Connecting Application Centric Infrastructure (ACI) to Outside Layer 2 and 3 Networks

Guide

Version 1.0

Contents

Introduction

Cisco® Application Centric Infrastructure (ACI) delivers centralized, application-driven policy automation, management, and visibility of both physical and virtual environments as a single system. It is optimized to support an “application anywhere” model, with complete freedom of application movement and placement. This novel approach empowers IT teams to offer cloud-based services to their customers directly with the associated service-level agreements (SLAs) and performance requirements for the most demanding business applications.

This document describes how to integrate ACI fabric with existing network infrastructure in the data center. This document requires that readers have basic knowledge about ACI building components, ACI basic concepts, and policy model.

The Cisco ACI solution allows users to use layer 3 technology (standard IP protocol) to connect to outside networks. You can use ACI to:

- Connect to an existing switch network infrastructure and provide a layer 3 connection between workloads in the ACI fabric and workloads outside of the ACI fabric.
- Connect to WAN routers in the data center so that a WAN router provides Layer 3 data-center interconnect (DCI) or Internet access for tenants. In some scenarios, a WAN router provides VPN connection to a tenant’s on-premise network.

The Cisco ACI solution also provides options to allow users to use layer 2 technology to connect the ACI fabric to an existing L2 network. At the time of this writing ACI support VLAN connection to an outside network. In the future, a virtual extensible LAN (VXLAN) will be another alternative to connect to an outside network. With VXLAN technology, a customer can extend layer 2 domains to remote devices that are reachable through an IP cloud.

The layer 2 connection between an ACI fabric and an outside network is required in the following scenarios:

- In the existing Data Centers, connect the existing switching network to an ACI leaf, and stretch the same VLAN and subnet across ACI and the existing network. This allows workloads to be distributed across the existing switching infrastructure and ACI fabric. Customers also have the choice to migrate the workloads from the existing networks to the ACI fabric.
- Extending the layer 2 domain from ACI to a DCI platforms so that the layer 2 domain of ACI can be extended to a remote Data Centers.

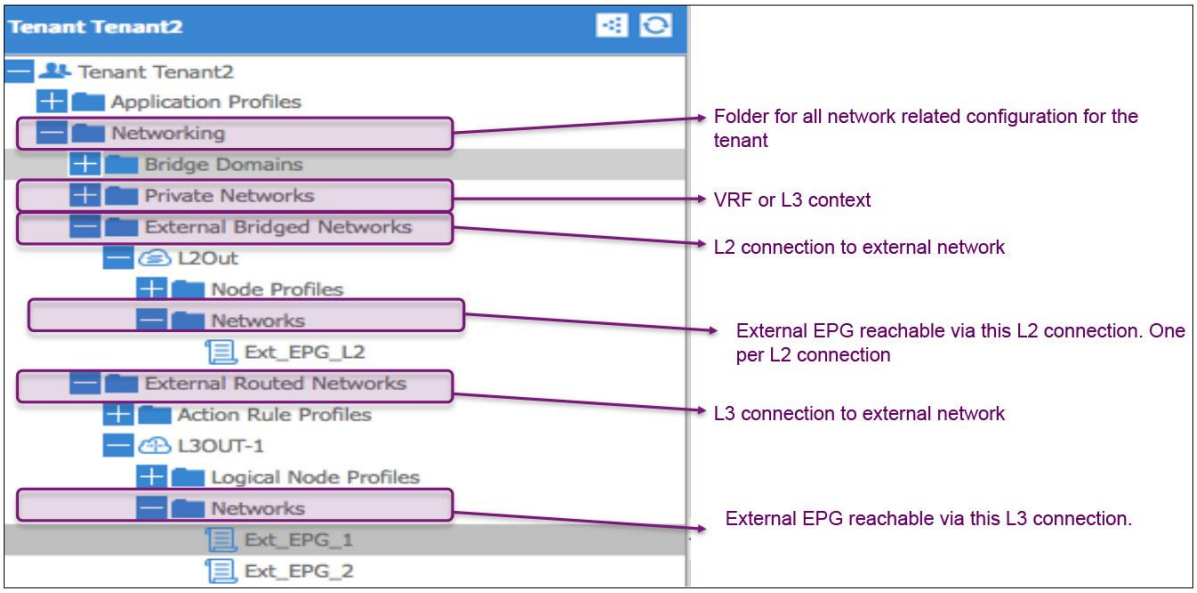
ACI Layer 3 Connection to an Outside Network

This section explains how ACI can connects to an outside networks using layer 3 technology. It explains the route exchange between ACI and the external routers, and supported dynamic routing protocols between the ACI border leaf and external routers. It also explores the forwarding behavior between internal and external endpoints, and how the policy is enforced for the traffic flow between them.

Before getting into the details of the implementation and supported features, these are some of the terms that are introduced in the ACI solution and on the APIC GUI.

Figure 1 depicts some terminologies and explanations in the context of ACI.

Figure 1. ACI Terminology and Explanations

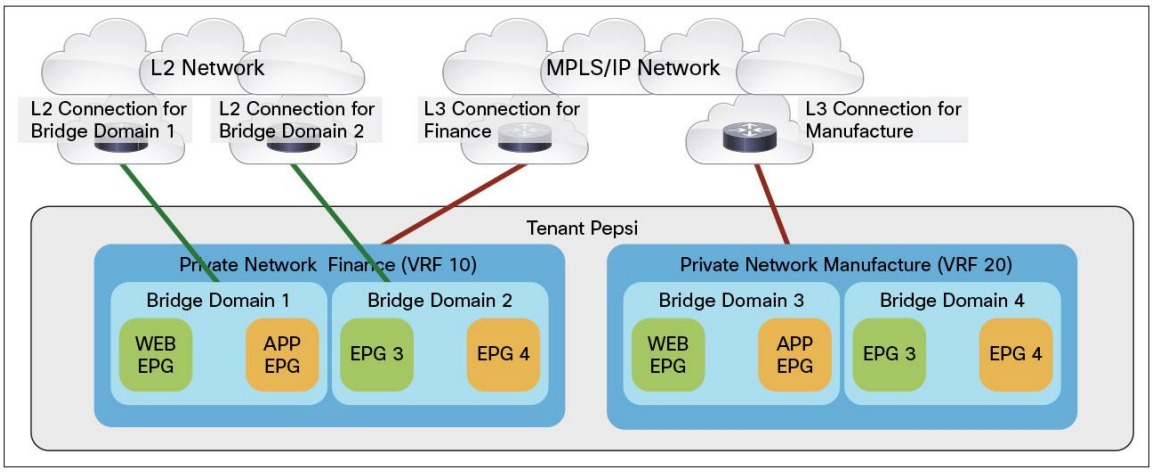


- **Networking** - The “Networking” tab under the tenant menu is the place where users can configure most network related objects for a tenant. It is a folder that contains the configuration for bridge domains, private networks, external bridged networks, and external routed networks.
- **Bridge domains** - The “Bridge Domain” tab allow a user to configure parameters for all bridge domains for this tenant, such as the subnet address, forwarding behavior for this bridge domain, whether ACI fabric provides routing for this bridge domain, whether the tenant allows a given subnet to be advertised to external routes, etc.
- **Private networks** - Private networks are also called layer 3 in some Cisco documents and you may see this in the object model document for ACI. A private network is essentially a Virtual Route Forwarding (VRF) that provides IP address space isolation for different tenants. Each tenant can have one or more private networks, or share one default private network with other tenants when there is no overlapping IP addressing being used in the ACI fabric.
- **External bridged networks** - An external bridged network is sometimes referred to as an layer 2 outside connection in this and other Cisco documents. It is one of the options to provide layer 2 extension from the ACI fabric to an outside network. This is one of topics covered in detail in this document. The “**Network**” menu under external bridged network is for users to configure the external endpoint group (EPG). Once external EPGs are defined, users can define the contract for the communication between the internal EPG (EPG defined under application profile) and the external EPGs.
- **External routed networks** - These are also called layer 3 outside connections in this and other ACI documents. This is where users configure the interfaces, protocols, and protocol parameters that are used to provide IP connectivity to external routers. The “**Network**” menu under the external routed network is for users to configure external EPGs. Once external EPGs are defined, users can define contracts for the communication between internal EPGs (EPG defined under application profile) and the external EPGs.

Figure 2 illustrates the relationship between layer 2 and layer 3 outside connections, and other components of ACI solution.

Layer 3 outside connections, or external routed networks, provide IP connectivity between a private network of a tenant and an external IP network. Each layer 3 outside connection is associated with one private network only. A private network may not have a layer 3 outside connection if IP connectivity to the outside is not required.

Figure 2. Relationship between Layer 2 Layer 3 Outside Connections and Other Components



Border Leaves

The border leaves are ACI leaves that provide layer 3 connections to outside networks. Any ACI leaf can be a border leaf. These can also simply be called leaf switches. There is no limitation in the number of leaf switches that can be used as border leaves. The border leaf can also be used to connect to compute, IP storage, and service appliances. In large-scale design scenarios it may be preferred to have border leaf switches separated from the leaves that connect to compute and service appliances for scalability reasons.

Three different types of interfaces are supported on a border leaf switches to connect to an external router:

- **Layer 3 interface**
- **Sub-interface with 802.1Q tagging -** With sub-interface, the same physical interface can be used to provide a layer 2 outside connection for multiple private networks
- **Switched Virtual Interface (SVI) -** With an SVI interface, the same physical interface that supports layer 2 and layer 3 and the same physical interface can be used for a layer 2 outside connection as well as a layer 3 outside connection

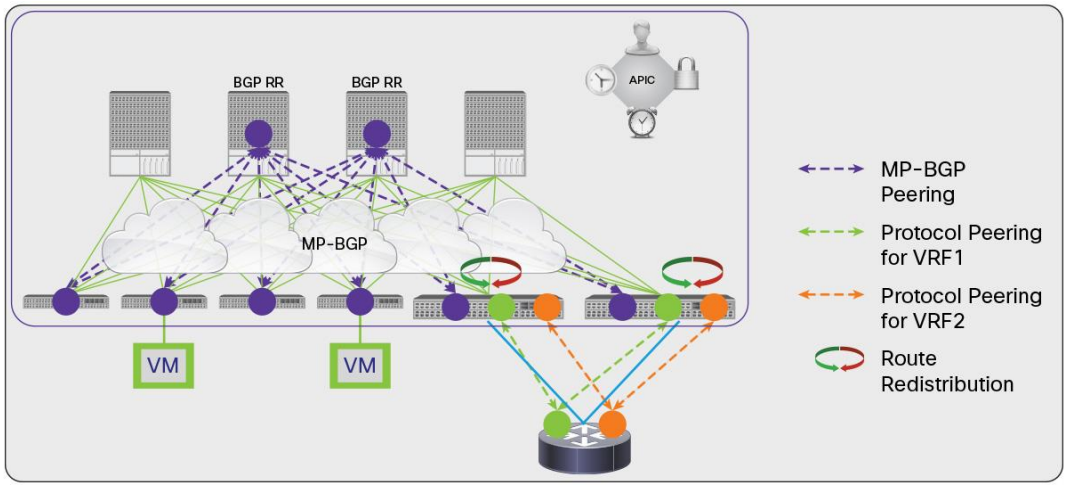
In addition to supporting routing protocols to exchange routes with external routers, the border leaf also applies and enforces policy for traffic between internal and external endpoints.

Route Distribution within the ACI Fabric

As of this writing, ACI supports the following routing mechanisms: static routes, OSPFv2, and IBGP protocol. ACI supports VRF-lite implementation when connecting to an external routers. By using sub-interfaces, border leaf can provide layer 3 outside connection for multiple tenants with one physical interface. VRF-lite does require one protocol session per tenant though. Future software versions of ACI will support BGP Ethernet EVPN (EVPN) between the border leaf and external router, and one BGP session can carry route updates for all tenants.

Within the ACI fabric, Multiprotocol BGP (MP-BGP) is implemented between leaf and spine switches to propagate external routes within the ACI fabric. The BGP route reflector technology is deployed in order to support a large number of leaf switches within a single fabric. All of the leaf and spine switches are in one single BGP autonomous system (AS). Once the border leaf learns the external routes, it can then redistribute the external routes of a given VRF to an MP-BGP address family VPN version 4 (or VPN version 6 when IPv6 routing is supported in ACI). With address family VPN version 4, MP-BGP maintains a separate BGP routing table for each VRF. 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 leaves where the VRFs (or private network in the APIC GUI's terminology) are instantiated. Figure 3 illustrates the routing protocol within the ACI fabric and the routing protocol between the border leaf and external router with VRF-lite.

Figure 3. Routing Protocols in ACI Fabric



Following is an output captured on a leaf that shows the two MP-BGP sessions with two spine nodes.

```
fab3-leaf1# show bgp sessions vrf overlay-1
Total peers 2, established peers 2
ASN 100
VRF overlay-1, local ASN 100
peers 2, established peers 2, local router-id 10.0.121.159
State: I-Idle, A-Active, O-Open, E-Established, C-Closing, S-Shutdown

Neighbor        ASN    Flaps LastUpDn|LastRead|LastWrit St Port(L/R)  Notif(S/R)
10.0.106.30       100 0     1d03h   |00:00:44|00:00:32 E  50933/179  0/0
10.0.121.158      100 0     1d03h   |00:00:44|00:00:32 E  37324/179  0/0
```

In this example, we have BGP autonomous system number (ASN) 100 for the fabric. The addresses, 10.0.106.30 and 10.0.121.158, are the two spine nodes addresses. With ACI, Cisco Application Policy Infrastructure Controllers (APICs) manage the infrastructure IP address space and automatically allocate proper IP addressing required for the leaf and spine. The infrastructure IP addresses are in a separate VRF than the tenants so that the infrastructure IP is contained within fabric and won't have an overlapping IP address issues. The infrastructure VRF is transparent to external traffic. Users can specify this IP address range during the APIC initial configuration.

The following output shows the leaf receiving external routes through MP-BGP from two spine nodes:

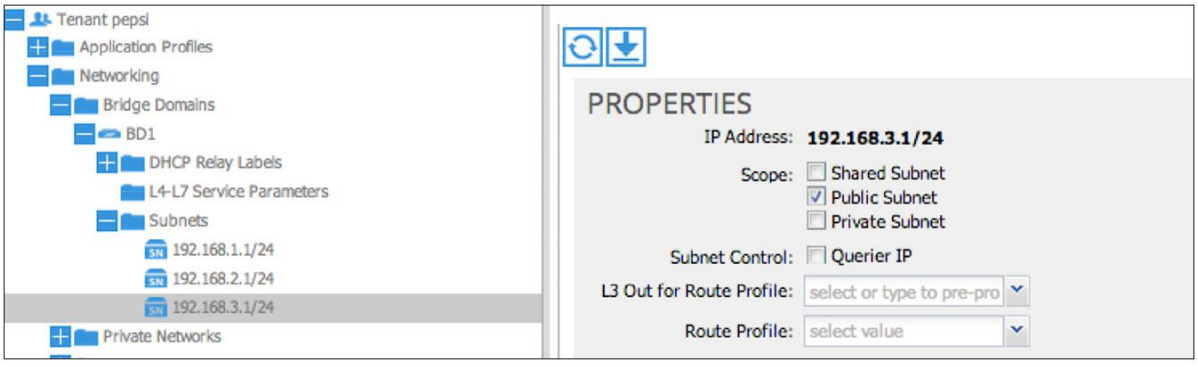
```
0.0.0.0/0, ubest/mbest: 2/0
    *via 10.0.117.190%overlay-1, [200/1], 23:43:44, bgp-100, internal, tag 100
    *via 10.0.117.191%overlay-1, [200/1], 22:24:00, bgp-100, internal, tag 100
7.7.7.7/32, ubest/mbest: 1/0
    *via 10.0.117.190%overlay-1, [200/4], 07:35:46, bgp-100, internal, tag 100
100.1.1.0/30, ubest/mbest: 1/0
    *via 10.0.117.190%overlay-1, [200/0], 23:43:59, bgp-100, internal, tag 100
100.1.1.4/30, ubest/mbest: 1/0
    *via 10.0.117.191%overlay-1, [200/0], 22:24:11, bgp-100, internal, tag 100
```

External routes received from two spine nodes via MP-BGP

Each layer 3 outside connection and its protocol sessions are associated with a given private network (essentially VRF within ACI). As a result, border leaf switches know which private network the external routes need to be redistributed to. The same rule applies to the reverse direction of route propagation (i.e., propagate ACI fabric tenant routes to external routers). In the ACI fabric, the private networks are dynamically instantiated where required. For compute leaf switches, private networks are created when end points are attached to the private network. On the border leaves, the private network is created when the layer 3 outside connection is configured for this private network.

Border leaves are the place where tenant subnets are injected into the protocol running between the border leaves and external routers. Users have control of which tenant subnets they want to advertise to the external routers. When specifying subnets under a bridge domain for a given tenant, the user has the choice to specify the scope of a subnet as indicated in Figure 4.

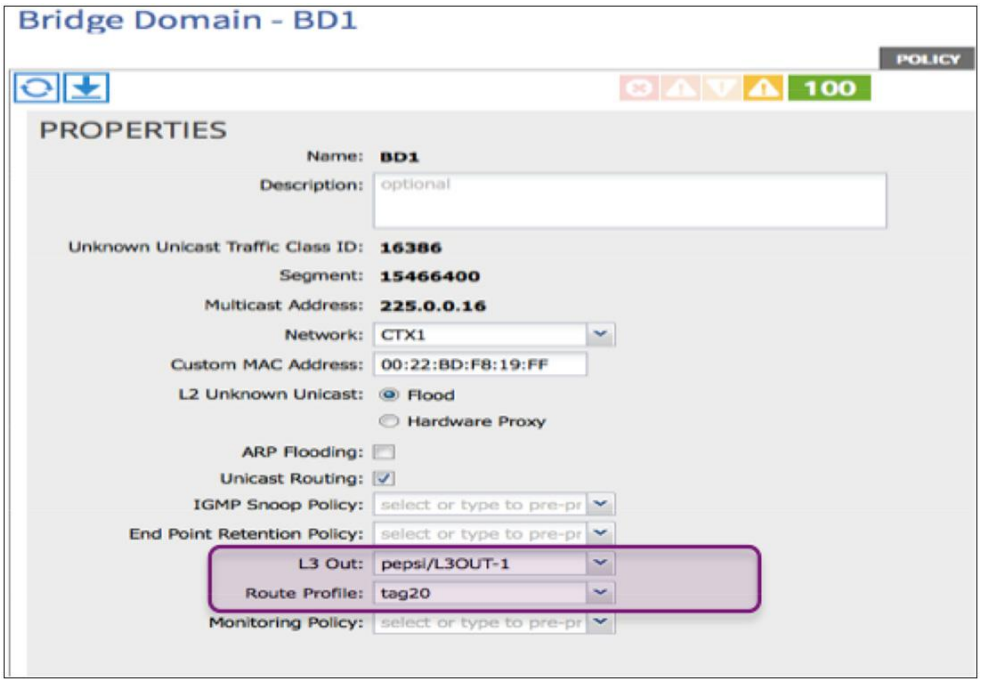
Figure 4. Subnet Configuration under Bridge Domain



- **Public subnet** indicates that this subnet will be advertised to the external router by the border leaf.
- **Private subnet** indicates that this subnet will be contained within the ACI fabric and will not be advertised to external routers by the border leaf.
- **Shared subnet** is for shared services. It is used to indicate that this subnet needs to be leaked to one or more private networks. The shared subnet attribute is applicable to both public and private subnets. Details of shared services are not covered in this document.

In addition to specifying a tenant subnet as public subnet, the user also needs to associate a layer 3 outside connection to a bridge domain in order for the border leaf to advertise the tenant subnet to an external router (Figure 5).

