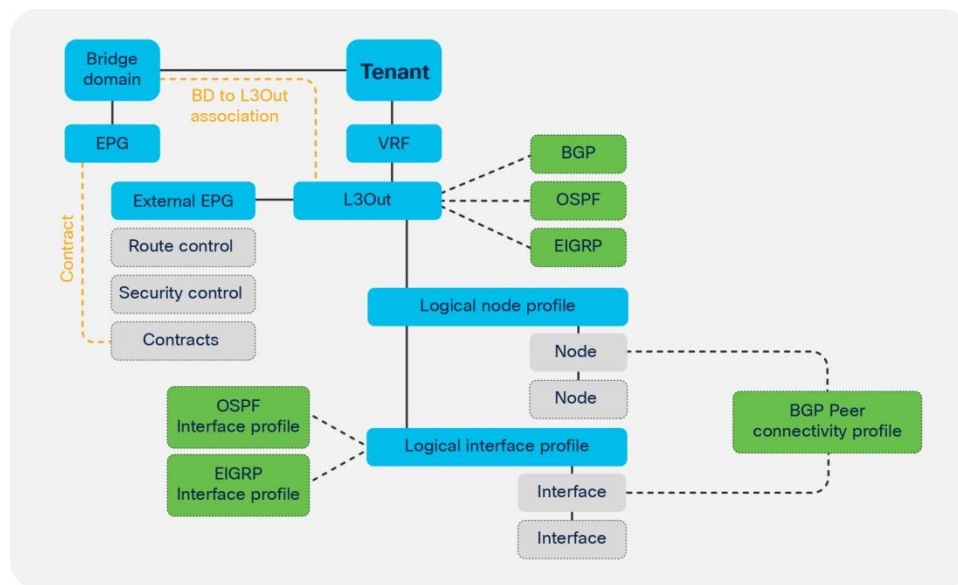


Figure 65.
Object model for L3Out

The L3Out policy is associated with a VRF and consists of the following:

- Logical node profile: This is the leafwide 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 is 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.

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 from the router ID.
- Each leaf switch should use a unique router ID per VRF. 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. Cisco ACI raises a fault if different router IDs are configured for L3Out connections on the same node for the same VRF.
- 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.

Route announcement options for the Layer 3 Outside (L3Out)

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 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) or a route-map must be configured matching the bridge domain subnet.
- 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 an 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 66 shows the bridge domain and subnet configuration with the relationships to an L3Out.

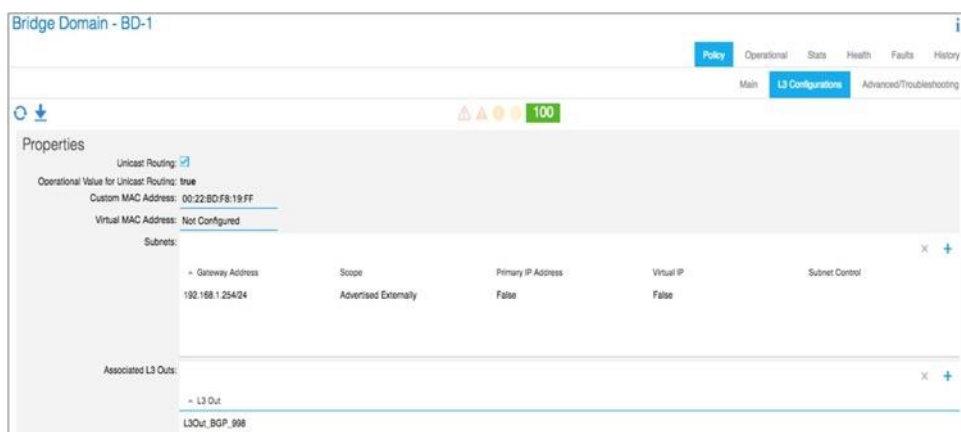


Figure 66.
Bridge domain relationships to L3Out connections

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 67).

These route maps apply to all routes:

- Directly connected subnets
- Static routes
- Transit routes

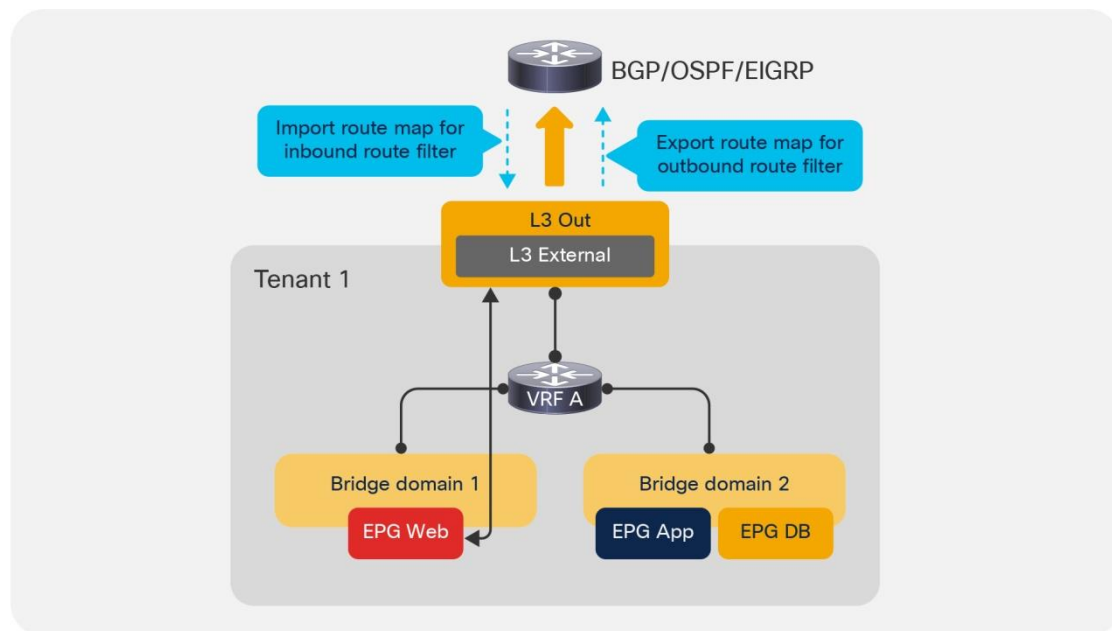


Figure 67.
L3Out configuration to control imported and exported routes

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 groups of external endpoints require different contract configurations.