Figure 5. Associate a L3 Outside Connection with a Bridge Domain

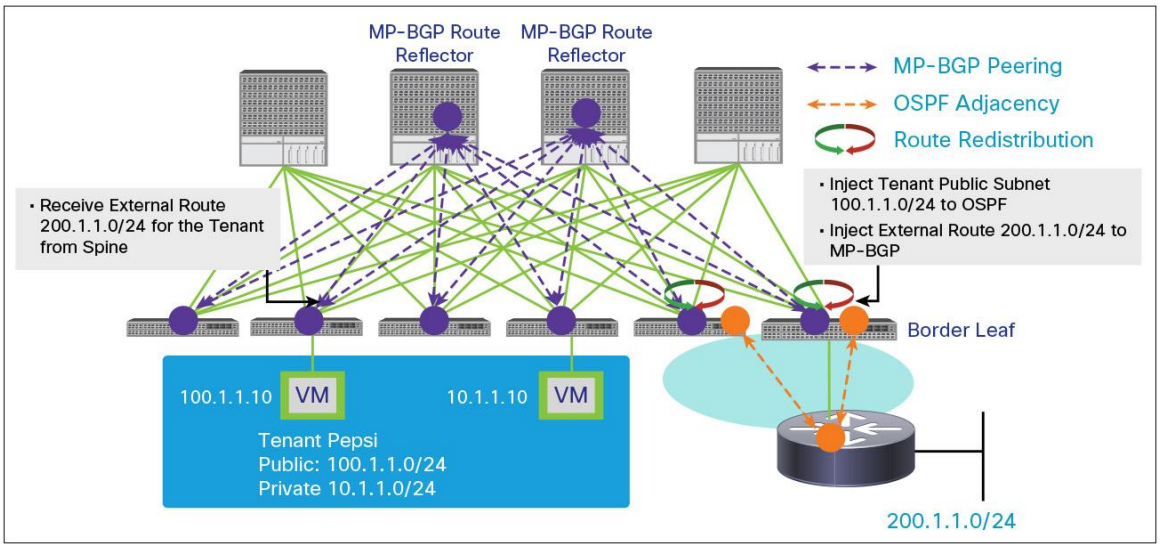


The following is a sample XML file that can be posted to the APIC in order to create a tenant (named “Tenant2”), the private network (“CTX1”), bridge domain (“bd1”), and its three subnets. The XML file will also associate the layer 3 outside connection, named “L3OUT-1” (not created by this XML post), and specify one subnet of the bridge domain to be a public subnet.

```
<fvTenant name='Tenant2'>
    <fvCtx name="CTX1"\>
    <!—Create bridge domain and enable routing-->
    <fvBD name="bd1" unicastRoute="yes">
         <!—Associate the bridge domain with L3 outside connection-->
        <fvRsBDToOut tnL3extOutName='L3OUT-1'/>
        <fvSubnet ip="1.1.1.1/16"/>
        <fvSubnet ip="1.2.1.1/16"/>
        <fvSubnet ip='40.1.1.1/24' scope='public'/>
        <fvRsCtx tnFvCtxName="CTX1"/>
    </fvBD>
</fvTenant>
```

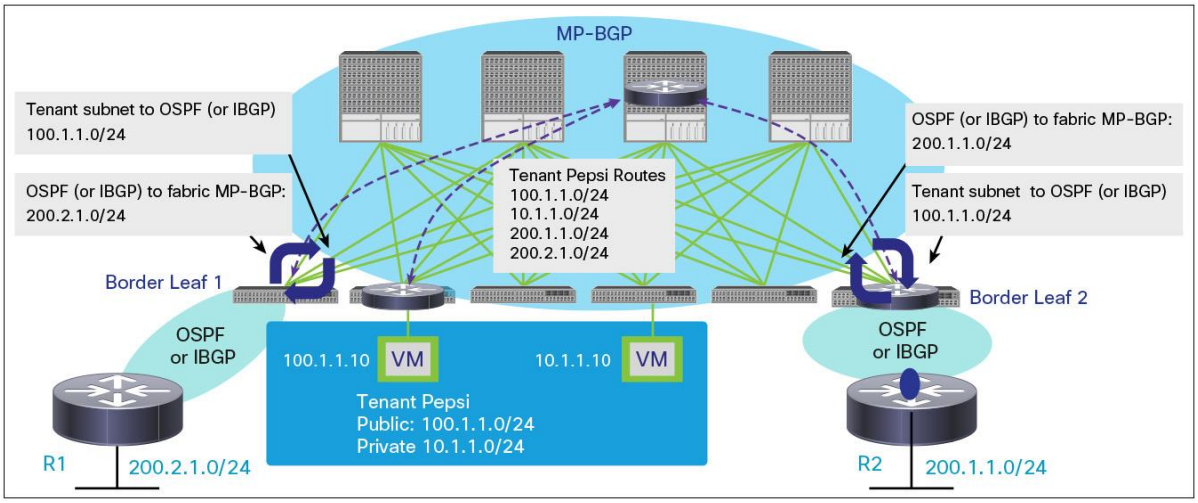
Figure 6 depicts how the external routes are propagated within the ACI fabric and how tenant subnets are announced to external routers.

Figure 6. Propagation of External Routes within ACI Fabric



With current software version, the ACI border leaf only advertises tenant subnets (the subnets under the bridge domain) to the external routers. It doesn't advertise transit routes (the routes learned from another external router) to external routers. In other words, external router 1 advertises routes to the ACI fabric, and the ACI fabric doesn't advertise these routes to external router 2. Figure 7 explains this behavior. Border leaf 1 learns external route 200.2.1.0/24 and injects it into the MP-BGP domain. Border leaf 2 learns this external route from BGP route reflector (RR) (the spine node). However, it doesn't inject the route 200.2.1.0 to the routing protocol (OSPF or iBGP) with external router R2. Both border leaves only redistribute the tenant public subnet (which is 100.1.1.0/24) to external router R1 and R2.

Figure 7. Route Exchange between ACI Border Leaf and External Routers



The current implementation of the fabric route distribution policy implies that the ACI fabric is not intended to be used as a transit network to carry traffic between two routers or two external IP networks. The enhancement to support ACI fabric being a transit network is planned for future software release. Please check the latest ACI software release note.

Note: ACI fabric is designed to be a stub network with current software version.

One important thing to note is that MP-BGP is not enabled by default in the ACI fabric. For the deployment scenario where ACI fabric is used as L2 fabric or there is no need for L3 outside connection, MP-BGP is not required. To enable MP-BGP please configure BGP policy on the APIC to specify the BGP ASN and specify spine nodes as BGP route reflectors. Once these two are configured, the APIC will take care of the rest, such as configuring IBGP peering between the leaf and spine, and specifying leaves as route reflector clients. It also automatically generates the required configuration for route redistribution on the border leaf.

Note: MP-BGP is not enabled by default within the ACI fabric. To enable MP-BGP, assign an AS number and specify spine nodes as RR.

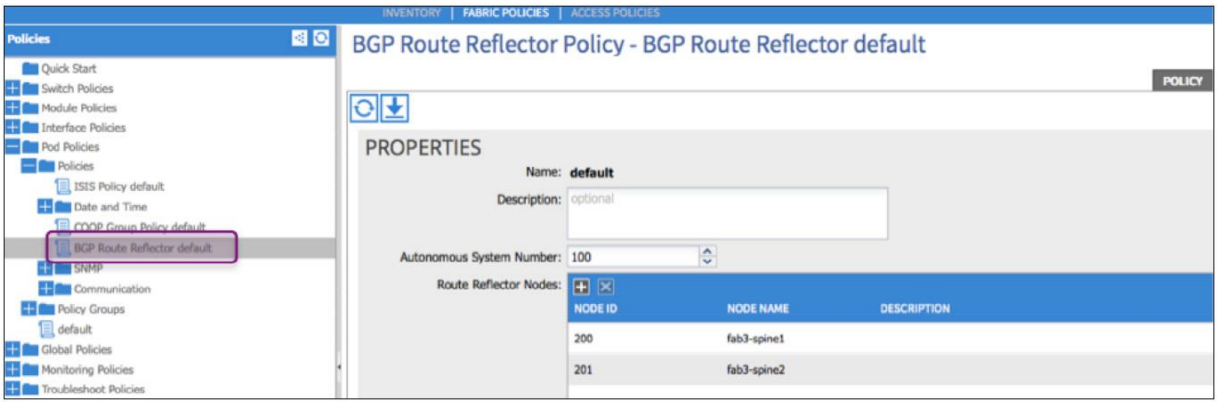
All external routers are connected to leaf nodes. The spine node has no physical connection to external routers. Spine nodes only have BGP sessions with leaf switches and they can't be used as BGP RR for external routers.

Configuring the BGP AS number and RR nodes involves three steps on the APIC GUI. First, configure BGP route reflector policy under menu **Fabric→Fabric Policies→ POD Policies**. Then combine the configured BGP route reflector policy, along with other policies (if required) to a policy group. Lastly, apply the policy group configured in the previous step to the POD as POD profile.

Here is a closer look at how to enable MP-BGP within the ACI fabric.

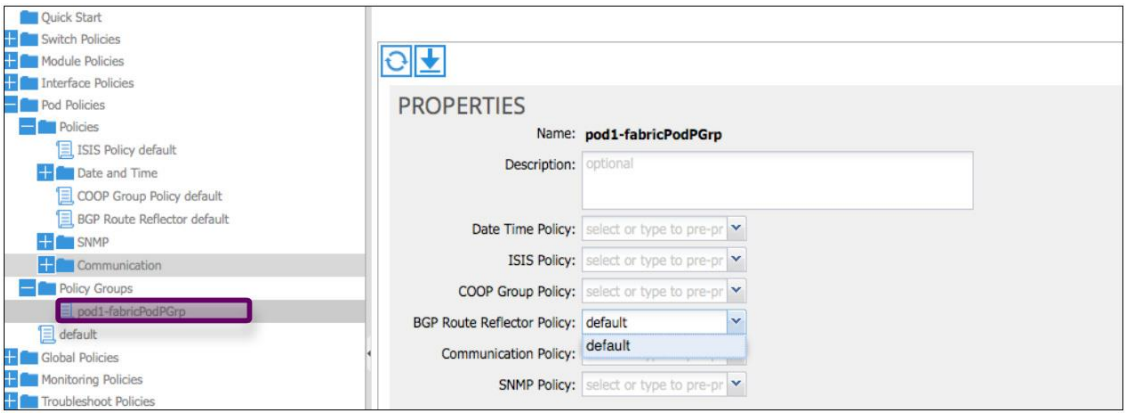
1. **Create a BGP Route Reflector policy** named "default" by going to the menu, **Fabric→Fabric Policies→Pod Policies→BGP Route Reflector default**. As shown below in Figure 8, specify the AS number and add a spine node ID as the route reflector.

Figure 8. Specifying BGP AS Number and Spine Nodes as Route Reflectors



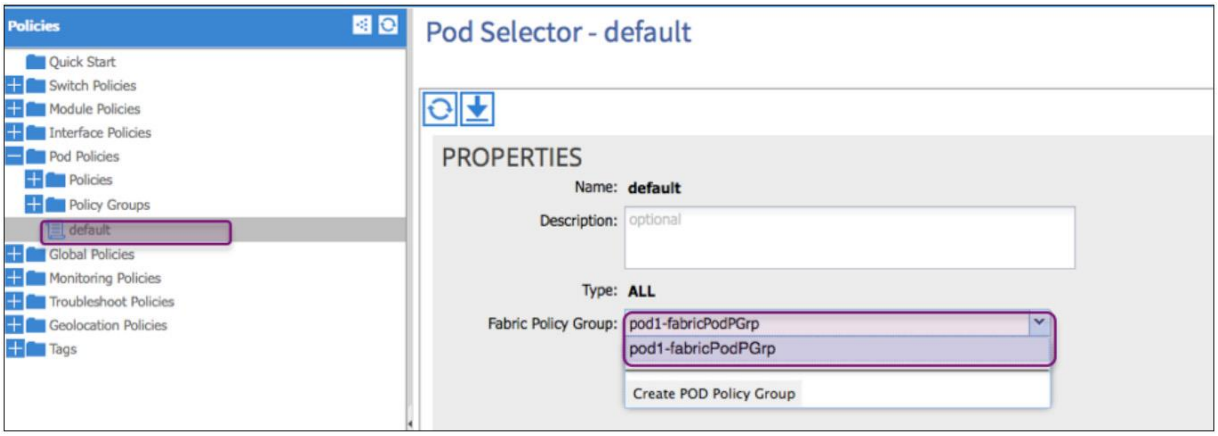
2. **Create a POD policy group** by going to menu **Fabric→Fabric Policies→Pod Policies→Policy Groups**. The policy group allows users to combine multiple policies, such as BGP policy, Integrated System to Integrated System (IS-IS) routing protocol policy, co-operative (COOP) policy, and others, to a policy group and apply it to the POD. Figure 9 offers an example, creating a policy group named “pod1-fabricPodPGrp” that includes the BGP route reflector policy named “default”.

Figure 9. Creating POD Policy Groups



3. Apply the policy group to the POD by going to menu **Fabric→Fabric Policies→Pod Policies→default**. Select the policy group “pod1-fabricPodPGrp” created in the previous step as fabric policy group (Figure 10).

Figure 10. Selecting the Policy Group as POD Policy



This configuration task can also be achieved by posting the following XML file to URI:

https://<apic_IP>/api/policymgr/mo/uni.xml

```
<fabricInst>
  <fabricFuncP>
    <fabricPodPGrp name="pod1-fabricPodPGrp">
      <fabricRsPodPGrpBGPRRP tnBgpInstPolName="default" />
    </fabricPodPGrp>
  </fabricFuncP>
```

```
  <bgpInstPol name='default' descr='Fabric Default BGP Policy'>
    <bgpAsP name='aspn' asn='100'/>
    <bgpRRP name='route-reflector'>
      <bgpRRNodePEp id='200'/>
      <bgpRRNodePEp id='201'/>
    </bgpRRP>
  </bgpInstPol>
  <fabricPodP name="default">
    <fabricPodS name="default" type="ALL">
      <fabricRsPodPGrp tDn="uni/fabric/funcprof/podpgrp-pod1-fabricPodPGrp"/>
    </fabricPodS>
  </fabricPodP>
</fabricInst>
```

OSPF Routing Protocol Peering between ACI and the External Router

OSPF Area Type

At the time of this writing border leaf switches only support the Not So Stubby Areas (NSSA). The fact that the ACI fabric is designed to be a stub network, supporting NSSA reduces the size of the OSPF database, and the border leaf switches need to maintain and reduce the overhead of the routing protocols. This also implies that the ACI border leaf switches will not be in area 0 and will not provide Area Border Router (ABR) functionality. Although the APIC GUI and object model for OSPF don't provide area-type configurations, users need to set the area type on the external routers to be a NSSA in order to bring up OSPF adjacency.

Supported Interface Type

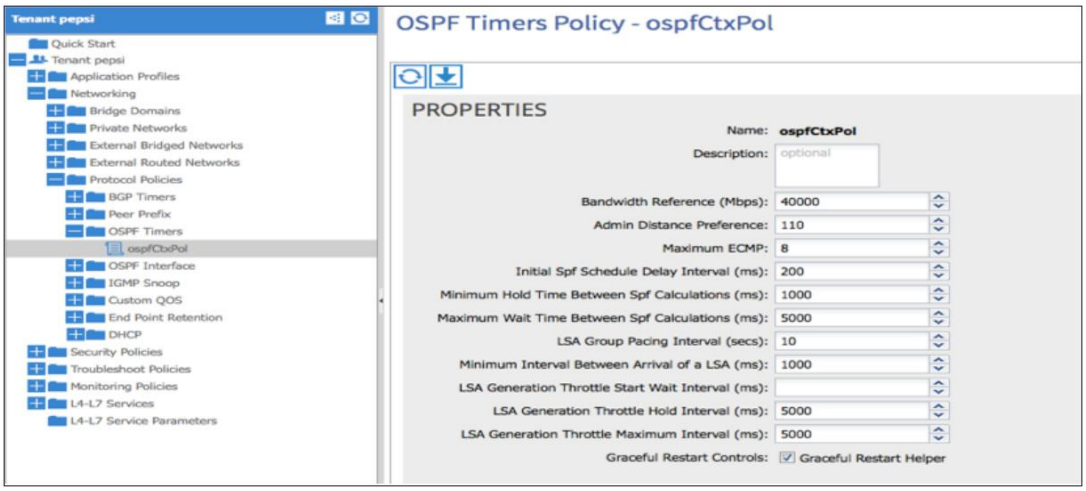
When peering with external routers with OSPF, users have the choice of three different types of interfaces: layer 3 interface, SVI, and sub-interface. Sub-interface is used to provide a layer 3 outside connection for multiple private networks on the same physical interface. An SVI interface is needed when same physical interface is used to provide layer 2 and 3 outside connections. Please that this SVI interface only presents on the border leaf and is not enabled on other leaves.

OSPF Protocol Parameters Tuning

The OSPF implementation of the ACI border leaf supports parameters tuning for the OSPF interface and for the OSPF global timer and parameters.

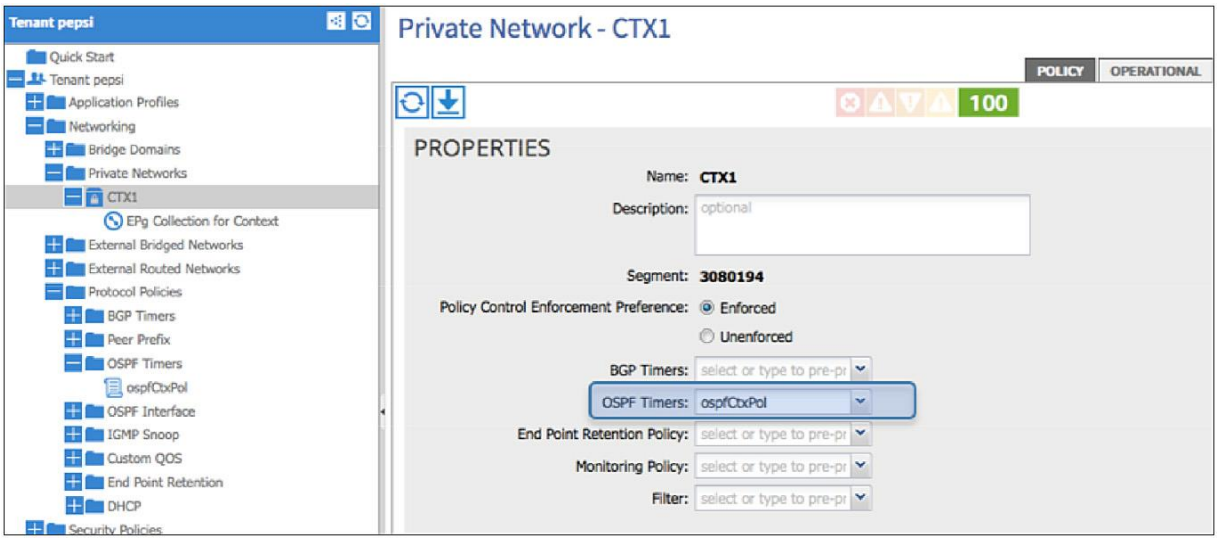
The OSPF timer and parameters displayed in Figure 11 can be tuned with policy configuration.

Figure 11. OSFP Timers and Parameters



User needs to keep in mind that these parameters are configurable, and applicable on a per-private network (or VRF) basis. To apply the configured OSPF timer policy as shown in Figure 11 go to the menu **Tenant→ Networking→ Private Networks** and choose the OSPF timer policy under drop-down menu "OSPF Timers" (Figure 12).

Figure 12. Applying Configured OSPF Timer Policy

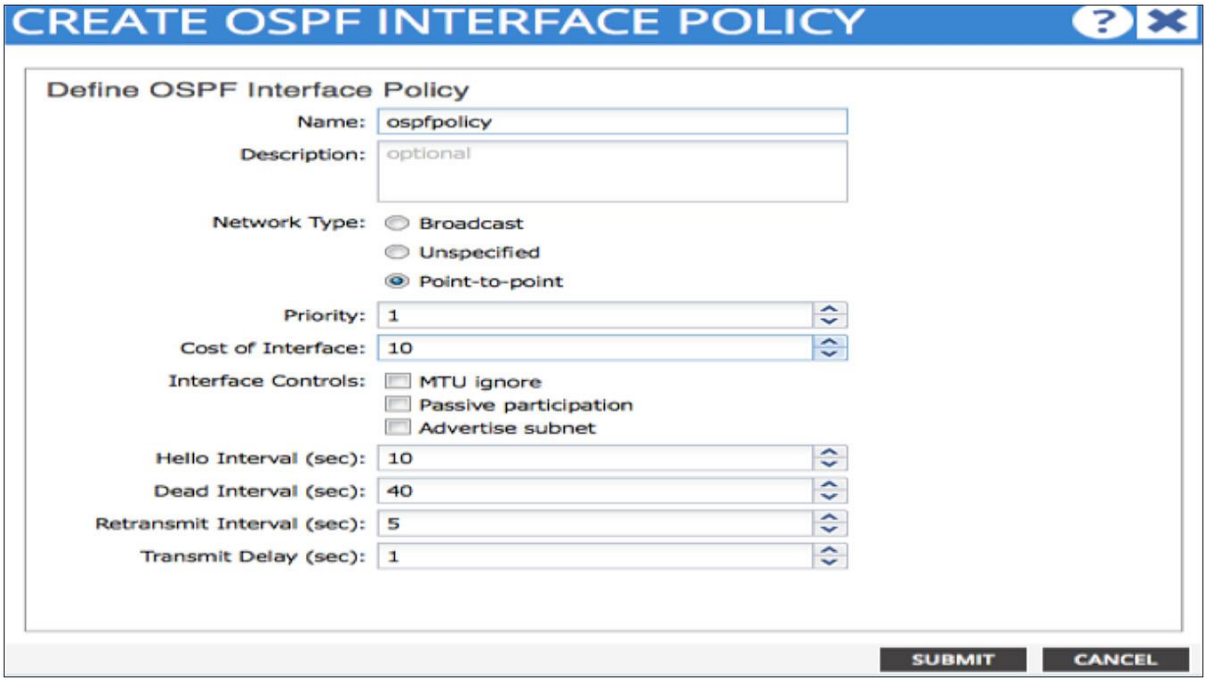


Under the OSPF interface policy the supported configuration parameters are:

- OSPF network type
- OSPF interface timer
- OSPF authentication type and key
- Designated Router (DR) priority
- OSPF interface cost
- Passive interface

Figure 13 shows how to create an OSPF interface policy.

Figure 13. OSPF Interface Policy Creation



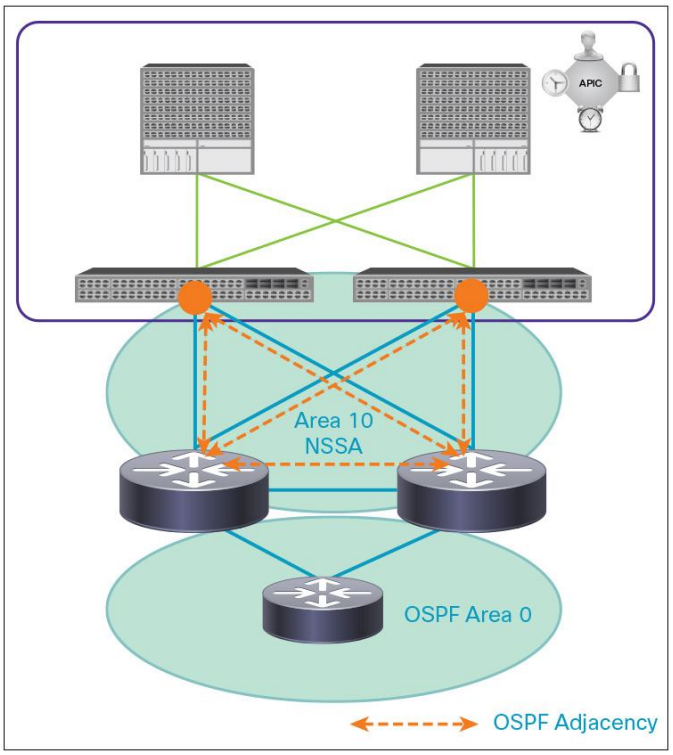
The OSPF adjacency authentication type and key configuration options are not displayed in Figure 13. Those options are configured under the OSPF interface profile. (See step 3 for a layer 3 outside connection with OSPF example).

OSPF High-Availability Design

An important consideration for OSPF design is that there is no OSPF adjacency between two ACI border leaf switches when layer 3 interfaces or the sub-interface are used to connect to the external routers. In the case of an SVI being used for external connections, both border leaf switches and external routers reside on the same VLAN and same subnet, and they do form OSPF adjacency.

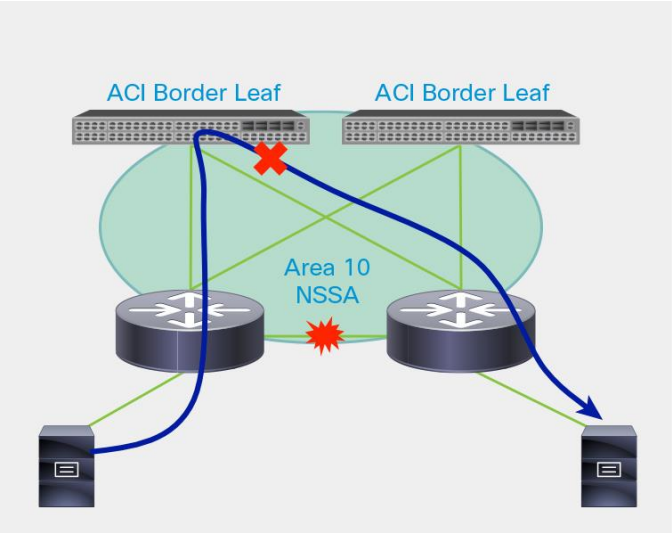
Figure 14 shows the OSPF adjacency among two border leaf switches and external routers with a layer 3 interface or sub-interface. It highlights the fact that there is no OSPF adjacency between two border leaves.

Figure 14. OSPF Adjacency between Border Leaf and External Routers



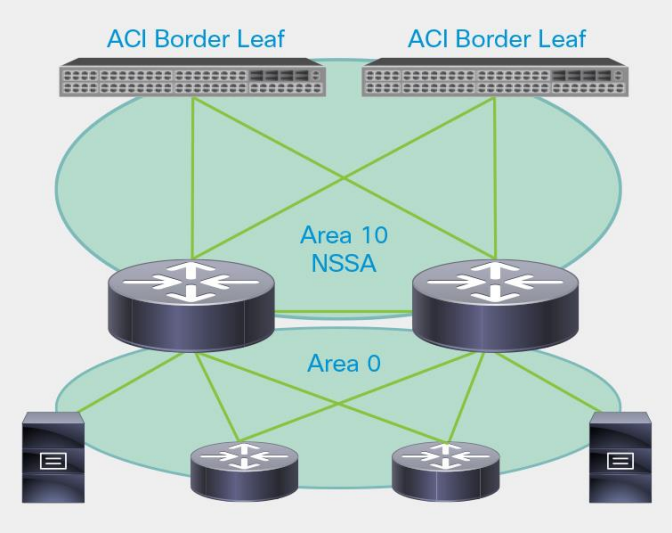
The proper design also requires the two external routers to have redundant links between one another. As explained earlier, the ACI fabric is designed to be a stub network and it can't be used to carry transit traffic between two external routers. The design in Figure 15 may obviate the traffic between two external routers when the only direct link between the two routes fails.

Figure 15. OSPF Failure Scenario



When there is only one direct connection between two external routers, the OSPF routing table for the subnet attached to another router will point to the ACI border leaf. However, when the ACI border leaf receives the traffic it may drop the packet due to the policy or security implementation. The policy model of ACI assumes either source or destination of the traffic stream is an endpoint learned by the ACI fabric. Because of this policy behavior, it is recommended to have redundant links between two external routers to prevent the routing protocol from choosing the ACI border leaf as a transit router. Figure 16 shows one alternative design to alleviate this situation.

Figure 16. OSPF Alternative Design



Tag Tenant Routes Using OSPF Route Policy

The ACI border leaf provides the capability to allow users to tag the tenant routes when it injects tenant public subnets into OSPF protocol. As explained previously, border leaf switches support NSSA, and the tenant public subnets are redistributed to OSPF as type 7 Link State Advertisement (LSA). The tagging mechanism allows the network architect to group the tenant routes and apply required route policy to these routes by matching the tag value instead of the IP address prefix. In doing so the external router doesn't need to change the route policy when a tenant subnet is added or removed in the ACI fabric.

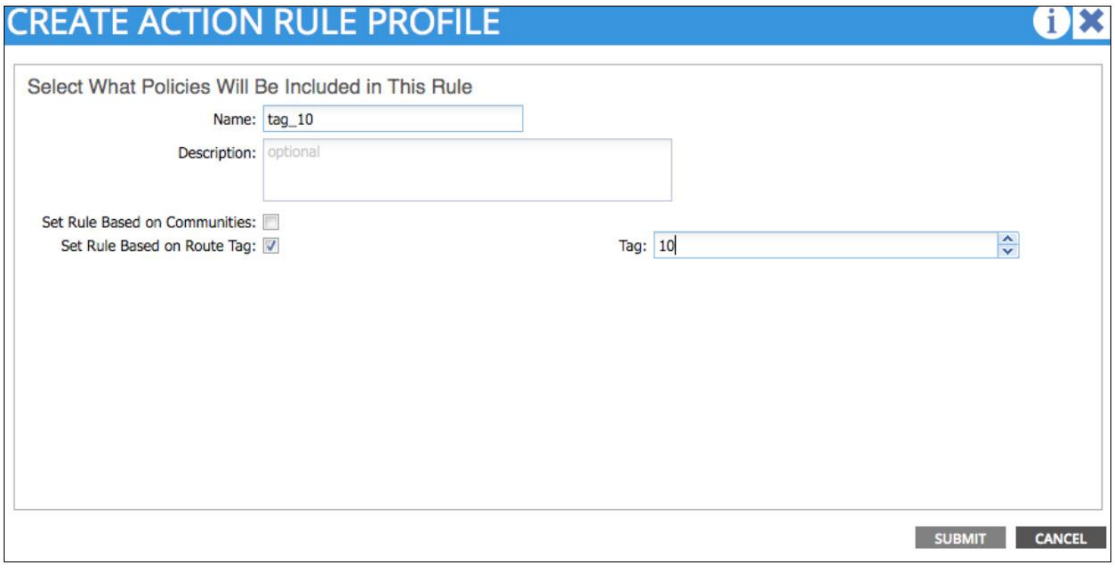
The configuration of the OSPF route tag involves the following steps:

- Configure action rules with the value that the user wants to have in the tag attribute of OSPF routes
- Create a route profile that includes the action rule
- Associate the route profile with a bridge domain or subnet of a bridge domain

Let's use a configuration example to explain these steps.

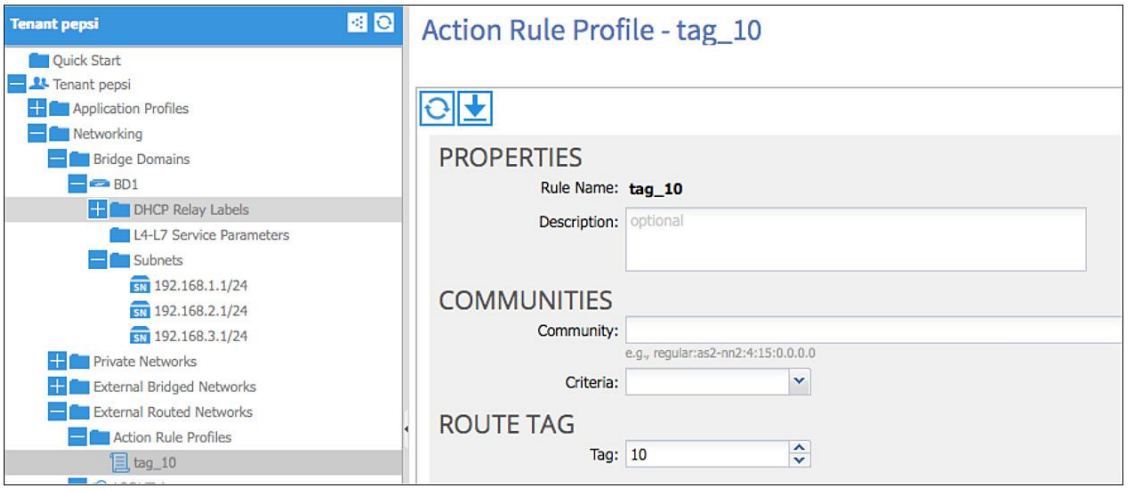
1. Go to menu **Tenant→External Routed Networks→Action Rule Profiles** to create a new action rule (Figure 17).

Figure 17. Example of How to Create an Action Rule Profile



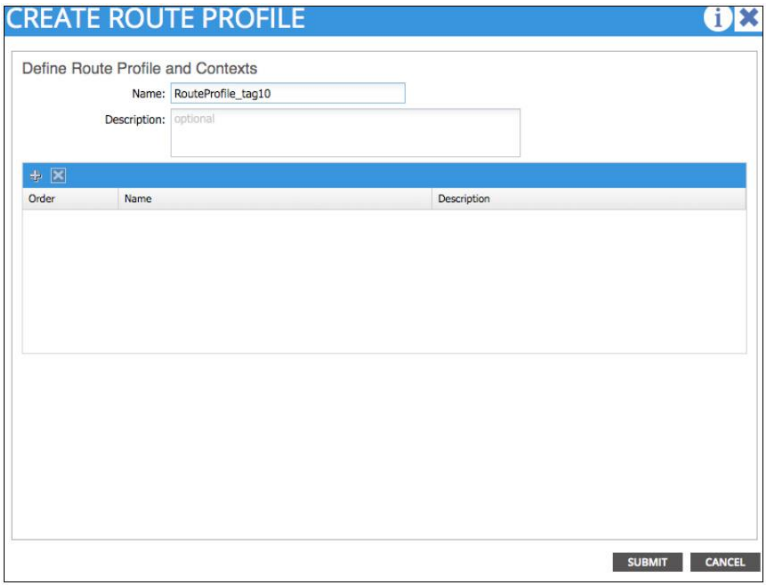
This action rule will mark route tag with value 10. Figure 18 shows how it looks after the configuration.

Figure 18. Route Action Rule Configuration Result



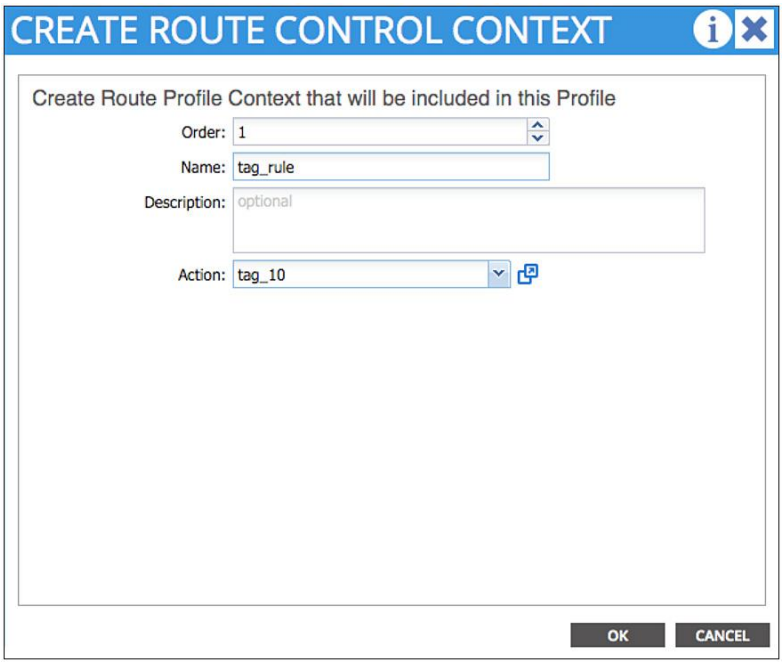
2. Go to menu **Tenant→External Routed Networks→Name of your layer 3 outside connection→Route Profiles→Actions** to create a new route profile for your layer 3 outside connection (Figure 19). Assume you have configured one layer 3 outside connection.

Figure 19. Creating a Route Profile for a Layer 3 Outside Connection



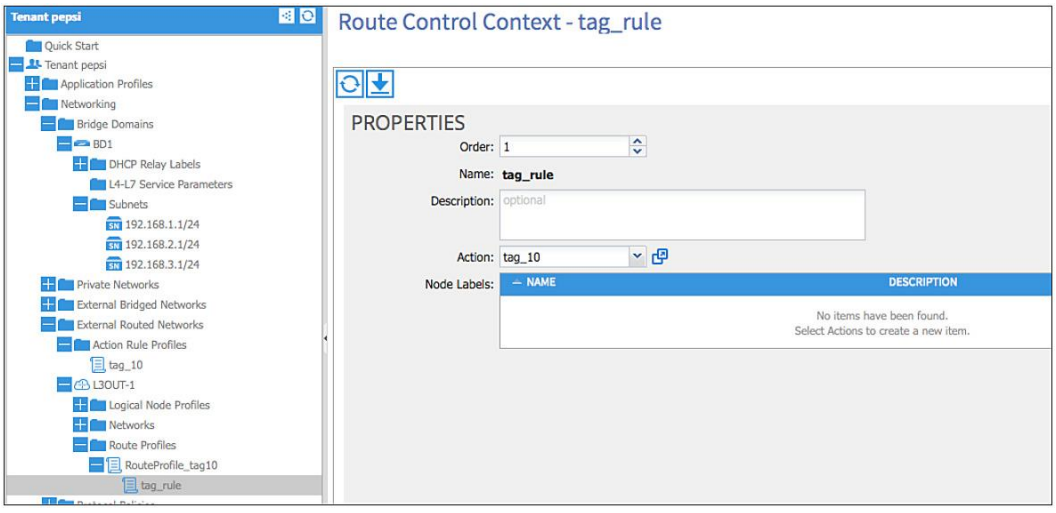
After you create the route profile, named “RouteProfile_tag10”, click the “+” sign to add action rules like shown in Figure 20.

Figure 20. Adding Action Rules



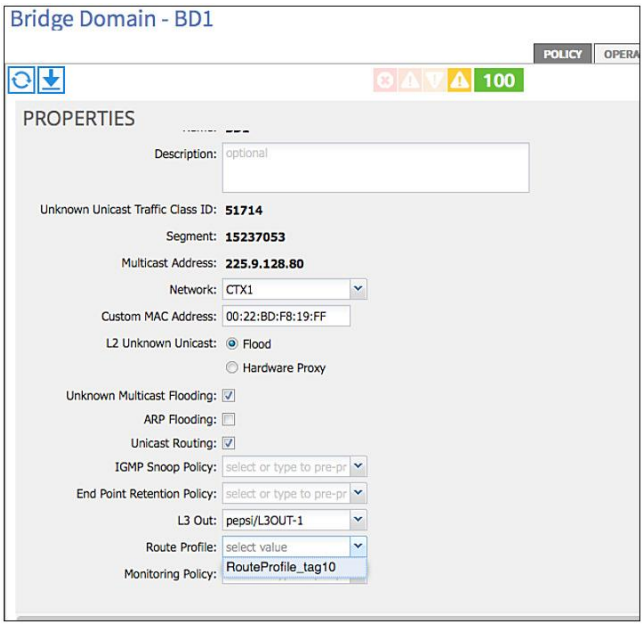
This step effectively includes the action rule “tag_10” to this route profile. With this action rule the route profile will mark the routes with tag value 10. Figure 21 shows how the route profile looks in the APCI GUI after completion of configuration.

Figure 21. Route Profile Configuration



3. Next, associate the route profile with a bridge domain or certain subnets of a bridge domain. Figure 22 displays an example that associates a route profile with a bridge domain.

Figure 22. Associating the Route Profile with a Bridge Domain



When associating a route profile with a bridge domain all of the subnets under the bridge domain will be marked with the same tag value. The software allows the user to associate a route profile with a subnet of a bridge domain; this provides flexibility to mark different tag values for different subnets. When a route profile is specified under both the bridge domain and the subnets of a bridge domain the route profile under the subnet takes precedence.