Under the Layer 3 external EPG (also referred to as L3ext) configurations, the user can map external endpoints to this EPG by adding IP prefixes and network masks. The network prefix and mask do not 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.

The following is a summary of the options 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.

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 in the section Route Announcement Options for the Layer 3 Outside: a subnet under the bridge domain defined as Advertised Externally, a reference to the L3Out from the bridge domain or a route-map matching the bridge domain subnet, and a contract between the external EPG and internal EPGs.

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.

Host routes announcement

You can announce /32s of the Endpoints attached to the fabric (or to be more specific to the Cisco ACI pod) via a L3Out in two ways:

- Using GOLF
- Using the regular L3Out with Border Leafs (This option is available starting with Cisco ACI Release 4.0.)

With the capability of announcing host routes from the Border Leaf, you can use BGP, OSPF, EIGRP to announce /32 routes.

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. 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
- Subinterface with IEEE 802.1Q tagging: With this option, multiple subinterfaces 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 subinterfaces or SVIs, border leaf switches can provide L3Out connectivity for multiple tenants with one physical interface.

Attachment of endpoints to border leaf switches is 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, but additional tuning of the dataplane learning may be required.

Please refer to the section titled “Placement of outside connectivity / using border leafs for server attachment.”

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 68. 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.

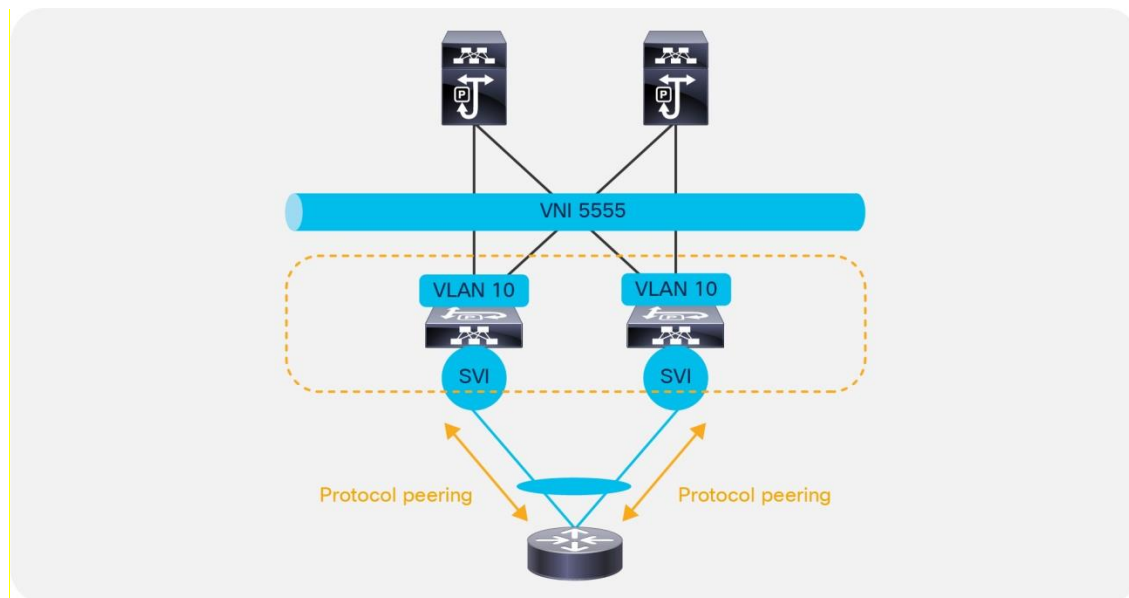


Figure 68.

Dynamic routing: peering over vPC

Additional design considerations are necessary when using a L3Out based on a vPC with more than two border leaf switches.

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 69.

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, subinterfaces, and SVIs, but it is used primarily with SVIs.
- Hot Standby Routing Protocol (HSRP): This option is available on routed interfaces and on subinterfaces (and not on SVIs). It is used primarily in conjunction with an external switching infrastructure that helps ensure Layer 2 connectivity between the subinterfaces.

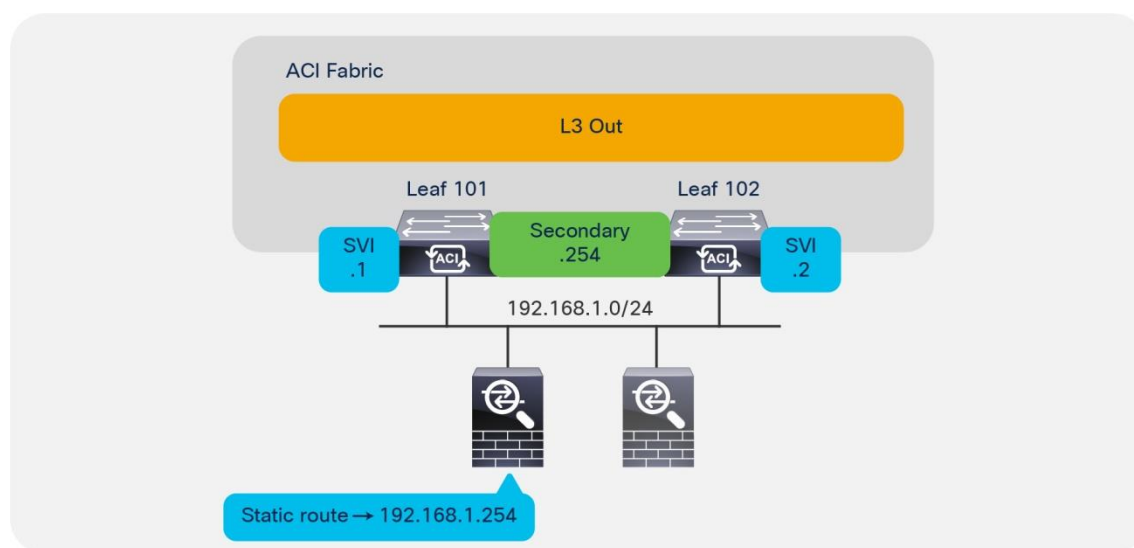


Figure 69.