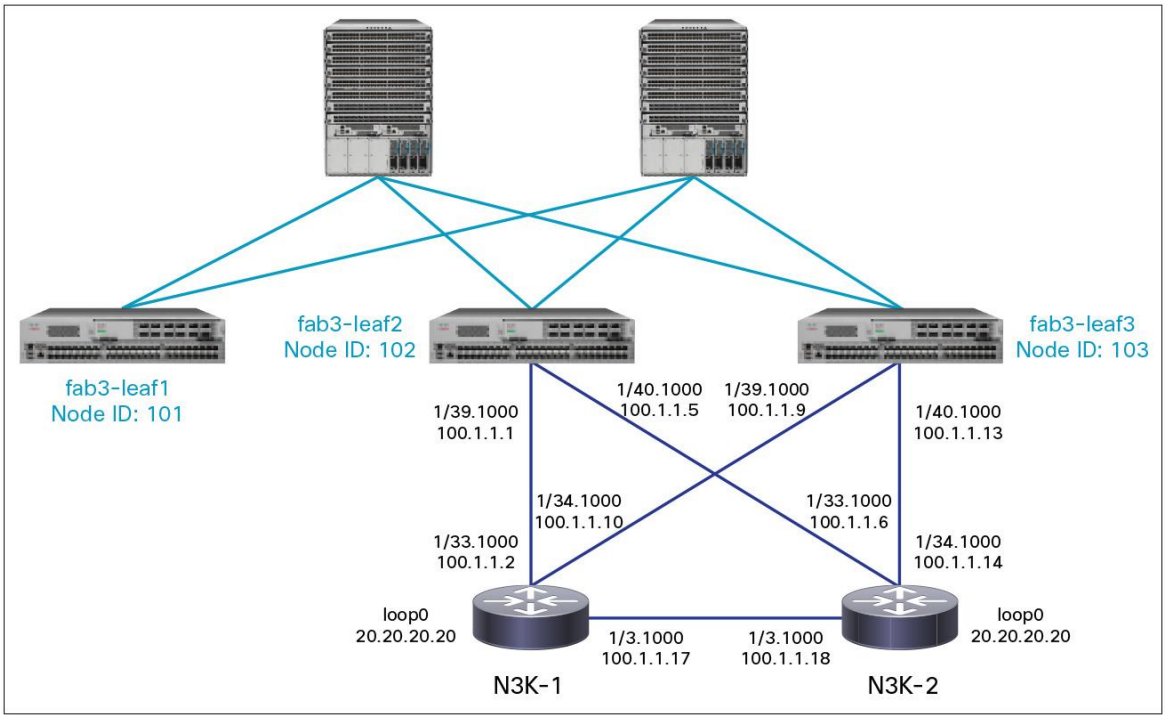
4. Finally, verify the OSPF route is tagged with value 10 on the external router that connects to the ACI border leaf.

```
N9396-4# sh ip ro ospf
IP Route Table for VRF "default"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

192.168.3.0/24, ubest/mbest: 1/0
    *via 100.1.1.1, Eth1/18, [110/20], 00:00:02, ospf-1, nssa type-2, tag 10
N9396-4#
```

Layer 3 Outside Connection with OSPF Example

In this section we will explore how to deploy a layer 3 outside connections with OSPF on the topology shown in Figure 23.

Figure 23. Topology for Layer 3 Outside Connection with OSPF Example

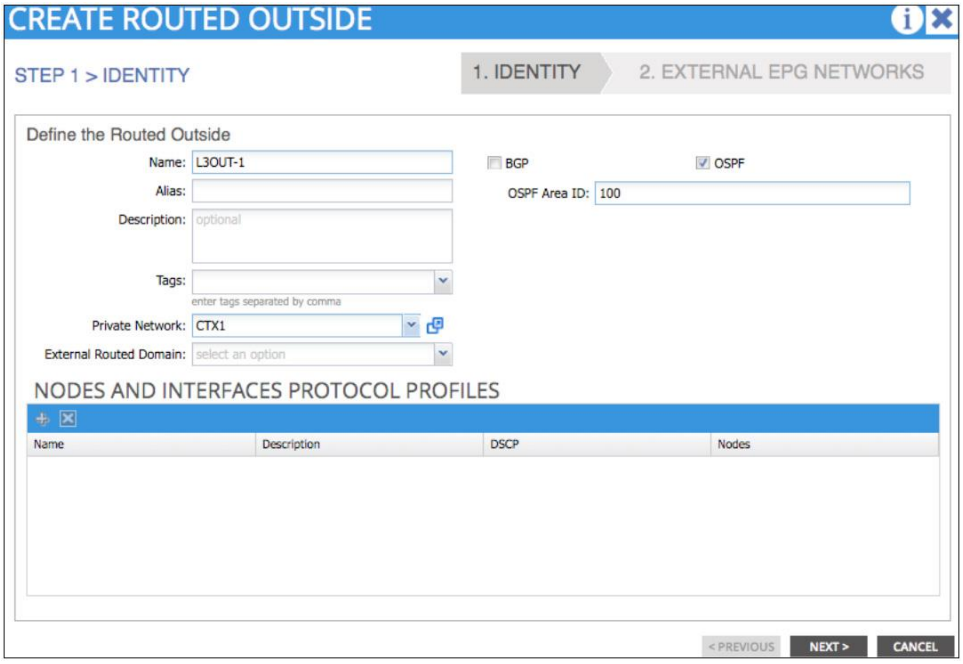


In this network are two spines and three leaves with two of them acting as border leaf switches. Two Cisco Nexus® 3000 switches simulate two external routers that connect the ACI fabric to the rest of the data center network or to the WAN. OSPF is deployed between the two border leaf switches and two Nexus 3000 switches, and they are in the same NSSA.

In this example, one tenant is created, named “pepsi”. The tenant has one EPG, called “WEB”, which resides in the bridge domain called “BD1”.

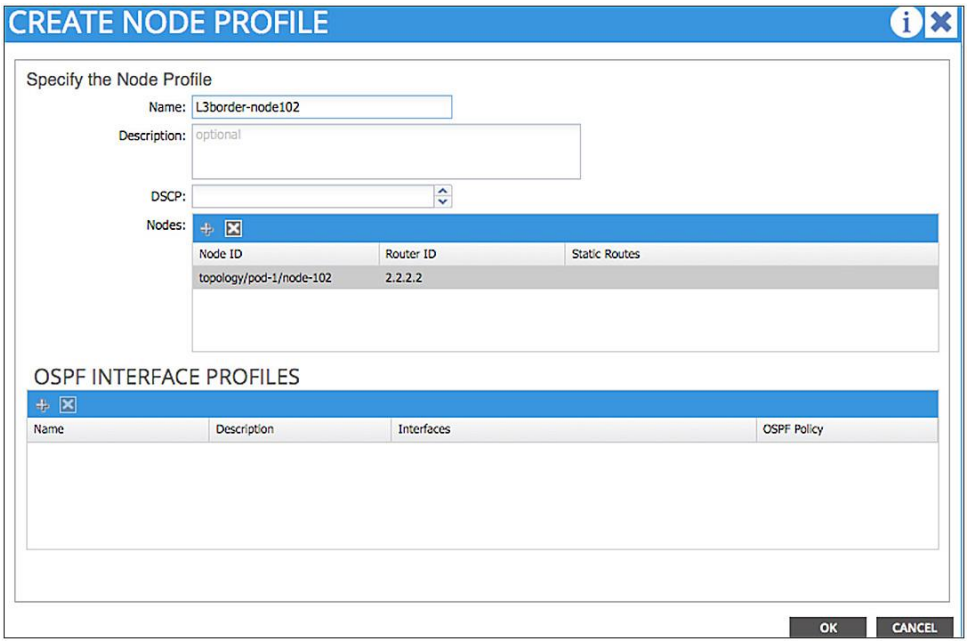
1. **Start to create a layer 3 outside connection**. Create a layer 3 outside connection, called"L3OUT-1", under menu Tenant→Networking→External Routed Networks. Select OSPF as the protocol, choose area ID of 100, and associate the layer 3 outside connection with the private network, "CTX1" (Figure 24).

Figure 24. Start to Create L3 Outside Connection



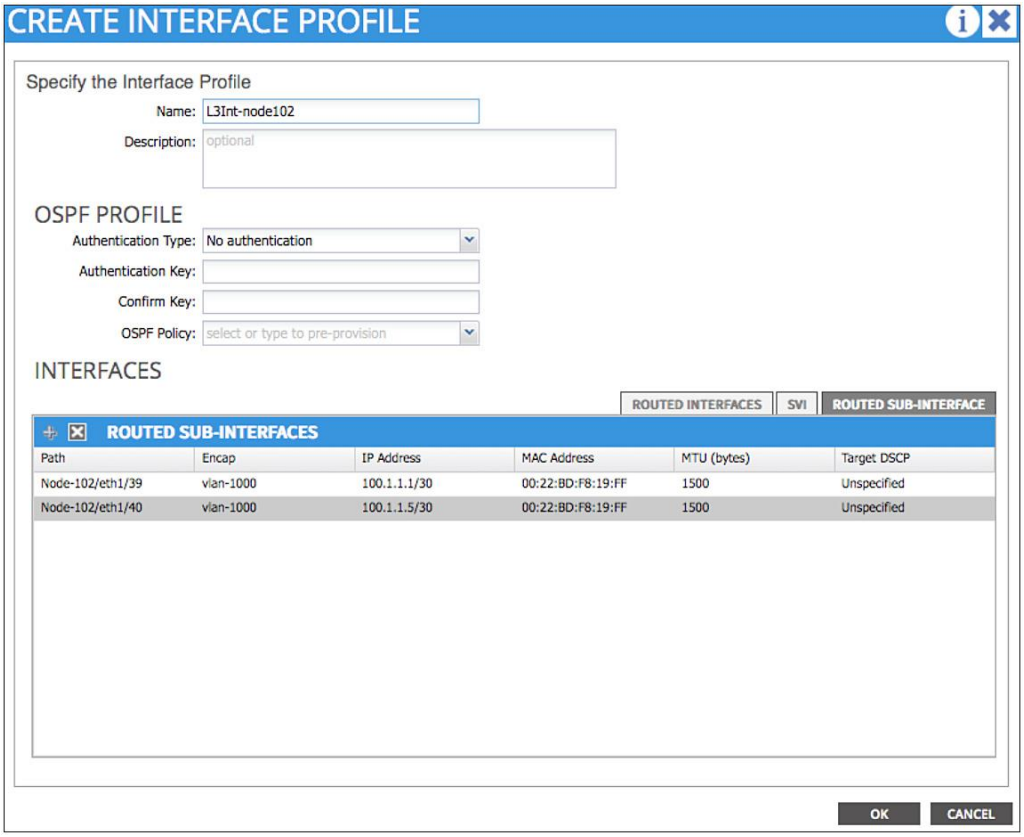
2. **Add the layer 3 border leaf node.** Click the "+" sign under the "**Nodes and Interface Profile Profiles**" and follow the wizard to start to a add border leaf node for the layer 3 outside connection (Figure 25).

Figure 25. Adding a Layer 3 Border Leaf



3. **Add layer 3 interfaces for this border leaf**. Click the "+" sign under "OSPF Interface Profiles". Figure 26 provides an example of how to add two layer 3 sub-interfaces (sub-interface for eth1/39 and eth1/40 with dot1q tag 1000) on border leaf node 102 to connect to the two Nexus 3000s. Specify the MTU to be 1500. Leave the OSPF policy empty for the time being.

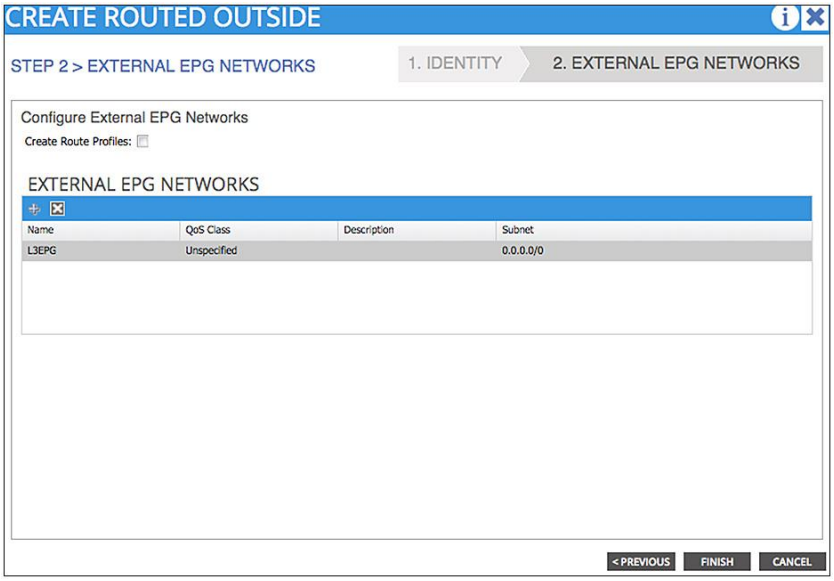
Figure 26. Adding Two Layer 3 Sub-Interfaces



4. Repeat steps 2 and 3 to add node 103 as a border leaf node for "L3OUT-1". Add two sub-interfaces (eth1/39 and eth1/40 with dot1q tag 1000) for the border leaf node.

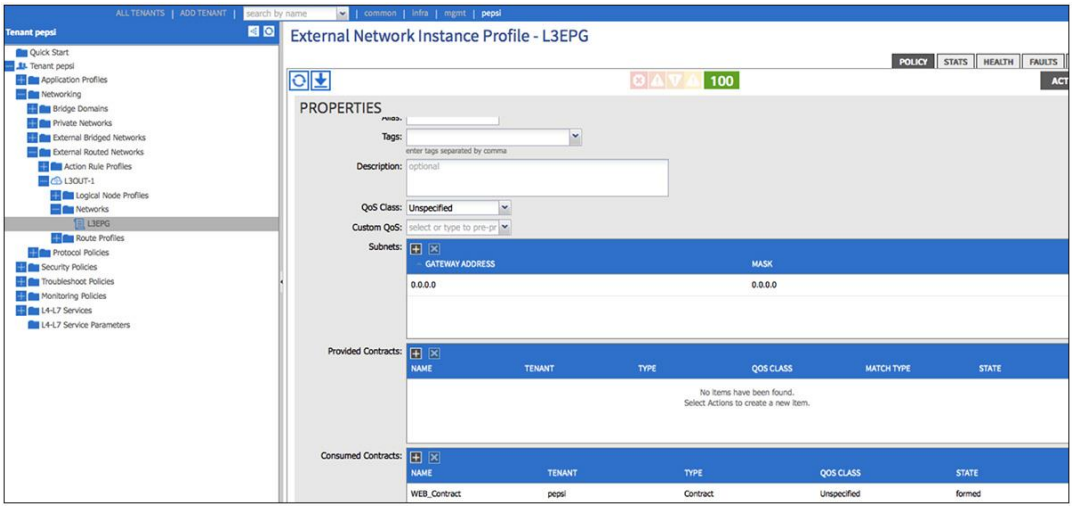
5. Click ‘Next” to start to configure the external EPG. The ACI fabric maps external layer 3 endpoints to the external EPG by using the IP prefix and mask. One or more external EPGs can be supported for each layer 3 outside connection, depending on whether the user wants to apply a different policy for different groups of external endpoints. In this example we treat all outside endpoints equally and create only one external EPG. The ACI policy model requires an external EPG and the contract between the external EPGs and inside EPGs. Without this, all connectivity to outside will be blocked, even if external routes are learned properly. This is part of the security model of ACI (Figure 27).

Figure 27. Configuring the External EPG



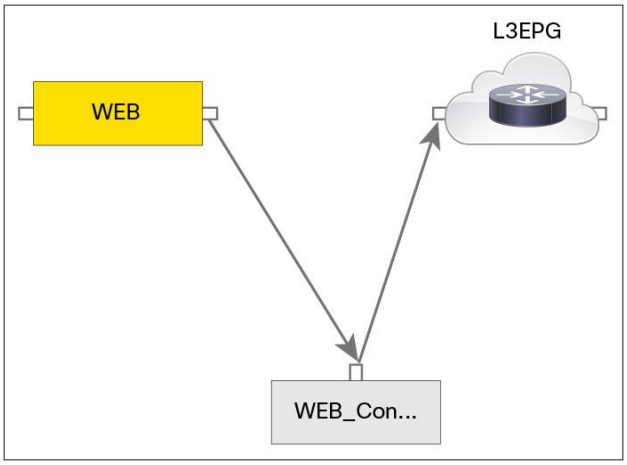
6. **Configure a contract between the external and internal EPG.** In this example, we specify “WEB_contract” as consumed contract for external EPG “L3EPG” (steps to configure this contract is skipped here). Click “**Finish**” in the previous step and go to menu **Tenant→Tenant Pepsi→Networking→External Routed Networks→L3OUT-1→Networks→L3EPG**. Click “+” under the section of “Consumed Contracts” to add “WEB_cotract” as the consumed contract (Figure 28).

Figure 28. Configuring a Contract for External EPG



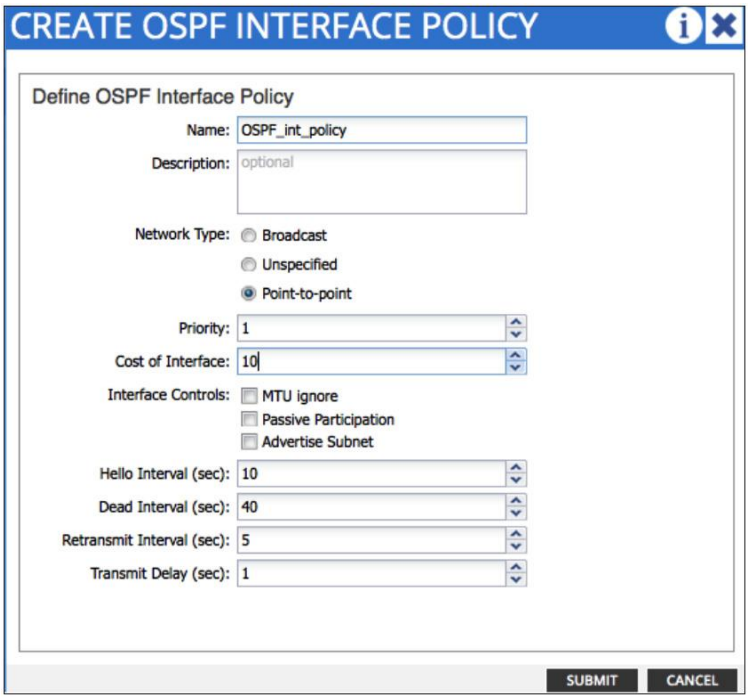
Next go to menu Application Profiles→**Application1→Application EPGs→WEB EPG** and add the same contract, "WEB_contract", as the provided contract for WEB EPG. Once the consumer-provider relationship is established, check the application profile; it should look like Figure 29 on the APIC GUI. It states that communication between WEB EPG and L3EPG is regulated by policy "WEB_contract". Note that without a contract all communications between EPGs are blocked, including the communication with external EPGs.

Figure 29. Contract Relationship with L3 External EPG



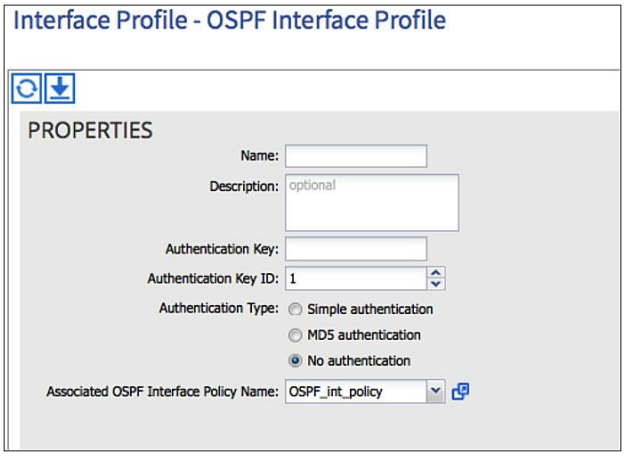
7. **Create OSPF interface policy** by going to menu **Tenant→Networking→Protocol Policies→OSPF Interface** (Figure 30).

Figure 30. Creating the OSPF Interface Policy



8. **Associate the OSPF interface policy with the sub-interfaces** on the two border leaf switches by going to menu **External Routed Networks→Logical Node Profiles→Logical Interface Profiles** (Figure 31). You will need to repeat this step for both border leaf switches.