L3Out secondary address configuration

In the example in Figure 69, a pair of Cisco ASA firewalls (running in active/standby mode) are attached to the Cisco ACI fabric. On the fabric side, the 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 70.

Figure 70.
SVI configuration

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.

Outside bridge domains

When configuring an SVI on an L3Out, you 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.

As already explained in the subsection titled “Do not use the L3Out to connect servers” the L3Out is meant to attach routing devices. It is not meant to attach servers directly on the SVI of an L3Out. Servers should be attached to EPGs and bridge domains.

There are multiple reasons for this:

- The Layer 2 domain created by an L3Out with SVIs is not equivalent to a regular bridge domain.
- The L3ext classification is designed for hosts that are multiple hops away.

Add L3out SVI subnets to the external EPG

When connecting devices to the L3out, such as L4–L7 devices, you should not just configure an L3ext of 0.0.0.0/0, but you should also add the L3Out SVI subnets. This is important if you have traffic destined to an IP address that is on the L3Out SVI; for instance, destined to the NAT address or the VIP of a firewall / load balancer. If you do not include the L3Out SVI, the IP addresses of the L4–L7 devices are assigned to class-id 1 instead of the L3ext class-id. In the specific case of traffic destined to an NAT or VIP address that belongs to the L3Out, if you did not add the L3Out SVI subnet to the L3ext, you may see that the traffic may be dropped even if a contract between the EPG and the L3ext is present.

Figure 71 illustrates this point.

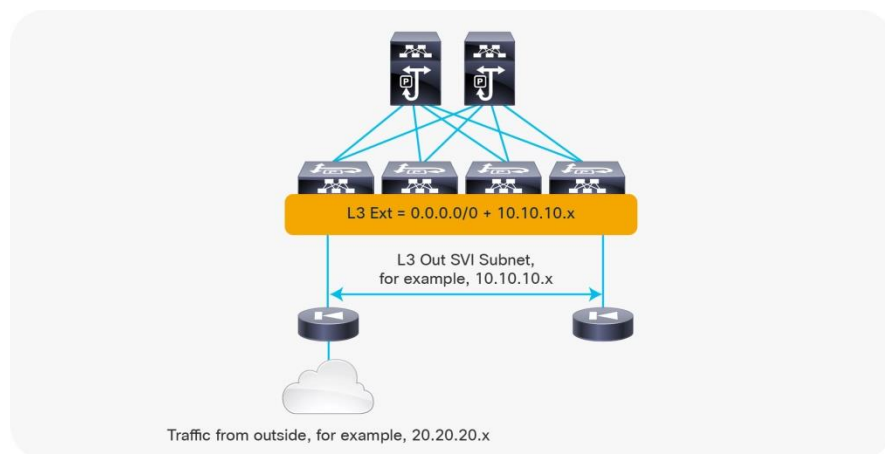


Figure 71.

Add the SVI subnet to the L3ext

Bidirectional Forwarding Detection (BFD) for L3Out

Cisco ACI Software Release 1.2(2g) added support for bidirectional forwarding detection (BFD) for L3Out links on border leafs. 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.

With Cisco ACI versions prior to Cisco ACI 3.1(1), BFD can be configured on L3Out interfaces only, where BGP, OSPF, EIGRP, or static routes are in use.

From Cisco ACI Release 3.1(1), BFD can also be configured between leaf and spine switches, and between spines and IPN links for GOLF, Multi-Pod, and Multi-Site connectivity (to be used in conjunction with OSPF or with static routes).

Note: BFD for spines is implemented for cloud-scale line cards:

<https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/datasheet-c78-736677.html>

Cisco ACI uses the following implementations of BFD:

- 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 BFD 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.

This global default policy can be overridden if required by creating a new nondefault 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.

Enabling BFD on L3Out SVIs helps ensure fast failure detection (assuming that the connected device supports it). For routed interfaces and subinterfaces, BFD may still be enabled, although physical interface mechanisms should ensure fast failure detection in most circumstances.

The verified scalability guide provides the BFD session scale that has been tested per leaf:

https://www.cisco.com/c/en/us/support/cloud-systems-management/application-policy-infrastructure-controller-apic/tsd-products-support-series-home.html#Verified_Scalability_Guides

Considerations for multiple L3Outs

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.

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.

External EPGs have a VRF scope

Even if Layer 3 external EPGs are under the L3out, when the VRF is configured for ingress filtering, Layer 3 external EPGs should be thought of as per-VRF classification criteria.

In the presence of multiple L3Outs and with VRF configured for ingress filtering (which is the default), the L3exts must be configured to be L3Out-specific by entering specific subnets instead of 0.0.0.0/0 (or by having, at the most, one 0.0.0.0/0 L3ext).

Note: If the VRF is configured with egress filtering instead, if the L3Outs are on different leaves, the L3ext of 0.0.0.0/0 would then be effectively referring to the specific L3Out where it is configured.

Consider the example shown in Figure 72.

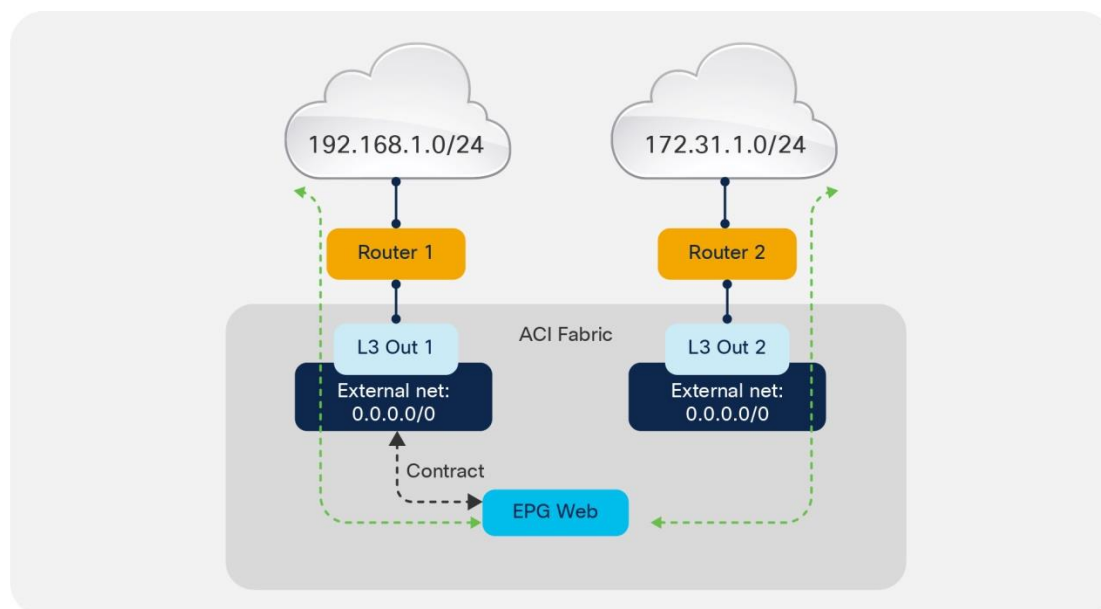


Figure 72.

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 a 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 73.

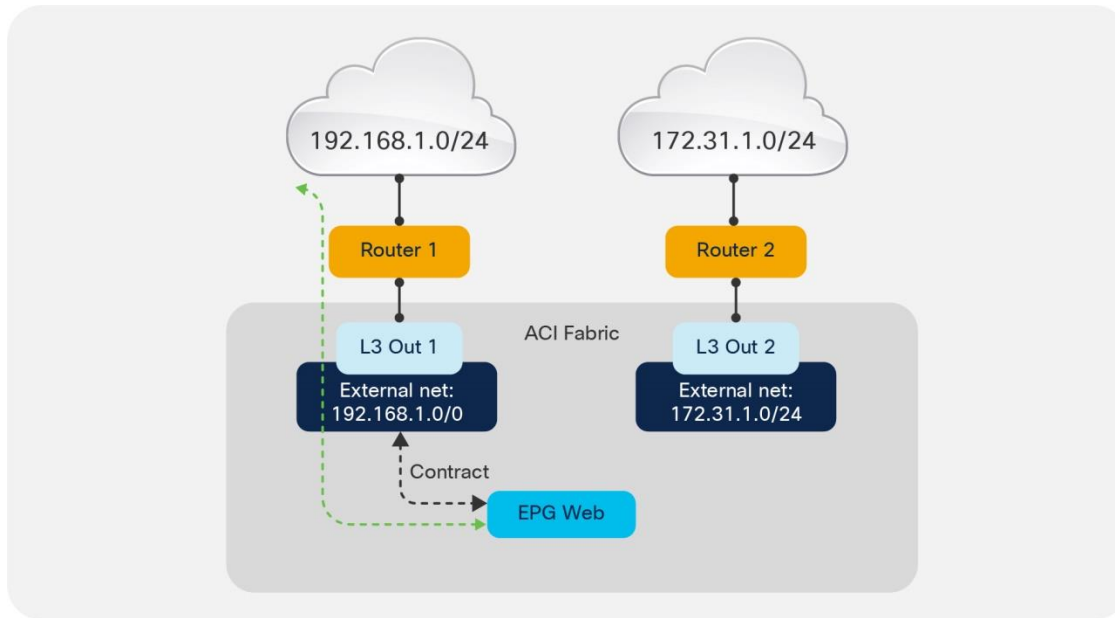


Figure 73.
Security enforcement with multiple EPGs: nonoverlapping subnet classification

Figure 74 should help in understanding how to use the L3ext.

The left of the figure shows how the L3ext is configured in Cisco ACI; it is under the L3Out. The right of the figure shows how you should think of the L3ext; that is, as a per-VRF configuration.

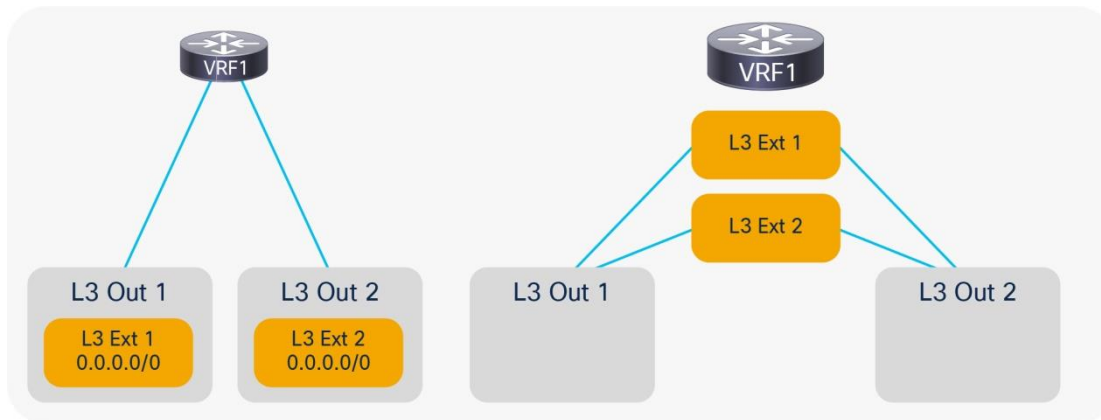


Figure 74.
The L3ext / external EPG classifies traffic for all L3Outs under the same VRF.