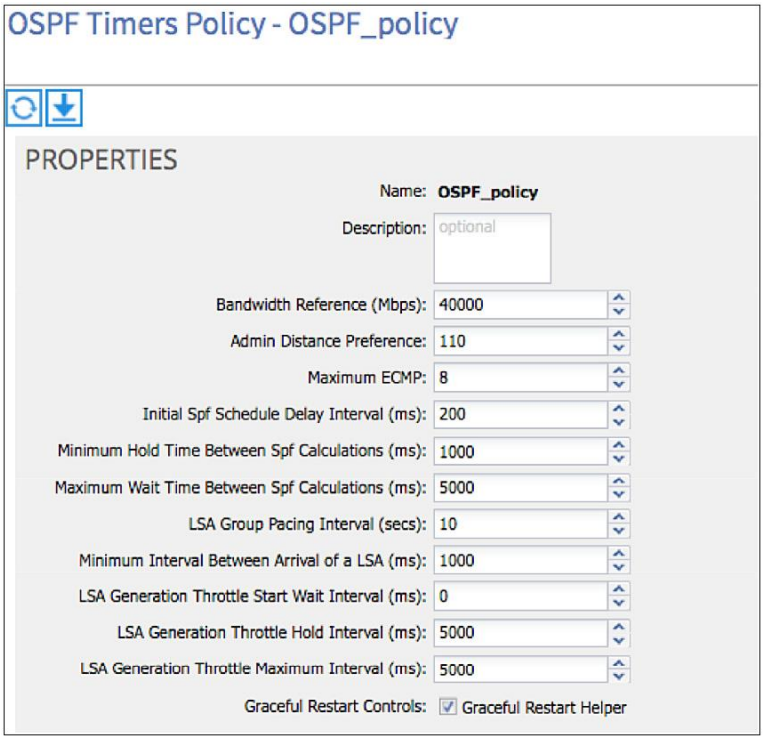
Figure 31. Associating the OSPF Interface Policy with Sub-Interfaces



9. **Configure the policy for OSPF protocol parameters and associate it with private networks** (Figure 32). This is equivalent to configuring OSPF parameters under the VRF of “router ospf”.

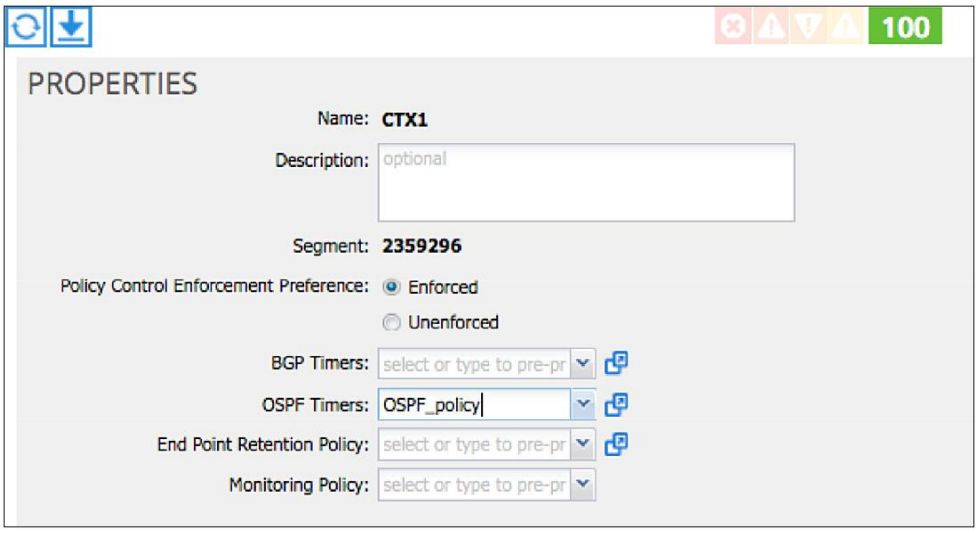
   Create the OSPF policy by going to menu **Tenant→Networking→Protocol Policies→OSPF Timers**.

Figure 32. OSPF Protocol Parameters Policy Configuration



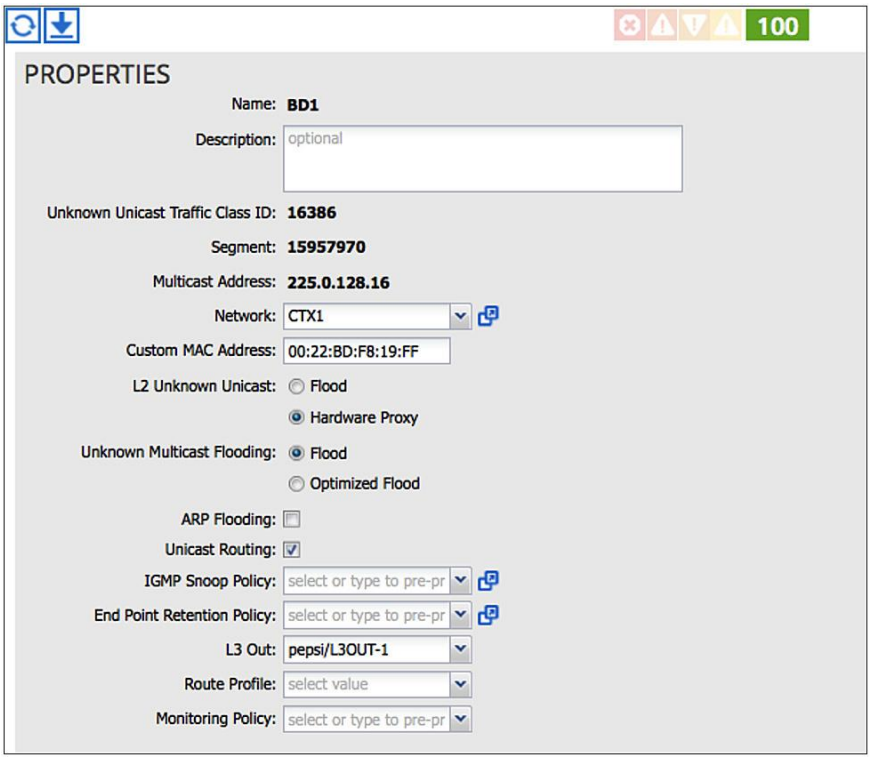
Associate the policy with private network CTX1 for this tenant by going to menu **Tenant→Networking→Private Networks** (Figure 33).

Figure 33. Associating Configured OSPF Protocol Policy with Private Network



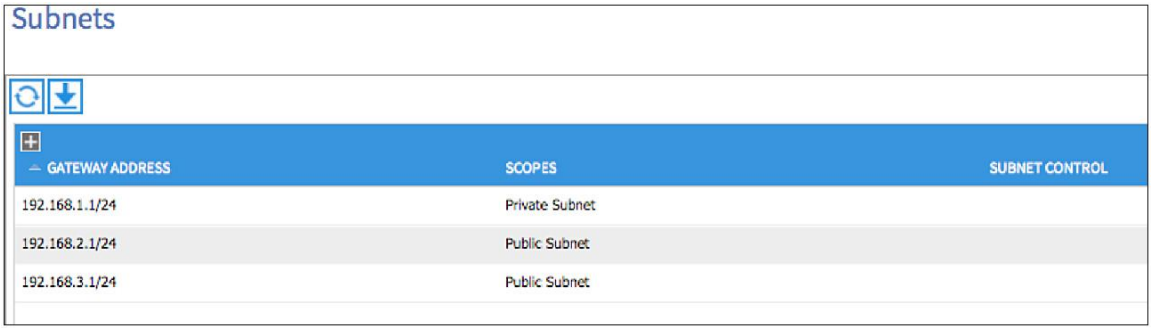
10. **Associate the layer 3 outside connection with bridge domain** “BD1” for this tenant (Figure 34). Repeat this step if there are multiple bridge domains for the tenant.

Figure 34. Associating Layer 3 Outside Connection with Bridge Domain



Under bridge domain “BD1” there are three subnets; two of them should be configured as public subnets (Figure 35). With this association, these two public subnets will be advertised to external routers by OSPF.

Figure 35. Subnet Scope of Bridge Domain



With above steps the Layer 3 outside connection configuration o APIC GUI is complete. If required by design, you have the option to configure policy to set the tag value for the tenant routes. Follow the steps explained in the section, “Tag Tenant Routes Using OSPF Route Policy.” Users must configure a BGP AS number and route reflector as explained in the section, “Route Distribution within the ACI Fabric.” Otherwise, the external routes won’t be propagated to non-border leaf switches.

The next step is to apply the related configuration on external routes. On the two Nexus 3000 switches, create VRF “pepsi” and use sub-interfaces with dot1q tag 1000 to connect to the two border leaves. The two switches generate default routes and advertise them to ACI border leaves. Apply the following configuration on Nexus 3000-1:

```
interface Ethernet1/3.1000
  encapsulation dot1q 1000
  vrf member pepsi
  ip address 100.1.1.17/30
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.100
interface Ethernet1/33
  description to leaf102 eth1/39
  no switchport
interface Ethernet1/33.1000
  encapsulation dot1q 1000
  vrf member pepsi
  ip address 100.1.1.2/30
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.100
interface Ethernet1/34
  description to leaf103 eth1/39
  no switchport
interface Ethernet1/34.1000
  encapsulation dot1q 1000
  vrf member pepsi
  ip address 100.1.1.10/30
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.100
interface loopback0
  vrf member pepsi
  ip address 20.20.20.20/32
  ip router ospf 1 area 0.0.0.100
router ospf 1
  vrf pepsi
    router-id 4.4.4.4
    area 0.0.0.100 nssa default-information-originate
```

Apply the same configuration (with the correct IP address) on the second Nexus 3000. Once the configuration is finished you can check the OSPF adjacency and routing table on both ACI border leaf switches and Nexus 3000 switches. As shown in the following output, Nexus 3000-1 has OSPF adjacency with two border leaves (with router ID 2.2.2.2 and 3.3.3.3) and the Nexus 3000-2 (with router ID 5.5.5.5).

```
N3K-1# sh ip ospf neighbors vrf pepsi
 OSPF Process ID 1 VRF pepsi
 Total number of neighbors: 3
 Neighbor ID     Pri State            Up Time  Address         Interface
 5.5.5.5           1 FULL/ -          10:19:04 100.1.1.18      Eth1/3.1000
```

```
 2.2.2.2           1 FULL/ -          10:21:08 100.1.1.1       Eth1/33.1000
 3.3.3.3           1 FULL/ -          10:21:16 100.1.1.9       Eth1/34.1000
```

In addition to all subnets for the links between Nexus 3000 and border leaves, Nexus 3000-1 learns the two public subnets: 192.168.2.0/24 and 192.168.3.0/24. The private subnet - 192.168.1.0/24 - is contained within the ACI fabric and is not propagated to the Nexus 3000 routing table, as expected.

```
N3K-1# sh ip ro vrf pepsi
IP Route Table for VRF "pepsi"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

0.0.0.0/0, ubest/mbest: 1/0
    *via Null0, [1/0], 10:38:56, static
2.2.2.2/32, ubest/mbest: 1/0
    *via 100.1.1.1, Eth1/33.1000, [110/5], 10:21:12, ospf-1, intra
3.3.3.3/32, ubest/mbest: 1/0
    *via 100.1.1.9, Eth1/34.1000, [110/5], 10:21:15, ospf-1, intra
20.20.20.20/32, ubest/mbest: 2/0, attached
    *via 20.20.20.20, Lo0, [0/0], 10:38:55, local
    *via 20.20.20.20, Lo0, [0/0], 10:38:55, direct
100.1.1.0/30, ubest/mbest: 1/0, attached
    *via 100.1.1.2, Eth1/33.1000, [0/0], 10:38:48, direct
100.1.1.2/32, ubest/mbest: 1/0, attached
    *via 100.1.1.2, Eth1/33.1000, [0/0], 10:38:48, local
100.1.1.4/30, ubest/mbest: 1/0
    *via 100.1.1.18, Eth1/3.1000, [110/8], 10:19:08, ospf-1, intra
100.1.1.8/30, ubest/mbest: 1/0, attached
    *via 100.1.1.10, Eth1/34.1000, [0/0], 10:38:48, direct
100.1.1.10/32, ubest/mbest: 1/0, attached
    *via 100.1.1.10, Eth1/34.1000, [0/0], 10:38:48, local
100.1.1.12/30, ubest/mbest: 1/0
    *via 100.1.1.18, Eth1/3.1000, [110/8], 10:19:08, ospf-1, intra
100.1.1.16/30, ubest/mbest: 1/0, attached
    *via 100.1.1.17, Eth1/3.1000, [0/0], 10:19:28, direct
100.1.1.17/32, ubest/mbest: 1/0, attached
    *via 100.1.1.17, Eth1/3.1000, [0/0], 10:19:28, local
192.168.2.0/24, ubest/mbest: 2/0
    *via 100.1.1.1, Eth1/33.1000, [110/20], 10:21:12, ospf-1, nssa type-2
    *via 100.1.1.9, Eth1/34.1000, [110/20], 10:21:15, ospf-1, nssa type-2
192.168.3.0/24, ubest/mbest: 2/0
    *via 100.1.1.1, Eth1/33.1000, [110/20], 10:21:12, ospf-1, nssa type-2
    *via 100.1.1.9, Eth1/34.1000, [110/20], 10:21:15, ospf-1, nssa type-2
```

OSPF Database for VRF pepsi on the Nexus 3000 Switches. Two border leave switches redistribute the tenant public subnets as type-7 LSA.

```
N3K-1# sh ip ospf database vrf pepsi
        OSPF Router with ID (4.4.4.4) (Process ID 1 VRF pepsi)

                Router Link States (Area 0.0.0.100)

Link ID         ADV Router      Age         Seq#       Checksum Link Count
2.2.2.2         2.2.2.2         847         0x80000034 0x38a0   5
3.3.3.3         3.3.3.3         849         0x8000003a 0x683e   5
4.4.4.4         4.4.4.4         718         0x8000001d 0x355b   7
5.5.5.5         5.5.5.5         716         0x8000001d 0xd4a6   7

                Type-7 AS External Link States (Area 0.0.0.100)

Link ID         ADV Router      Age         Seq#       Checksum Tag
0.0.0.0         4.4.4.4         728         0x80000017 0xbd95    0
0.0.0.0         5.5.5.5         766         0x80000018 0x9db0    0
192.168.2.0     2.2.2.2         1617        0x8000001b 0x61bf    0
192.168.2.0     3.3.3.3         1620        0x8000001a 0x77a2    0
192.168.3.0     2.2.2.2         1617        0x8000001b 0x56c9    0
192.168.3.0     3.3.3.3         1620        0x8000001a 0x6cac    0
```

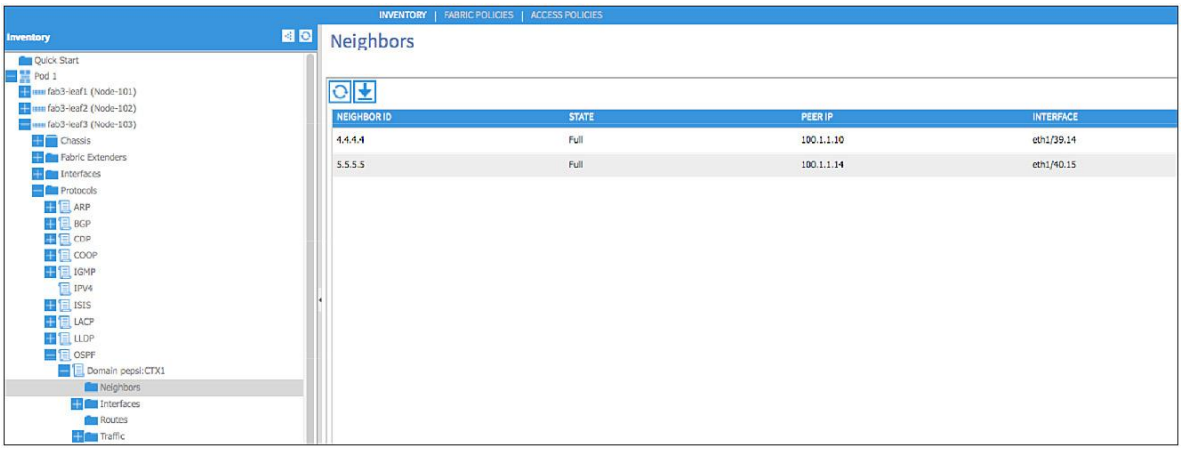
Verify the connectivity by pinging the default gateway IP provided by the ACI fabric.

```
N3K-1# ping 192.168.3.1 vrf pepsi
PING 192.168.3.1 (192.168.3.1): 56 data bytes
64 bytes from 192.168.3.1: icmp_seq=0 ttl=57 time=1.894 ms
64 bytes from 192.168.3.1: icmp_seq=1 ttl=57 time=0.673 ms
64 bytes from 192.168.3.1: icmp_seq=2 ttl=57 time=0.633 ms
64 bytes from 192.168.3.1: icmp_seq=3 ttl=57 time=0.629 ms
64 bytes from 192.168.3.1: icmp_seq=4 ttl=57 time=0.616 ms


--- 192.168.3.1 ping statistics ---
5 packets transmitted, 5 packets received, 0.00% packet loss
round-trip min/avg/max = 0.616/0.888/1.894 ms
N3K-1#
N3K-1# ping 192.168.2.1 vrf pepsi
PING 192.168.2.1 (192.168.2.1): 56 data bytes
64 bytes from 192.168.2.1: icmp_seq=0 ttl=57 time=0.948 ms
64 bytes from 192.168.2.1: icmp_seq=1 ttl=57 time=0.618 ms
64 bytes from 192.168.2.1: icmp_seq=2 ttl=57 time=0.604 ms
64 bytes from 192.168.2.1: icmp_seq=3 ttl=57 time=0.599 ms
64 bytes from 192.168.2.1: icmp_seq=4 ttl=57 time=0.736 ms
```

```
--- 192.168.2.1 ping statistics ---
5 packets transmitted, 5 packets received, 0.00% packet loss
round-trip min/avg/max = 0.599/0.701/0.948 ms
```

Next check the OSPF adjacency and routes on the ACI border leaf. The user has a choice of using APIC GUI, command-line interface (CLI), or representational state transfer (RESTful) API to accomplish this task. Figure 36 shows the border leaf forms adjacency with two Nexus 3000 switches on APIC GUI menu **Fabric→Inventory→Pod-1→node→Protocols→OSPF**. You may notice that there is no OSPF adjacency between the two OSPF border leaves, as explained earlier.

Figure 36. Check OSPF Adjacency between Border Leaf and External Routes on APIC GUI



The user can also check the OSPF routes on the APIC GUI (Figure 37). The output on border leaf fab3-leaf3 (with node ID 103) shows that the border leaf learns two default routes from two Nexus 3000s (Figure 37). The flag "in-rib, v4" indicates that the route is installed in the routing information base (RIB) as best routes. The two tenant public subnets - 192.168.2.0 and 192.168.3.0 - also show up in the OSPF routing table. This is because another border leaf - fab3-leaf2 (node ID 102) - also redistributes the same subnets to the OSPF domain. Border leaf fab3-leaf3 learns the LSA via the Nexus 3000. These two routes are not considered best routes and are not installed in the RIB as indicated by the flag for the routes.