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 data center to the local L3Out and then bridged on the external bridge domain to the L3Out in another data center.

Figure 75 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 or newer 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.

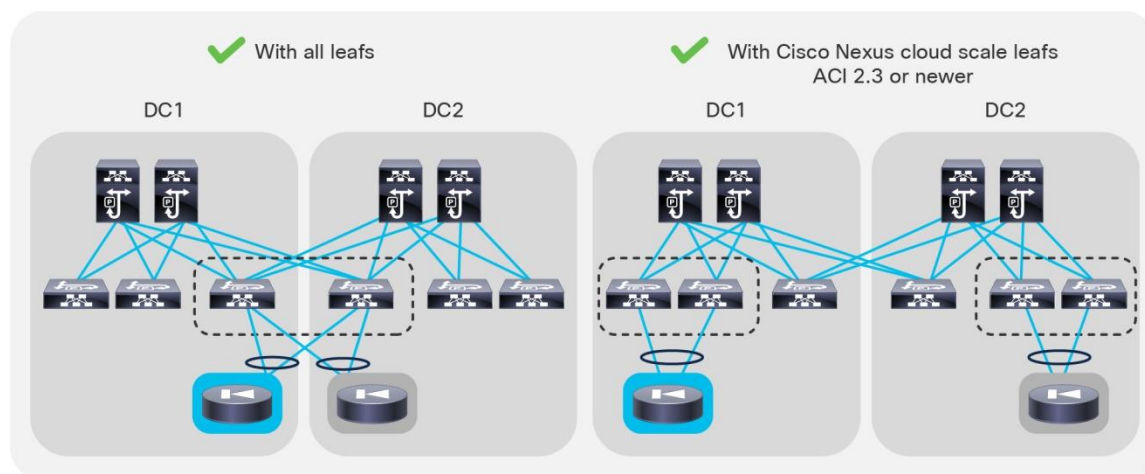


Figure 75.
Design considerations with static routing L3Out with SVI and vPC

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 76 illustrates this point.

Figure 76 shows four border leaf switches: two in each data center. There are two L3Outs or a single L3Out that uses different VLAN encapsulations for data center 1 (DC1) and data center 2 (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.

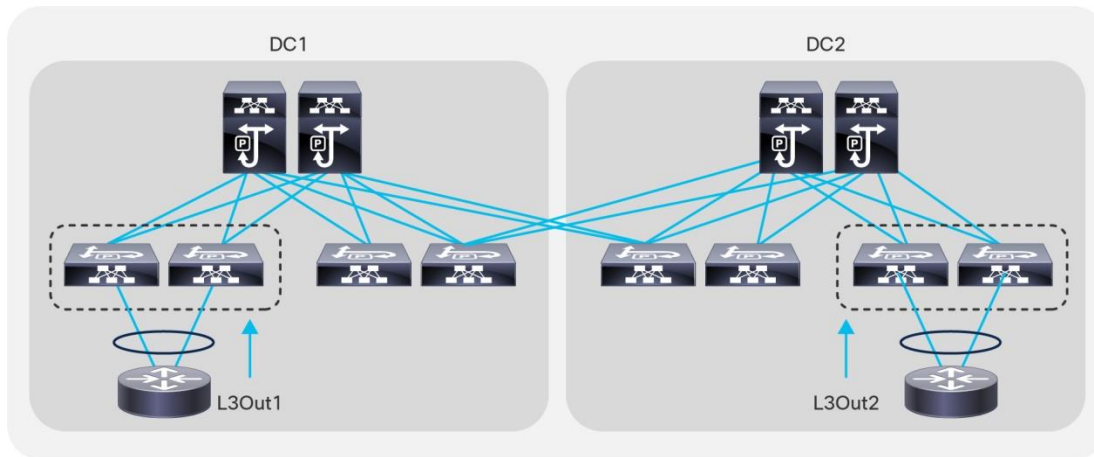


Figure 76.
Design considerations with dynamic routing L3Out with SVI and vPC

Using BGP for external connectivity



BGP Autonomous System (AS) number

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 “BGP Route Reflector Policy” section.

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 77.

Peer Connectivity Profile - BGP Peer Connectivity Profile 10.10.10.2




Properties

Address: 10.10.10.2
 Description: optional

BGP Controls:

☐ Allow Self AS
☐ Disable Peer AS Check
☐ Next-hop Self
☐ Send Community
☐ Send Extended Community

CHECK ALLUNCHECK ALL

Password:
 Confirm Password:

Allowed Self AS Count: 3

Peer Controls:

☐ Bidirectional Forwarding Detection
☐ Disable Connected Check

EBGp Multihop TTL: 1

Weight for routes from this neighbor: 0

Private AS Control:

☒ Remove all private AS
☒ Remove private AS
☒ Replace private AS with local AS

BGP Peer Prefix Policy: select a value

Remote Autonomous System Number: 100

Local-AS Number Config: no options

Local-AS Number: 3333
This value must not match the MP-BGP RR policy

Figure 77.
L3Out BGP configuration

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.

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).

Route summarization

Support for route summarization was introduced in Cisco ACI 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 78, Figure 79, and Figure 80.