Figure 37. Checking Routes through the APIC GUI

Routes

ACTIONS

NAME	PFX	PATH TYPE	AREA	FLAGS	UNICAST COST	MULTICAST COST	ADDR	IF
Route 0.0.0.0/0, Flags:in-rib,v4, Unicast Cost: 1	0.0.0.0/0	nssa2	0.0.0.100	in-rib,v4	1	1		
Nexthop eth1/39.14-100.1.1.10				v4			100.1.1.10	eth1/39.14
Nexthop eth1/40.15-100.1.1.14				v4			100.1.1.14	eth1/40.15
Route 100.1.1.0/30, Flags:in-rib,v4, Unicast Cost: 14	100.1.1.0/30	intra	0.0.0.100	in-rib,v4	14	14		
Route 100.1.1.12/30, Flags:direct,v4, Unicast Cost: 10	100.1.1.12/30	intra	0.0.0.100	direct,v4	10	10		
Route 100.1.1.4/30, Flags:in-rib,v4, Unicast Cost: 14	100.1.1.4/30	intra	0.0.0.100	in-rib,v4	14	14		
Route 100.1.1.8/30, Flags:direct,v4, Unicast Cost: 10	100.1.1.8/30	intra	0.0.0.100	direct,v4	10	10		
Route 192.168.2.0/24, Flags:v4, Unicast Cost: 20	192.168.2.0/24	nssa2	0.0.0.100	v4	20	20		
Nexthop eth1/39.14-100.1.1.10				v4			100.1.1.10	eth1/39.14
Nexthop eth1/40.15-100.1.1.14				v4			100.1.1.14	eth1/40.15
Route 192.168.3.0/24, Flags:v4, Unicast Cost: 20	192.168.3.0/24	nssa2	0.0.0.100	v4	20	20		
Route 2.2.2.2/32, Flags:in-rib,v4, Unicast Cost: 15	2.2.2.2/32	intra	0.0.0.100	in-rib,v4	15	15		
Route 20.20.20.20/32, Flags:in-rib,v4, Unicast Cost: 11	20.20.20.20/32	intra	0.0.0.100	in-rib,v4	11	11		
Nexthop eth1/39.14-100.1.1.10				v4			100.1.1.10	eth1/39.14
Nexthop eth1/40.15-100.1.1.14				v4			100.1.1.14	eth1/40.15
Route 3.3.3.3/32, Flags:direct,v4, Unicast Cost: 1	3.3.3.3/32	intra	0.0.0.100	direct,v4	1	1		

Next look at the routing table on the non-border leaf. Login to fab3-leaf1 (with node ID 101) and check the routing table for the private network, CTX1, for tenant pepsi. As expected, fab3-leaf1 learns all the external routes through MP-BGP. The APIC controller instantiates the required VRF for a tenant private network only when there are endpoints attached to the leaf. It is designed in such a way as to preserve the hardware resources on the leaf switches.

```
fab3-leaf1# show ip route vrf pepsi:CTX1
IP Route Table for VRF "pepsi:CTX1"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

0.0.0.0/0, ubest/mbest: 2/0
    *via 10.0.82.221%overlay-1, [200/1], 1d10h, bgp-100, internal, tag 100 (mpls-
vpn)
    *via 10.0.106.27%overlay-1, [200/1], 1d10h, bgp-100, internal, tag 100 (mpls-
vpn)
2.2.2.2/32, ubest/mbest: 1/0
    *via 10.0.82.221%overlay-1, [200/0], 1d10h, bgp-100, internal, tag 100 (mpls-
vpn)
3.3.3.3/32, ubest/mbest: 1/0
    *via 10.0.106.27%overlay-1, [200/0], 1d10h, bgp-100, internal, tag 100 (mpls-
vpn)
20.20.20.20/32, ubest/mbest: 2/0
    *via 10.0.82.221%overlay-1, [200/11], 1d10h, bgp-100, internal, tag 100
(mpls-vpn)
    *via 10.0.106.27%overlay-1, [200/11], 1d10h, bgp-100, internal, tag 100
(mpls-vpn)
100.1.1.0/30, ubest/mbest: 1/0
    *via 10.0.82.221%overlay-1, [200/0], 1d10h, bgp-100, internal, tag 100 (mpls-
vpn)
100.1.1.4/30, ubest/mbest: 1/0
    *via 10.0.82.221%overlay-1, [200/0], 1d10h, bgp-100, internal, tag 100 (mpls-
vpn)
100.1.1.8/30, ubest/mbest: 1/0
    *via 10.0.106.27%overlay-1, [200/0], 1d10h, bgp-100, internal, tag 100 (mpls-
vpn)
100.1.1.12/30, ubest/mbest: 1/0
    *via 10.0.106.27%overlay-1, [200/0], 1d10h, bgp-100, internal, tag 100 (mpls-
vpn)
100.1.1.16/30, ubest/mbest: 2/0
    *via 10.0.82.221%overlay-1, [200/14], 1d10h, bgp-100, internal, tag 100
(mpls-vpn)
    *via 10.0.106.27%overlay-1, [200/14], 1d10h, bgp-100, internal, tag 100
(mpls-vpn)
192.168.1.0/24, ubest/mbest: 1/0, attached, direct, pervasive
    *via 10.0.47.224%overlay-1, [1/0], 1d10h, static
192.168.1.1/32, ubest/mbest: 1/0, attached
```

```
    *via 192.168.1.1, Vlan14, [1/0], 1d10h, local
192.168.2.0/24, ubest/mbest: 1/0, attached, direct, pervasive
    *via 10.0.47.224%overlay-1, [1/0], 1d10h, static
192.168.2.1/32, ubest/mbest: 1/0, attached
    *via 192.168.2.1, Vlan14, [1/0], 1d10h, local
192.168.3.0/24, ubest/mbest: 1/0, attached, direct, pervasive
    *via 10.0.47.224%overlay-1, [1/0], 1d10h, static
192.168.3.1/32, ubest/mbest: 1/0, attached
    *via 192.168.3.1, Vlan14, [1/0], 1d10h, local
fab3-leaf1#
```

IBGP Routing Protocol Peering between the ACI and External Router

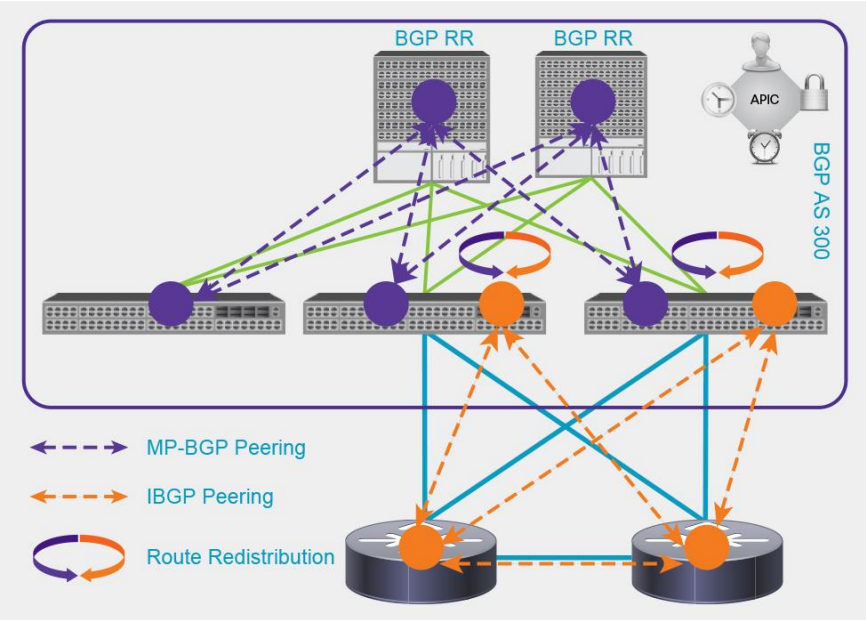
Besides static routes and OSPF customers also have the choice to run the BGP protocol between the ACI border leaf switches and external routers. As of this writing, the ACI border leaf switches supports IBGP only with eBGP support planned for future software release.

BGP AS Number

The ACI fabric supports one AS number. The same AS number is used for internal MP-BGP as well as for the iBGP session between the border leaf switches and external routers. The BGP AS number and RR need to be configured in order to enable MP-BGP within the ACI fabric (refer to the “Route Distribution Within ACI Fabric” section for details). Without MP-BGP, the external routes (static, OSPF, or BGP) for the layer 3 outside connections will not be propagated within the ACI fabric, and the ACI leaves that are not part of the border leaf will not have IP connectivity to an outside networks. Given that the same ASN is used for both cases and currently only IBGP is supported, the user needs to find out the ASN on the router that the ACI border leaf will connect to and use it as the BGP ASN for the ACI fabric.

Figure 38 depicts the BGP peering relationship on the ACI border leaf. The figure highlights that the same BGP ASN is being used for MP-BGP and the iBGP with the external router. Additionally, it also implies that the spine node only has MP-BGP sessions with the leaf. It doesn’t have BGP sessions with the external router and it can’t be used as BGP RR for that purpose.

Figure 38. BGP Peering Relationship on the ACI Border Leaf



BGP Route Policy

At the time of this writing, border leaf switches accept all the BGP route updates from its iBGP peer without applying any inbound route policy. The best routes will then be redistributed to the MP-BGP for the given private network (VRF) that the BGP session belongs to. MP-BGP subsequently distributes these routes to ACI leaf switches where the private network is instantiated. When both BGP and OSPF are deployed between the border leaf and external routers only routes learned through BGP will be injected into MP-BGP, and only BGP routes are distributed within the ACI fabric. Similarly, tenant routes are injected into BGP only when both BGP and OSPF are deployed.

Note: BGP routes take precedence when both OSPF and BGP are enabled between the ACI border leaf switches and external routers.

The ACI border leaf switches support outbound BGP policy to set community or extended community values for tenant routes. The BGP community and extended community attributes are commonly used by network architects to group together certain BGP routes and apply route policy by matching community values instead of other attributes (such as prefix or BGP AS number).

BGP Peering Consideration

iBGP design best practices need to be followed for the iBGP deployment between the ACI border leaf switches and external routers. The ACI border leaf needs to have IBGP sessions with all BGP speakers within the AS. In cases where the route reflector technology is deployed, ACI border leaf switches need to have IBGP sessions with all route reflectors in the BGP RR cluster.

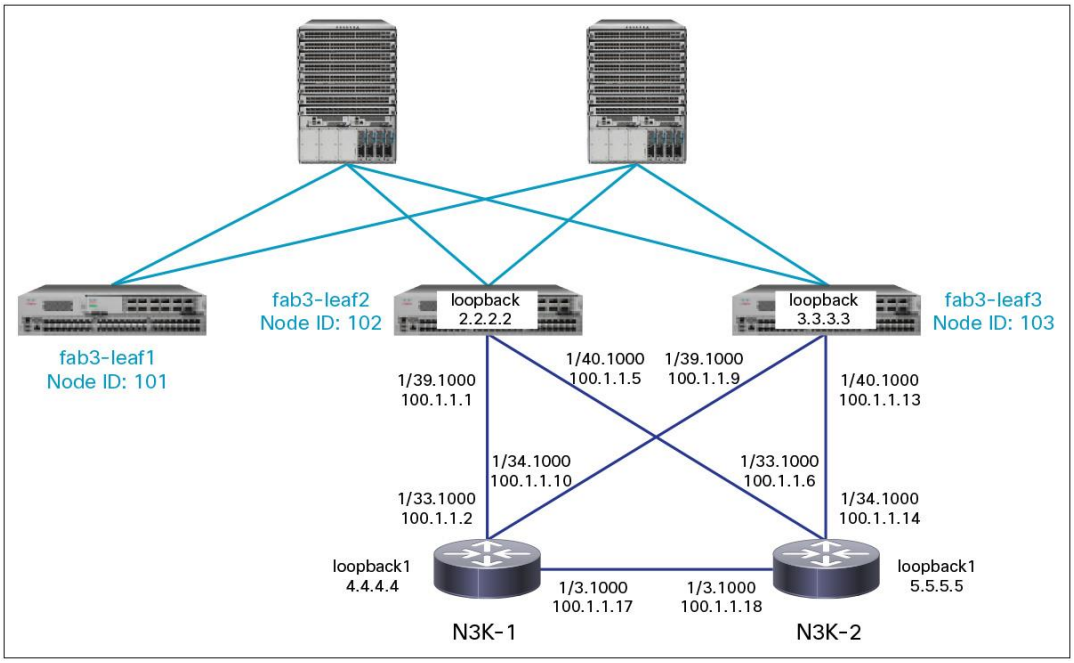
Notice that border leaves don't have iBGP sessions among themselves. This is not required because border leaf switches can learn routes from each other through MP-BGP.

At the time of this writing please follow the VRF-lite best practices for the multi-tenant deployment scenarios. When the layer 3 outside connection is required for each tenant, configure separate iBGP sessions for each tenant.

BGP Deployment Example

This section explains how to create a layer 3 outside connection with iBGP by using one example (Figure 39). This section uses the same topology as the one used in the section, OSPF Deployment Example.

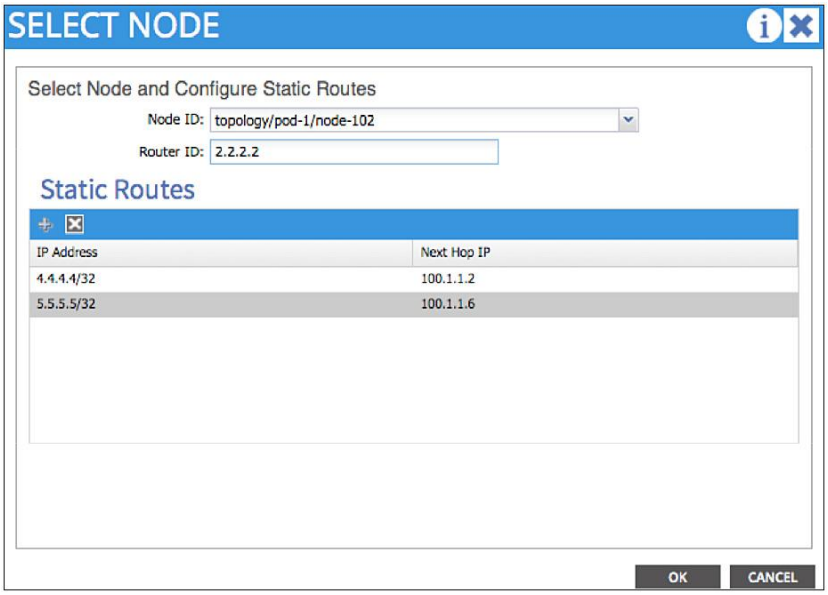
Figure 39. Topology for Layer 3 Outside Connection with IBGP Example



Creating a layer 3 outside connection with IBGP shares many common steps with configuring a layer 3 outside connection with OSPF. Each step below provides a brief description on how to create to create a layer 3 outside connection with IBGP, without actual screen captures.

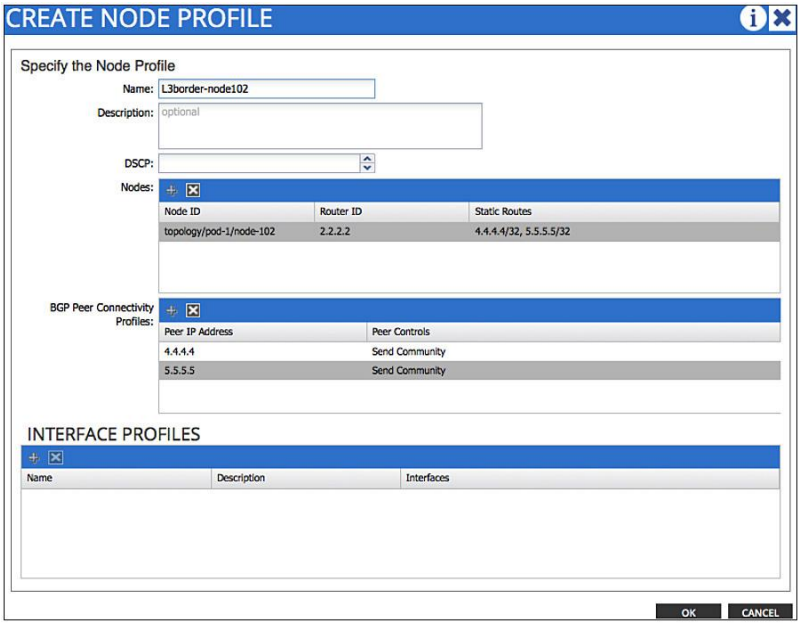
1. To start, create a layer 3 outside connection for the tenant. Choose BGP as the protocol. Associate it with a private network.
2. Click the “+” sign at the “Nodes and Interfaces Protocol Profiles” to add border leaf nodes. Start with border leaf node 102. This step specifies the router ID for this border leaf. Note that this IP address is also used as the source IP address for the BGP connection. As a result, it should be used as the BGP peer address on the external router. In this example, we assume OSPF is not deployed. The iBGP peer addresses of the Nexus 3000s are the loopback addresses and are not directly reachable. Therefore we also need to configure two static routes for the BGP peer address of the two Nexus 3000 switches, which are 4.4.4.4 and 5.5.5.5 (Figure 40).

Figure 40. Add Border Leaf and Static Routes for L3 Outside Connection



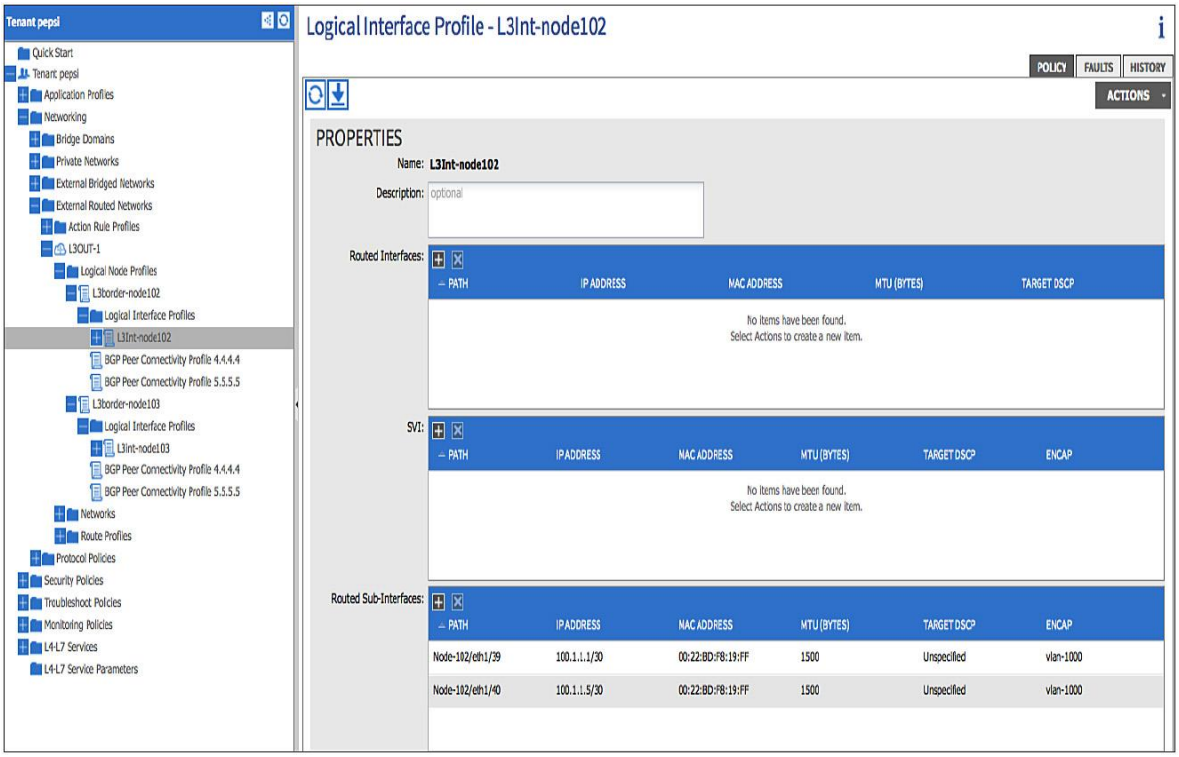
3. Add BGP peers 4.4.4.4 and 5.5.5.5 (the two Nexus switches). Figure 41 provides an example.

Figure 41. Adding BGP Peers



4. Click the “+” sign under Interface Profiles to add two sub-interfaces - eth1/39 and eth1/40 (with dot1q tag 1000). This step is same as the one explained in the OSPF deployment example section.
5. Repeat steps 3 and 4 for border leaf node 103. Figure 42 shows how the configuration looks like after adding two border leaves, BGP peers, and sub-interfaces.

Figure 42. Adding Sub-Interfaces to L3 Outside Connection



6. Create the EPG for the layer 3 outside connection and its related contract.
7. Associate the layer 3 outside connection with the bridge domain “BD1” for tenant pepsi. (This is same procedure as the one explained for OSPF deployment).

Now complete the configuration for the layer 3 outside connection with IBGP. Keep in mind that the BGP AS number and BGP route reflector need to be configured in order to enable MP-BGP within the ACI fabric. Without it the external routes won’t be distributed to non-border leaves.