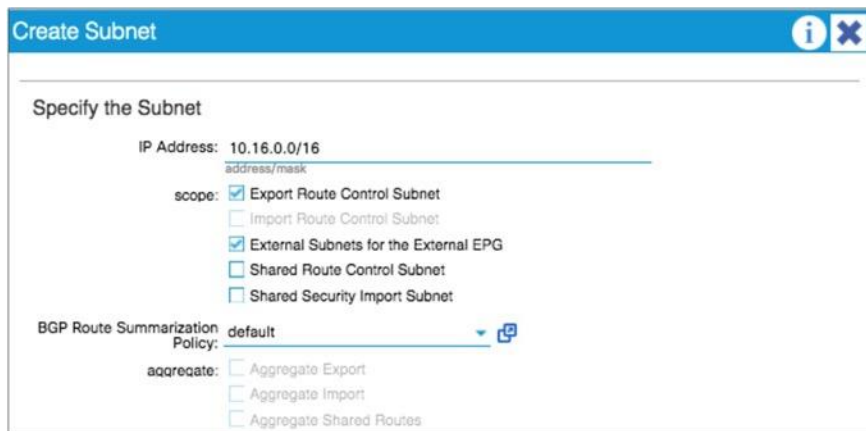


Figure 78.
BGP Route Summarization configuration

Create Subnet

Specify the Subnet

IP Address: 10.16.0.0/16
address/mask

scope:
☒ Export Route Control Subnet
☐ Import Route Control Subnet
☒ External Subnets for the External EPG
☐ Shared Route Control Subnet
☐ Shared Security Import Subnet

OSPF Route Summarization Policy: default

aggregate:
☐ Aggregate Export
☐ Aggregate Import
☐ Aggregate Shared Routes

Figure 79.
OSPF Route Summarization configuration

Create Subnet

Specify the Subnet

IP Address: 10.16.0.0/16
address/mask

scope:
☒ Export Route Control Subnet
☐ Import Route Control Subnet
☒ External Subnets for the External EPG
☐ Shared Route Control Subnet
☐ Shared Security Import Subnet

EIGRP Route Summarization: ☒

Figure 80.
EIGRP Route Summarization configuration

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 81 shows the configuration of the AS-SET option under the BGP summarization policy.

Create BGP Route Summarization Policy

Define BGP Route Summarization Policy

Name: BGP-Summarize

Description: optional

Control State: ☒ Generate AS-SET information

Figure 81.
BGP AS-SET configuration

OSPF route summarization

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 Cisco NX-OS).

When tenant routes or transit routes are injected into OSPF, the Cisco ACI leaf node where the L3Out resides acts as an OSPF Autonomous System Boundary Router (ASBR). In this case, the **summary-address** configuration (that is, the external route summarization) should be used. This concept is shown in Figure 82.

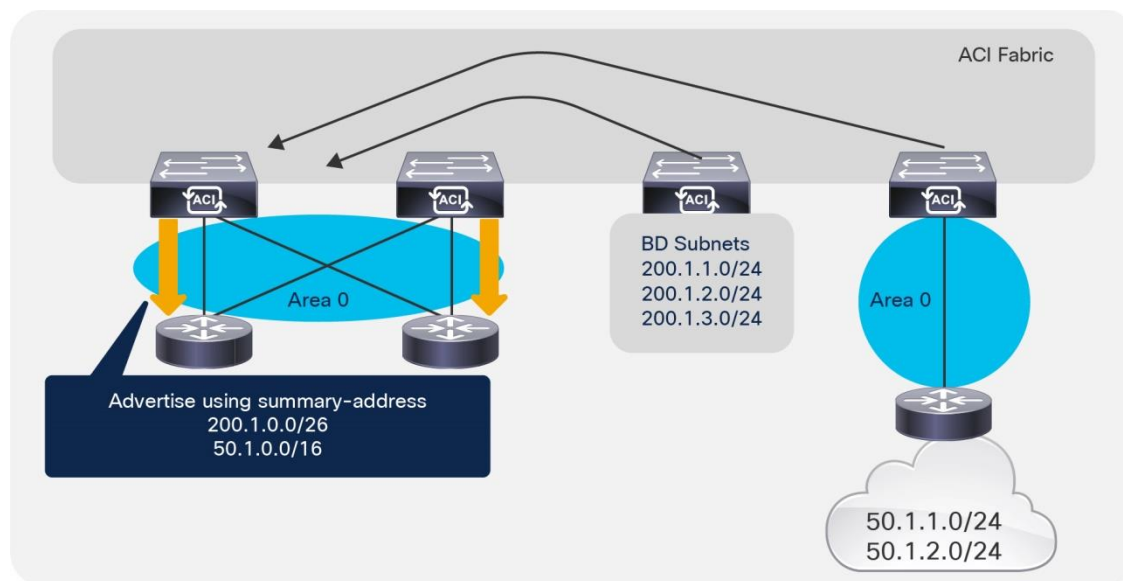


Figure 82.
OSPF summary-address operation

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 83.

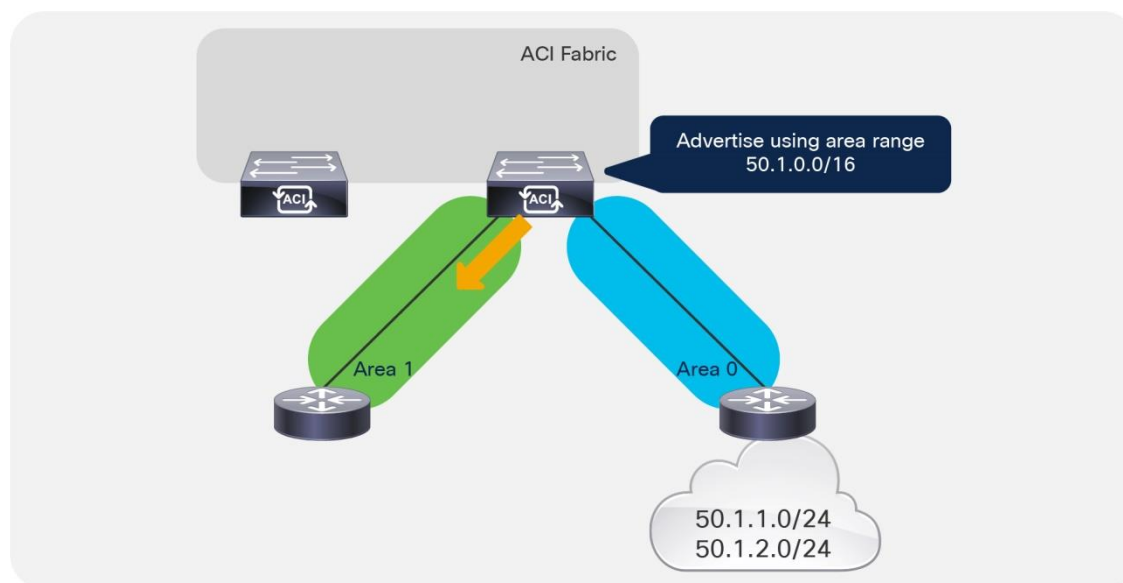


Figure 83.
OSPF area range operation

The OSPF route summarization policy is used to determine whether the summarization will use the area range or the summary-address configuration, as shown in Figure 84.



Create OSPF Route Summarization Policy

Define OSPF Route Summarization Policy

Name:

Description:

Inter-Area Enabled: ☒

Cost:

Figure 84.
OSPF Route Summarization

In the example in Figure 84, 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.

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 an 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 85.

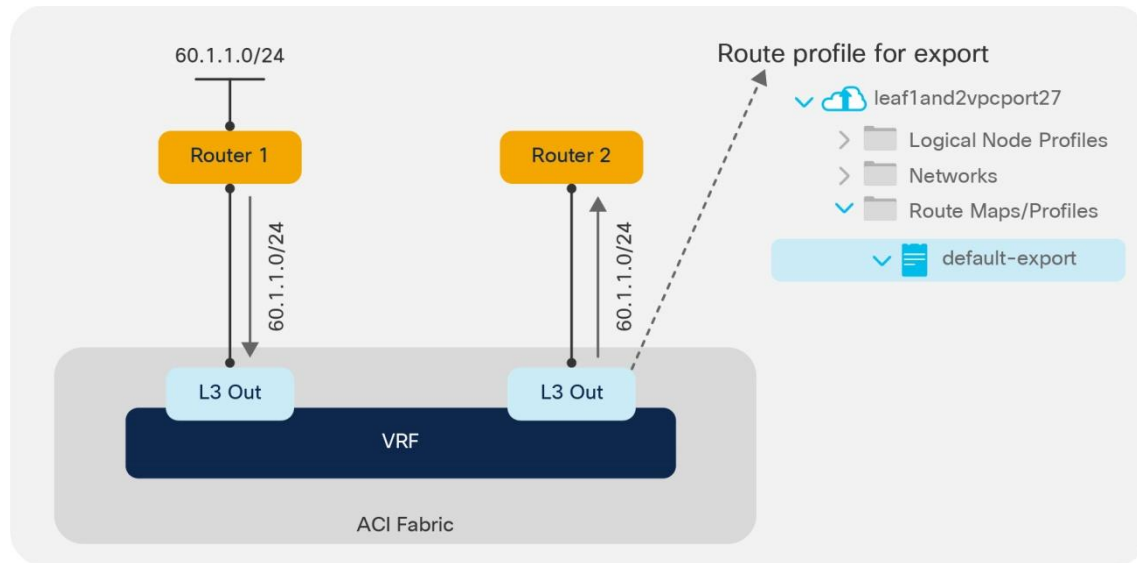


Figure 85.
Export route control operation

In the example in Figure 85, 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 86.

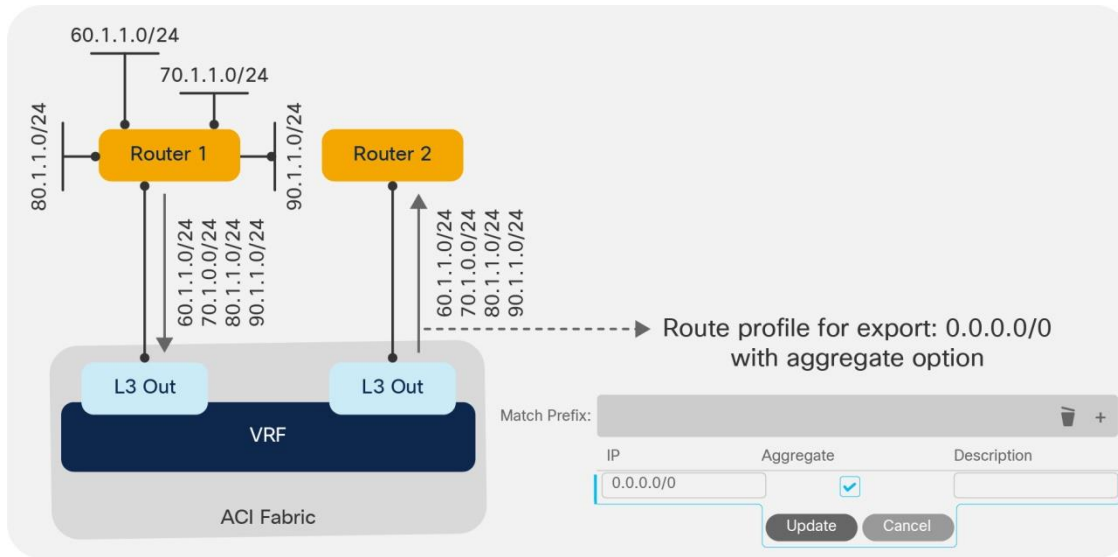


Figure 86.
Aggregate export option

In the example in Figure 86, 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: 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 87.