Next, look at the configuration on the Nexus 3000 Switch. On the Nexus 3000 configure three IBGP peers - two for ACI border leaves and the third for its peer Nexus 3000. Static routes are configured for IBGP peer IP addresses.

```
vrf context pepsi
  ip route 2.2.2.2/32 100.1.1.1
  ip route 3.3.3.3/32 100.1.1.9
  ip route 5.5.5.5/32 100.1.1.18

interface Ethernet1/3.1000
  encapsulation dot1q 1000
  vrf member pepsi
  ip address 100.1.1.17/30

interface Ethernet1/33
  description to leaf102 eth1/39
  no switchport
```

```
interface Ethernet1/33.1000
  encapsulation dot1q 1000
  vrf member pepsi
  ip address 100.1.1.2/30

interface Ethernet1/34
  description to leaf103 eth1/39
  no switchport

interface Ethernet1/34.1000
  encapsulation dot1q 1000
  vrf member pepsi
  ip address 100.1.1.10/30


interface loopback0
  vrf member pepsi
  ip address 20.20.20.20/32

interface loopback1
  vrf member pepsi
  ip address 4.4.4.4/32
router bgp 100
  vrf pepsi
    router-id 4.4.4.4
    address-family ipv4 unicast
      maximum-paths 8
      maximum-paths ibgp 8
    neighbor 2.2.2.2 remote-as 100
      update-source loopback1
      address-family ipv4 unicast
        default-originate
    neighbor 3.3.3.3 remote-as 100
      update-source loopback1
      address-family ipv4 unicast
        default-originate
    neighbor 5.5.5.5 remote-as 100
      update-source loopback1
      address-family ipv4 unicast
```

Check the BGP sessions and the routing table.

```
N3K-1# sh ip bgp summary vrf pepsi
BGP summary information for VRF pepsi, address family IPv4 Unicast
BGP router identifier 4.4.4.4, local AS number 100
```

```
BGP table version is 22, IPv4 Unicast config peers 3, capable peers 3
2 network entries and 4 paths using 320 bytes of memory
BGP attribute entries [1/136], BGP AS path entries [0/0]
BGP community entries [0/0], BGP clusterlist entries [0/0]

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
2.2.2.2         4   100      97      87       22    0    0 00:04:31 2
3.3.3.3         4   100      75      76       22    0    0 00:04:31 2
5.5.5.5         4   100      17      16       22    0    0 00:04:20 0
N3K-1# sh ip bgp vrf pepsi
BGP routing table information for VRF pepsi, address family IPv4 Unicast
BGP table version is 22, local router ID is 4.4.4.4
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath

   Network            Next Hop            Metric     LocPrf     Weight Path
*|i192.168.2.0/24     3.3.3.3                  0        100          0 ?
*>i                   2.2.2.2                  0        100          0 ?
*|i192.168.3.0/24     3.3.3.3                  0        100          0 ?
*>i                   2.2.2.2                  0        100          0 ?
```

The routing table proves that the border leaves advertise two public subnets of the bridge domain.

```
N3K-1# sh ip ro vrf pepsi
IP Route Table for VRF "pepsi"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

2.2.2.2/32, ubest/mbest: 1/0
    *via 100.1.1.1, [1/0], 01:21:11, static
3.3.3.3/32, ubest/mbest: 1/0
    *via 100.1.1.9, [1/0], 01:06:10, static
4.4.4.4/32, ubest/mbest: 2/0, attached
    *via 4.4.4.4, Lo1, [0/0], 01:52:09, local
    *via 4.4.4.4, Lo1, [0/0], 01:52:09, direct
5.5.5.5/32, ubest/mbest: 1/0
    *via 100.1.1.18, [1/0], 00:07:51, static
20.20.20.20/32, ubest/mbest: 2/0, attached
    *via 20.20.20.20, Lo0, [0/0], 1d18h, local
    *via 20.20.20.20, Lo0, [0/0], 1d18h, direct
100.1.1.0/30, ubest/mbest: 1/0, attached
    *via 100.1.1.2, Eth1/33.1000, [0/0], 01:21:20, direct
100.1.1.2/32, ubest/mbest: 1/0, attached
```

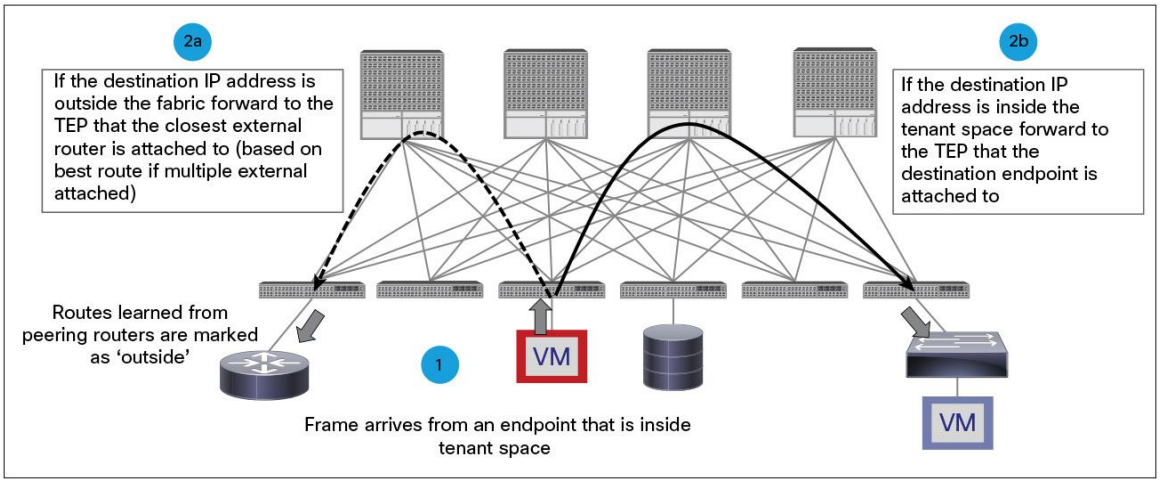
```
    *via 100.1.1.2, Eth1/33.1000, [0/0], 01:21:20, local
100.1.1.8/30, ubest/mbest: 1/0, attached
    *via 100.1.1.10, Eth1/34.1000, [0/0], 01:21:21, direct
100.1.1.10/32, ubest/mbest: 1/0, attached
    *via 100.1.1.10, Eth1/34.1000, [0/0], 01:21:21, local
100.1.1.16/30, ubest/mbest: 1/0, attached
    *via 100.1.1.17, Eth1/3.1000, [0/0], 1d18h, direct
100.1.1.17/32, ubest/mbest: 1/0, attached
    *via 100.1.1.17, Eth1/3.1000, [0/0], 1d18h, local
192.168.2.0/24, ubest/mbest: 2/0
    *via 2.2.2.2, [200/0], 00:04:43, bgp-100, internal, tag 100
    *via 3.3.3.3, [200/0], 00:04:43, bgp-100, internal, tag 100
192.168.3.0/24, ubest/mbest: 2/0
    *via 2.2.2.2, [200/0], 00:04:43, bgp-100, internal, tag 100
    *via 3.3.3.3, [200/0], 00:04:43, bgp-100, internal, tag 100
N3K-1#  ping 192.168.3.1 vrf pepsi
PING 192.168.3.1 (192.168.3.1): 56 data bytes
64 bytes from 192.168.3.1: icmp_seq=0 ttl=57 time=0.98 ms
64 bytes from 192.168.3.1: icmp_seq=1 ttl=57 time=0.635 ms
64 bytes from 192.168.3.1: icmp_seq=2 ttl=57 time=0.61 ms
64 bytes from 192.168.3.1: icmp_seq=3 ttl=57 time=1.251 ms
64 bytes from 192.168.3.1: icmp_seq=4 ttl=57 time=0.624 ms

--- 192.168.3.1 ping statistics ---
5 packets transmitted, 5 packets received, 0.00% packet loss
round-trip min/avg/max = 0.61/0.819/1.251 ms
N3K-1# ping 192.168.2.1 vrf pepsi
PING 192.168.2.1 (192.168.2.1): 56 data bytes
64 bytes from 192.168.2.1: icmp_seq=0 ttl=57 time=0.982 ms
64 bytes from 192.168.2.1: icmp_seq=1 ttl=57 time=0.691 ms
64 bytes from 192.168.2.1: icmp_seq=2 ttl=57 time=0.644 ms
64 bytes from 192.168.2.1: icmp_seq=3 ttl=57 time=0.675 ms
64 bytes from 192.168.2.1: icmp_seq=4 ttl=57 time=3.054 ms

--- 192.168.2.1 ping statistics ---
5 packets transmitted, 5 packets received, 0.00% packet loss
round-trip min/avg/max = 0.644/1.209/3.054 ms
N3K-1#
```

Forwarding and Policy Model with ACI Layer 3 Outside Connection

Inside and Outside

When a leaf switch receives a frame from the host it needs to determine whether the destination IP is inside the fabric or outside the fabric. If the destination IP matches with any /32 host route entry in the global station table, it means the destination is an endpoint inside the fabric and the leave switch already learned the endpoint. If the destination IP doesn't match with any /32 host route entry, the global station table leaf switch will check if the destination IP is within the IP address range of the tenant. If the address is within range, then the destination IP is inside of the fabric but the leaf switch hasn't yet learned the destination IP. The leaf switch then encapsulates the frame to VXLAN frame format with the spine proxy IP as the destination IP of the VXLAN outer IP header. The spine proxy checks the inner destination IP against its proxy database and finds the egress leaf switch IP and forward frame to the egress leaf. When the destination of the packet is outside of the fabric, it will match with one of the routes in the external routing table. The external routing table provides the VTEP (VXLAN Tunnel End Point) address of the border leaf (Figure 43).

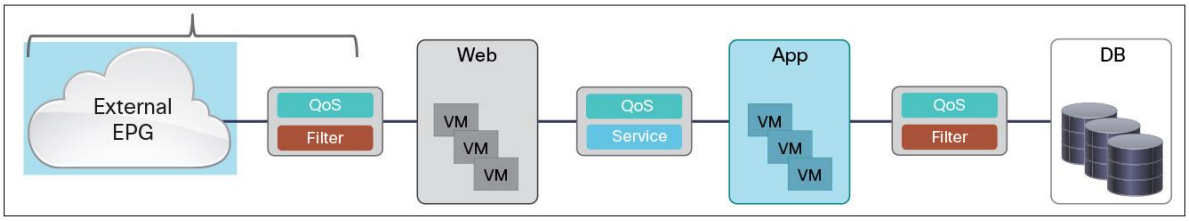
Figure 43. Forwarding to Inside End Points and Outside End Points



External EPG and Policy Model

The ACI fabric implements group-based policy. The endpoints that share the same policy requirement are assigned to an EPG. Instead of applying policy based on the IP address or applying policy on the per-endpoint basis, the group-based policy model applies the policy for the communication between groups. The same policy model is extended to include the communication between internal endpoints and external endpoints (endpoints that are outside of fabric). The external endpoints are assigned to an external EPG. For the layer 3 outside connections, the external endpoints are mapped to an external EPG based on IP prefixes. Figure 44 provides an example of an application profile, of three-tier applications, that includes one external EPG and its associated contracts.

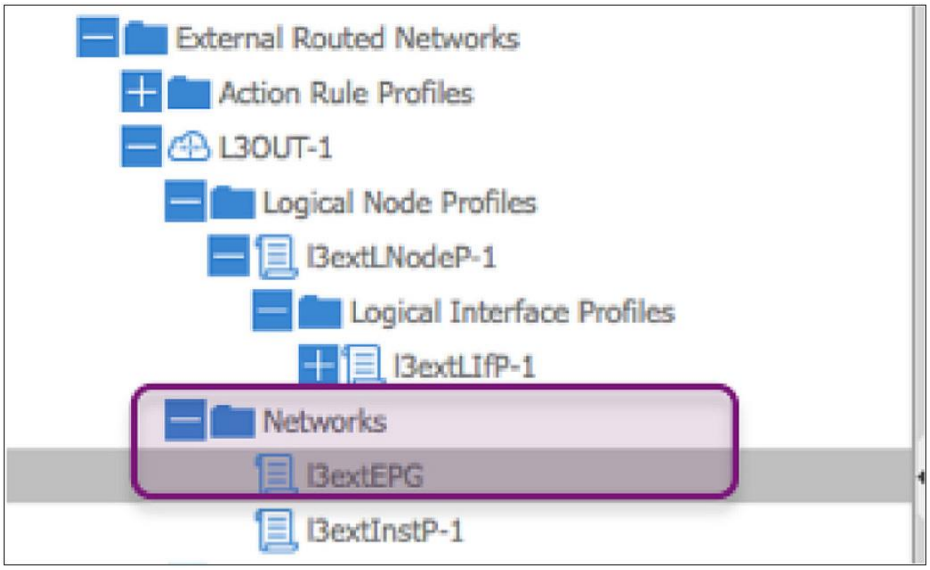
Figure 44. Application Profile Example with External EPG



For each layer 3 outside connection, the user has the option to create one or multiple external EPGs based on whether they need different policy treatments for different groups of external endpoints.

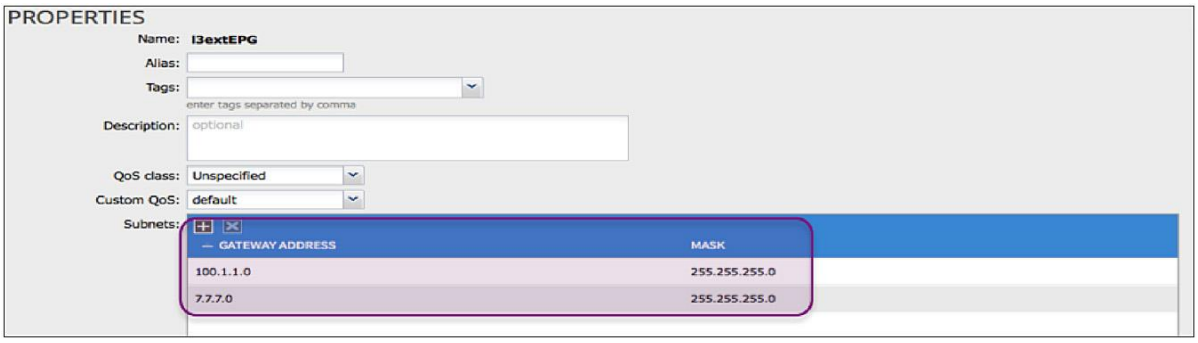
The external EPG configuration menu on the APIC GUI is the “Networks” menu under the layer 3 outside connection created by the user (Figure 45). In this example, it is the layer 3 outside connection named “L3OUT-1”.

Figure 45. External EPG Configuration Menu



Under the layer 3 external EPG configuration, the user can map external endpoints to this EPG by adding IP prefixes and network masks. The network prefix and mask don’t need to be the same as the ones in the routing table. When only one external EPG is required, simply use 0.0.0.0/0 to assign all external endpoints to this external EPG.

Figure 46. Adding Prefix and Mask to External EPG

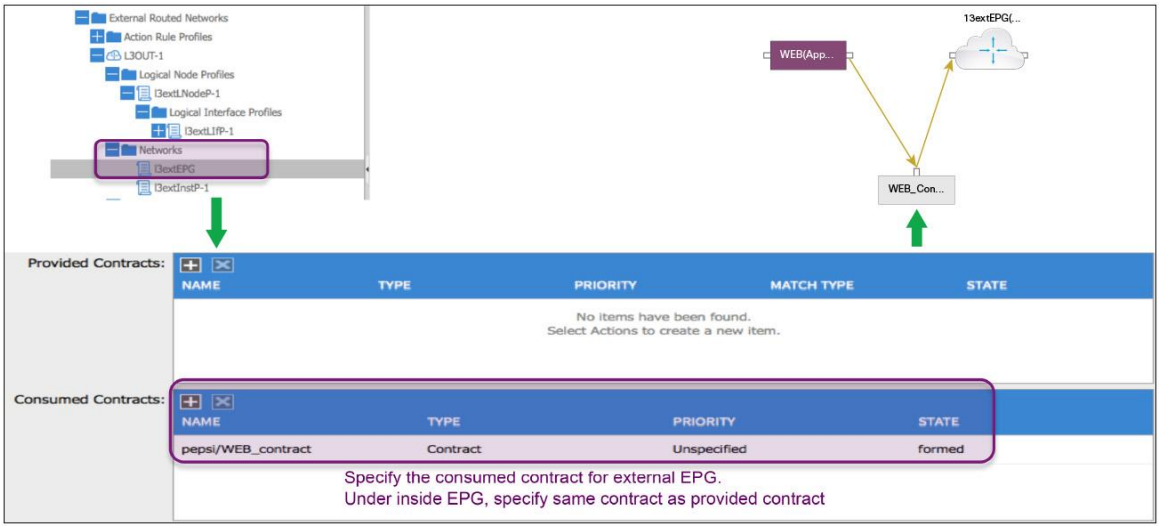


When mapping IP prefixes to an external EPG, one important thing to remember is to make sure the IP address ranges specified by the IP prefixes for the external EPG don't include the tenant IP address space. Otherwise, the fabric endpoints that are attached to the border leaf may be assigned to the external EPG. For example, if the tenant is assigned subnet 100.1.1.0/24 and the user has the IP prefix 100.1.0.0/16 in the external EPG configuration, if there is an endpoint with IP 100.1.1.10 attached to the border leaf, this endpoint may be classified to the external EPG. As a result, the wrong policy will be applied for the traffic related to this endpoint. The user can use 0.0.0.0/0 to define the external EPG. In such a case the border leaf will derive the external EPG based on the incoming interface. As a result, it won't lead to a policy issue even though 0.0.0.0/0 does overlap with the tenant IP address space.

Note: When using an IP prefix other than 0.0.0.0/0 to define the external EPG, make sure those IP prefixes and masks don't overlap with the tenant IP address space.

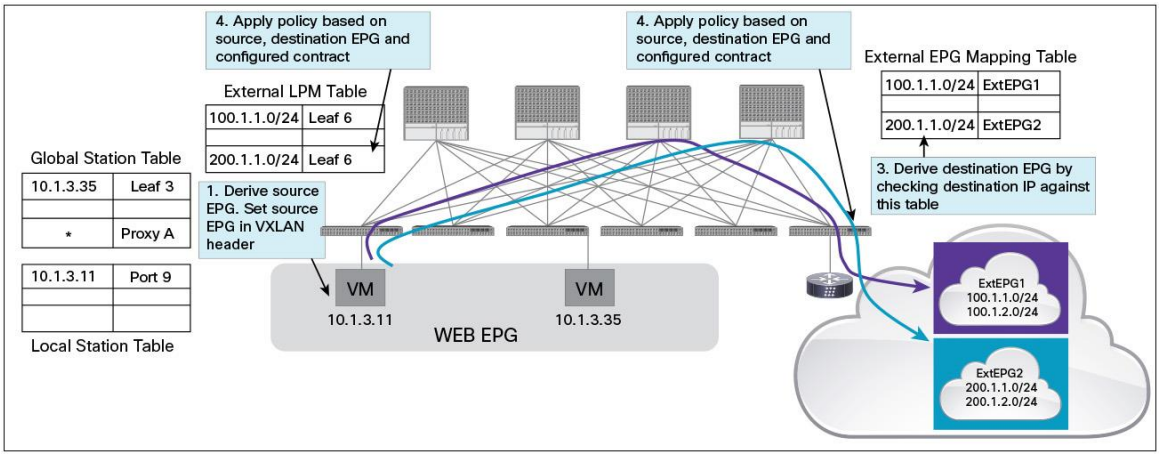
After creating the external EPG, the proper contract can be applied between the external EPG and other EPG as shown in Figure 46. By specifying the contract "WEB_contract" as the consumed contract for the external EPG "l3extEPG", and the provided contract for EPG "WEB" the contract "WEB_contract" is inserted between "WEB" EPG and the external EPG (as shown in the upper right corner of Figure 47). The traffic flow between "WEB" EPG and "l3extEPG" will be enforced by the policy defined in the contract "WEB_contract".

Figure 47. Applying the Proper Contract



Following is an example to explain how the data forwarding and policy enforcement work for the traffic flow between the internal EPG and external EPG. In this example, the internal EPG is named “WEB” and there are two external EPGs: ExtEPG1 and ExtEPG2. Hosts with the IP addresses of 100.1.1.0/24 and 100.1.2.0/24 are mapped to ExtEPG1. Those with the IP addresses of 200.1.1.0/24 and 200.1.2.0/24 are mapped to ExtEPG2.