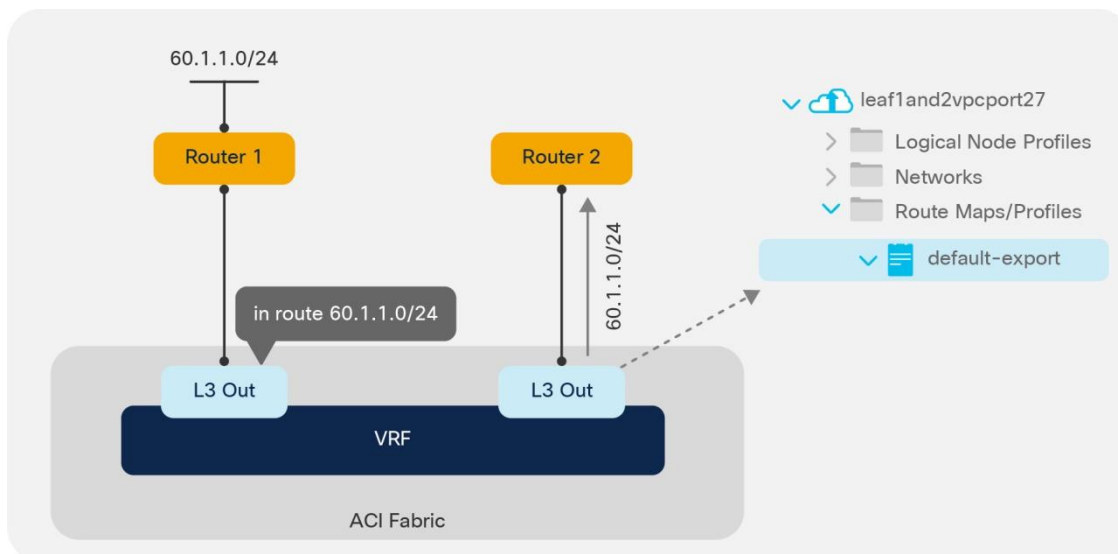


Figure 87.
Exporting static routes

In the example in Figure 87, there is a static route to 60.1.1.0 configured on the left L3Out. If you need to advertise the static route through the right L3Out, you must specify a route profile to allow it.

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:

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/aci/apic/sw/kb/b_KB_Transit_Routing.html.

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 88.

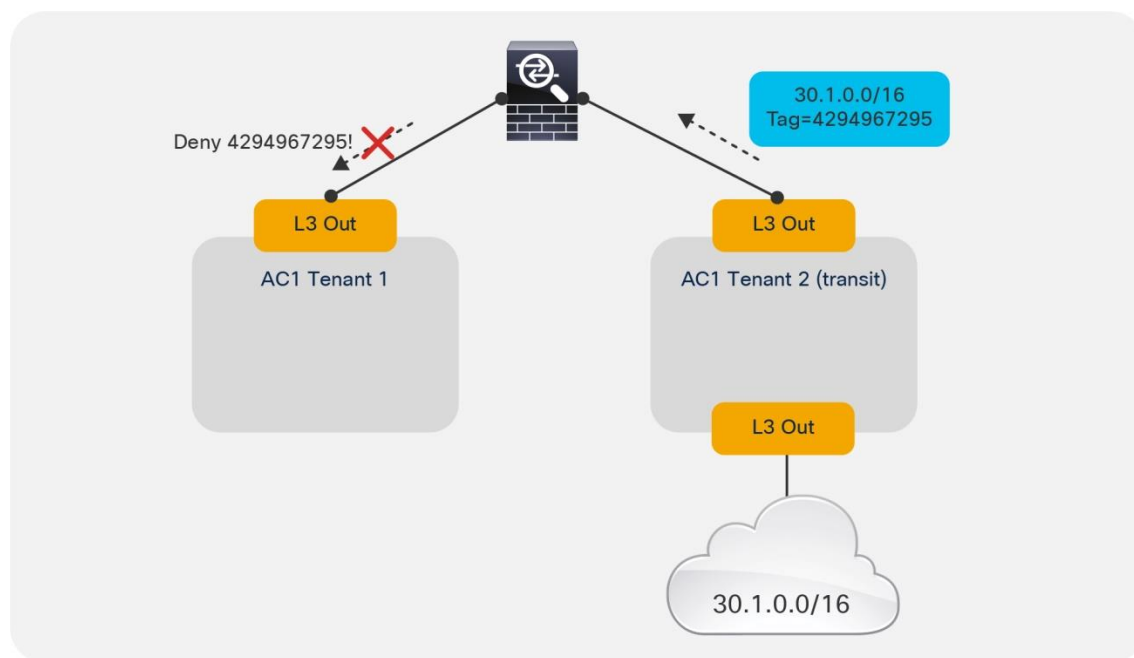


Figure 88.

Loop prevention with transit routing

In the example in Figure 88, an external route (30.1.0.0/16) is advertised in Cisco ACI Tenant 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 ACI Tenant 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 89.

Create Route Tag Policy

Define Route Tag Policy

Name: New-Route-Tag

Description: optional

Tag: 4294967222

Figure 89.
Changing route tags

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 (GOLF) 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: If you migrate an existing network to Cisco ACI, you can configure tenants with a network-centric approach, which makes migration easier. If you plan to migrate later with more segmentation, you should consider instead reducing the number of bridge domains by merging more subnets into the same bridge domain. When merging bridge domains, you should consider Flood in Encapsulation to limit the scope of flooding.
- 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. At the time of this writing, most features are designed to work with VRF configured for ingress filtering.

- **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 Cisco ACI dataplane learning works 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, and a subnet configured in the bridge domain.
- **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 optimization 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.
- **Dataplane learning:** In the presence of active/active NIC teaming (other than vPC) and in the presence of floating IP addresses, you may need to tune the VRF or the bridge domain for dataplane learning.
- **EPGs and access ports:** When associating EPGs with bare-metal servers, use Access (untagged) as the access-port option with Cisco Nexus 9300-EX and Cisco Nexus 9300-FX (or newer) platform switches. With first-generation leafs, you should use 802.1p for access ports.
- **EPG and contracts:** For migration purposes, make sure you understand the options of VRF unenforced, Preferred Groups, and vzAny.
- **Contracts:** Make sure you understand the relative priorities of contracts between EPGs, or between vzAny and the rule priority of permit, deny, and redirect.
- **Deployment and Resolution Immediacy:** 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. Configure different vDSs for management connectivity and for virtual workload connectivity.
- **VRF sharing:** 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 L3Out, vPC, SVI encapsulation, and routing in order to define correctly the L3Out configuration.
- **Multiple L3Outs and L3External:** Make sure you understand how the L3external works with multiple L3Outs.

For more information

For more information, please refer to <https://www.cisco.com/go/aci>.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at <https://www.cisco.com/go/offices>.

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: <https://www.cisco.com/go/trademarks>. Third-party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1110R)