Figure 48. Data Forwarding and Policy Enforcement for the Direction of Inside to Outside End Points

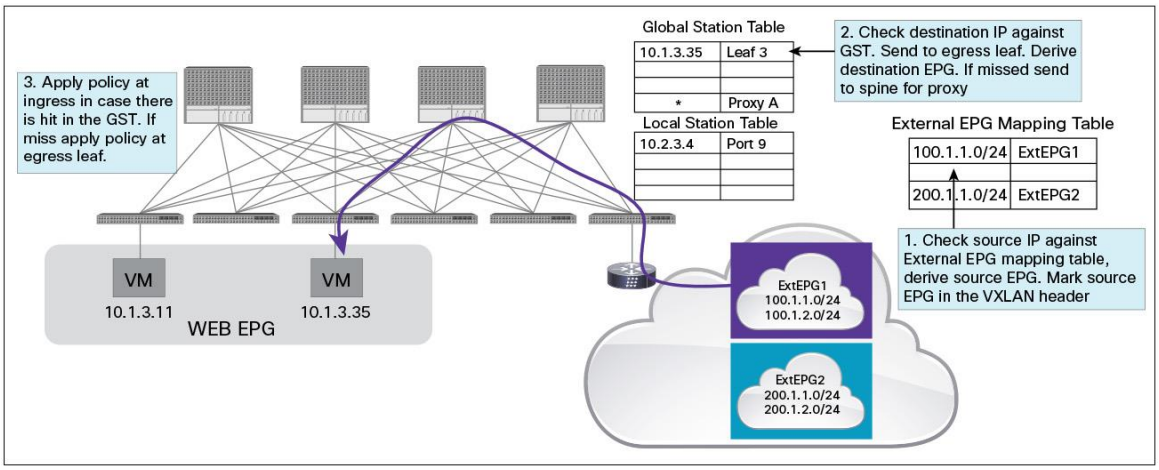


Following are actions taken for the traffic from the internal EPG to external EPG:

1. When the ingress leaf switch receives the frame, it learns the source MAC and source IP and programs them into the local station table. The leaf switch derives the source EPG based on the VLAN ID or VXLAN VNID. The MAC and IP addresses in the local station table also contain the EPG information and they can be used to derive EPG information for the subsequent packets.
2. The ingress leaf switch checks the destination IP against the external LPM table. The external LPM table stores the external summary routes learned from the border leaf. The matched entry provides the border leaf TEP IP address. The ingress leaf encapsulates the original frame in the VXLAN frame with the border leaf TEP IP address as the destination IP of the outer header. It also includes the source EPG information (identified in step 1) in the VXLAN header.
3. The VXLAN frame is forwarded by spine node to the border leaf switch. On the border leaf, the destination IP address of the original frame is checked against the “External EPG Mapping Table”. This table provides the IP prefix/mask to the external EPG mapping information. The lookup result provides the destination EPG of this frame.
4. With both source (carried in the VXLAN header) and destination EPG identified, the border leaf then applies the proper policy between the two EPGs. The border leaf processes the packets based on policy.

Let’s explore how the data forwarding and policy enforcement work for the reversed direction, which is the traffic flow from the external EPG to the internal EPG (Figure 49).

Figure 49. Data Forwarding and Policy Enforcement for the Direction of Outside to Inside End Points



1. The border leaf receives a frame destined for one of the internal endpoints. The border leaf checks the source IP address against the External EPG Mapping Table. The lookup result provides the source EPG of the frame.
2. The border leaf performs the lookup in the Global Station Table with the destination IP address. The Global Station Table provides cache entries for the remote endpoints (the endpoints attached to other leaf switches). If there is a hit for the lookup, the entry in the table provides the egress leaf TEP IP address, as well as the destination EPG information. In case the lookup doesn't match any entry in the table, the border leaf sends the frame to the spine by using the spine TEP IP address as the destination IP for the outer header. The spine switch does the lookup with the inner IP address and forwards the frame to the proper egress leaf switch. In this process, the source EPG identified in step 1 is carried in the VXLAN header.
3. In case there is a hit for the IP lookup against the global station table in step 2, the border leaf has both the source EPG and destination EPG, and it applies the proper contract configured for these two EPGs. The border leaf then encapsulates the frame in VXLAN format with the remote leaf TEP IP as the destination address for the outer header.
4. In case there is no hit for the IP lookup in step 2, the border leaf sends the frame to the spine and the spine then finds out the proper egress leaf by checking its hardware-mapping database. Because there is no destination EPG information, the border leaf can't apply the policy. When the frame is received on the egress leaf, the egress leaf checks the destination IP in the inner header against its local station table and identifies the egress interface as well as the destination EPG. With both source EPG (carried in the VXLAN header) and destination EPG available, the egress leaf then applies the policy configured for these two EPGs. The return traffic from the inside EPG to the outside EPG will have the border leaf learn the endpoint in the global station table. And all the subsequent frames traveling to this endpoint will have policy enforced at the border leaf.

ACI Layer 2 Connection to the Outside Network

In addition to the layer 3 outside connection to the outside network (described in previous sections), a network designer may need to extend a layer 2 domain beyond the ACI fabric. These circumstances may include connecting the existing layer 2 network to the ACI fabric, or extending the layer 2 domain to a data center infrastructure (DCI) platform (such as a Cisco Nexus 7000 Series Switch or Cisco ASR 9000 Series Router) that provides layer 2 DCI service to a remote site. Sometimes users simply need to assign a port to an EPG in order to connect a switch to the ACI fabric, or connect a hypervisor to the fabric.

There are several different ways to extend layer 2 domain beyond the ACI fabric:

- **Extend the EPG out of the ACI fabric** - A user can extend an EPG out of the ACI fabric by statically assigning a port (along with VLAN ID) to an EPG. The leaf will learn the endpoint information and assign the traffic (by matching the port and VLAN ID) to the proper EPG, and then enforce the policy. The endpoint learning, data forwarding, and policy enforcement remain the same whether the endpoint is directly attached to the leaf port or if it is behind a layer 2 network (provided the proper VLAN is enabled in the layer 2 network).
- **Extend the bridge domain out of the ACI fabric** - Another option to extend the layer 2 domain is to create a layer 2 outside connection (or external bridged network, as called in the APIC GUI) for a given bridge domain. It effectively extends the bridge domain to the outside network.
- **Extend the layer 2 domain with remote VTEP (future)** - In the previous two options the incoming traffic from outside is tagged with a VLAN ID. The ACI leaf classifies the traffic to the proper EPG by checking the port and VLAN ID. In future software releases, the remote VTEP will be supported, and can be used to extend the EPG or bridge domain.

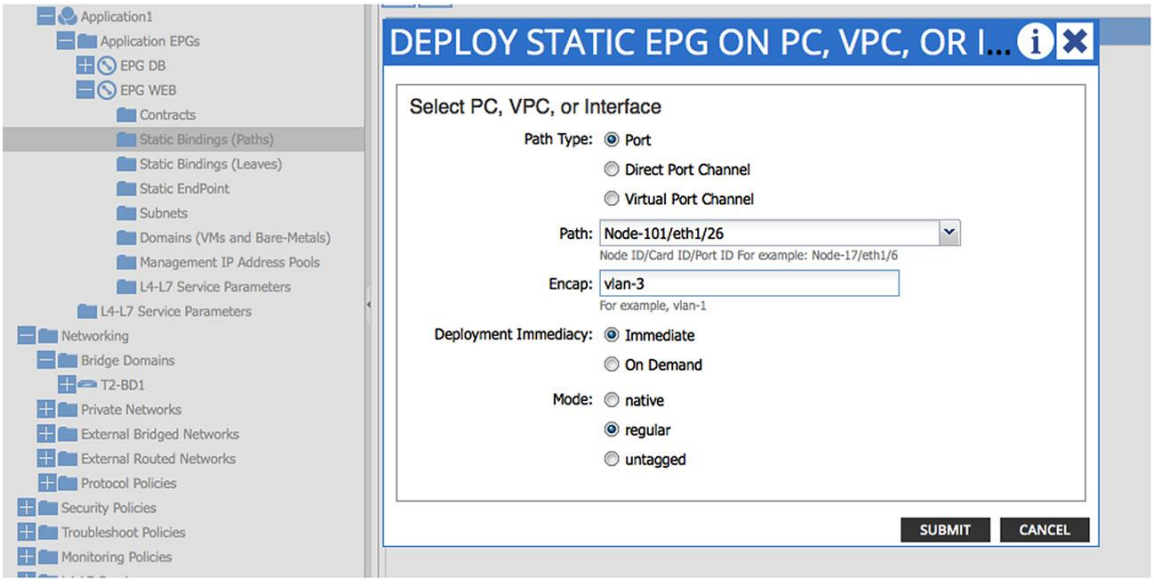
The following sections explain these three options in greater detail.

Extend the EPG Out of the ACI Fabric

The user can extend an EPG beyond an ACI leaf by statically assigning a leaf port (along with a VLAN ID) to an EPG. After doing so, all the traffic received on this leaf port with the configured VLAN ID will be mapped to the EPG and the configured policy for this EPG will be enforced. The endpoints need not be directly connected to the ACI leaf port. They can be behind a layer 2 network as long as the VLAN associated with the EPG is enabled within the layer 2 network that connects the remote endpoint to the ACI fabric.

To statically assign port to an EPG, go to menu Tenant→Application Profiles→EPG→Static Binding (Paths). Click the Action menu on the right side to start to assign port to an EPG. Figure 50 provides an example that assigns interface eth1/26 from the leaf node 101 along with VLAN 3 to EPG WEB.

Figure 50. Mapping VLAN Traffic Received on a Port to an EPG



In addition, to specify a port and VLAN ID, there are two configuration options:

- **Deployment immediacy** - Deployment immediacy determines when the actual configuration will be applied on the leaf switch hardware. It also determines when the hardware resource (such as VLAN resource and policy content-addressable memory [CAM] to support the related contract for this EPG) will be consumed on the leaf switch. The option “immediate” means the EPG configuration and its related policy configuration will be programmed in the hardware right away. The option “on demand” instructs the leaf switch to program the EPG and its related policy in the hardware only when the data frame is received for this EPG.
- **Mode** - This option specifies whether the ACI leaf expects incoming traffic to be tagged with a VLAN ID or not. The option “regular” means the leaf expects incoming traffic is tagged with the VLAN ID specified in this configuration task. The option “native” means the leaf switch expects to receive a frame without VLAN tagging. The meaning of “native” is same as the “native vlan” in legacy L2 switch. With the option “native” the same interface (physical port, Port-Channel or vPC) can carry traffic for other VLANs. In other words, you can statically assign the same interface(with different VLAN ID) to another EPG. On the outside L2 switch use the CLI “switchport native vlan **vlan_ID**”. The option “untagged” means the leaf switch expects to receive untagged traffic. Similar to the “switchport access vlan **vlan_ID**”, with this option you can only assign the interface to one EPG. This option can be used to connect a leaf port to a bare metal server whose network interface cards (NICs) typically generate untagged traffic.

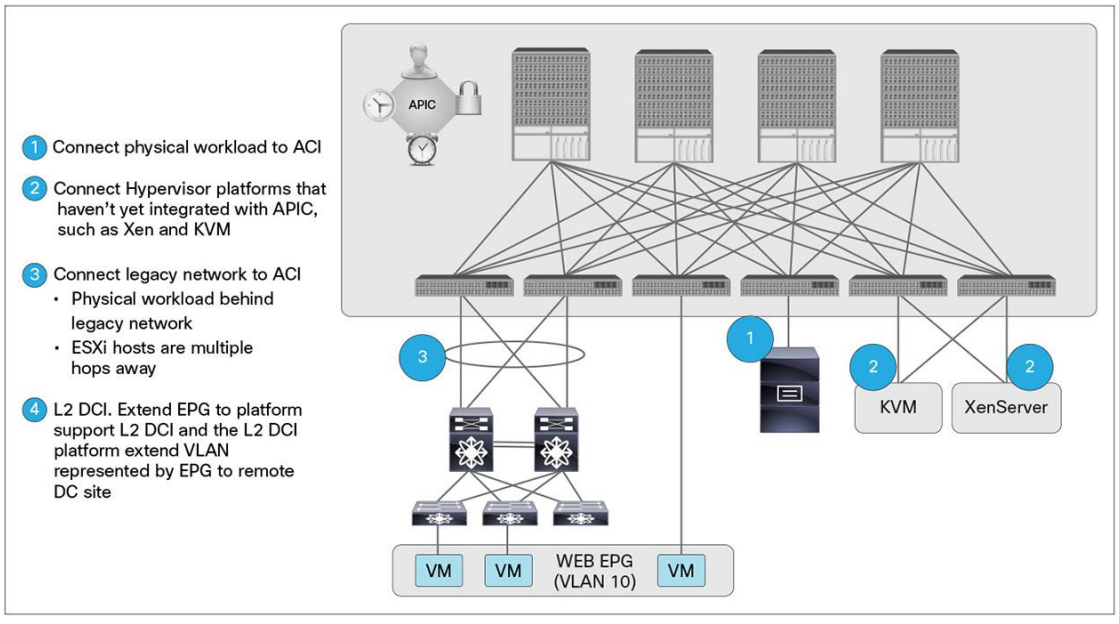
Besides the configuration tasks explained above, there are two additional steps required to finish the configuration.

- Create a physical domain and VLAN pool for this physical domain. Associate the physical domain with the EPG WEB
- Create an Attachable Access Entity Profile

Check step 6 in the section, **Extend Bridge Domain with Layer 2 Outside Connection Example**, for details on how to create a physical domain and attachable access entity profile.

Figure 51 summaries the use cases of extending the EPG out of the ACI fabric.

Figure 51. Use Cases of Extending EPG to Outside Network

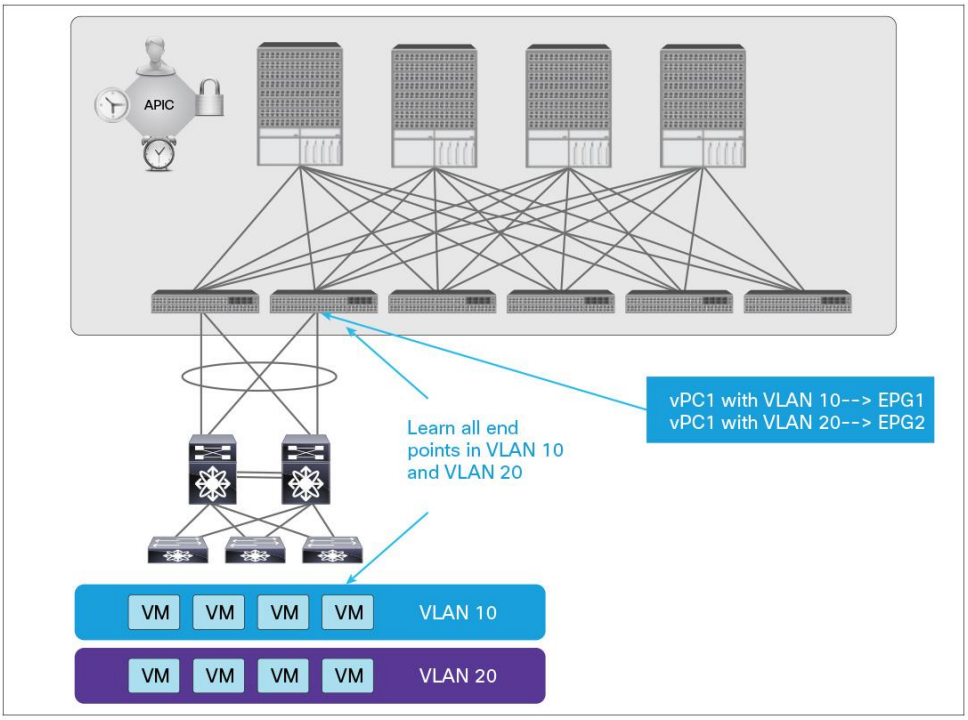


Each of the VLANs in the legacy network is mapped to an EPG in the ACI fabric. ACI leaves provide layer 3 forwarding between VLANs (cross-EPG), and the contract between EPGs is enforced.

Pay close attention to the endpoint count when connecting a large, layer 2 network to the ACI leaf. The ACI leaf ports that connect to the legacy network will learn all the endpoints behind the layer 2 network. When the routing is enabled for the bridge domain that the EPG belongs to, both the MAC and IP addresses of endpoints are learned. In case the routing is not enabled for the bridge domain, only the MAC address is learned. There is a 32,000-entry local station table (hardware table size, checking the software release document for supported endpoints per leaf) on every leaf node. Each MAC address and IPv4 address consumes one entry in the local station table, respectively. In the future, when IPv6 routing is supported by ACI software, the leaf port also learns the IPv6 address of an endpoint when IPv6 routing is enabled for the bridge domain. And each IPv6 address requires two hardware entries in the local station table.

In Figure 52, a layer 2 network consists of a Cisco Nexus switch connected to an ACI leaf through vPC1. On the ACI leaf, EPG1 and EPG2 are created to represent VLAN 10 and VLAN 20 in the layer 2 network. The two leaves learn all the MAC and IP addresses (assume routing is enabled for the bridge domain) of the hosts from VLAN 10 and VLAN 20.

Figure 52. Connecting ACI to L2 Network



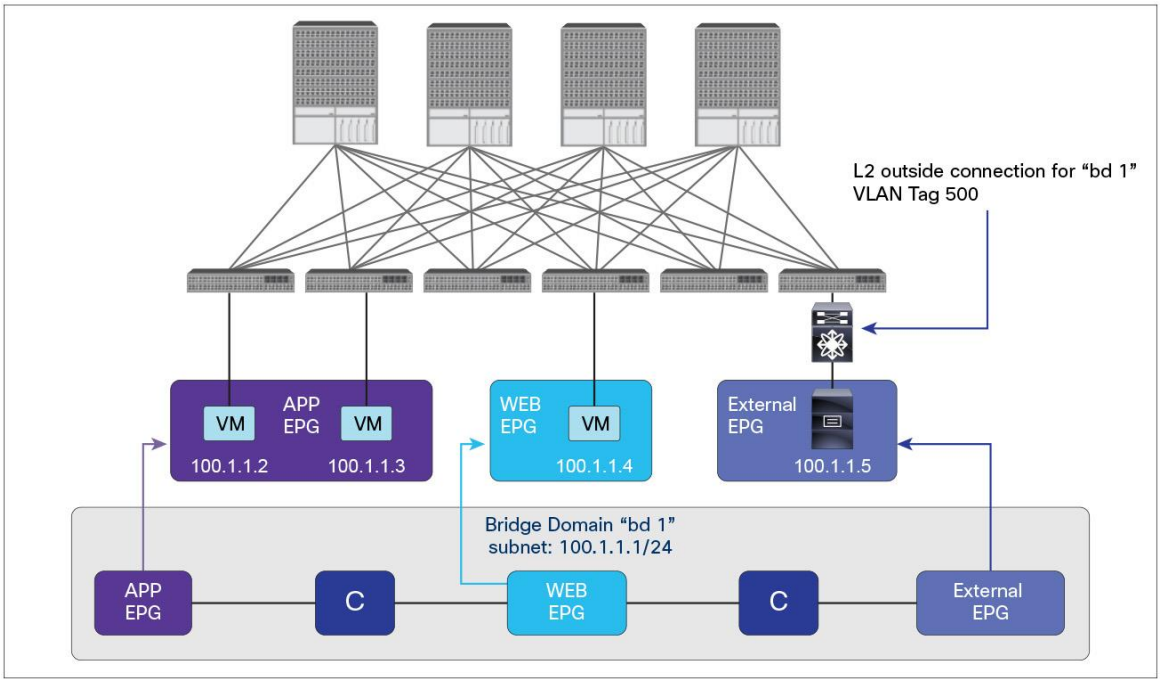
Although the size of the local station table is limited, the total amount of endpoints supported by the whole ACI fabric can be much larger than the size of the local station table. All the endpoints learned on the non-fabric uplinks are stored in the local station table. All the endpoints learned on the fabric uplink ports are kept in the global station table. The global station table is a cache table and each leaf only learns the remote endpoints with which its local endpoint has a conversation. When the remote endpoint is not the learned leaf switch, send the packet to the spine switch and the spine switch will search the endpoint in its inline hardware mapping database. It will then forward the frame to the remote leaf without penalty of performance and latency.

When the ACI fabric needs to connect to multiple layer 2 networks or layer 2 switches it is better to distribute them among multiple leaves to reduce the likelihood of filling up the local station table on the leaves.

Extend the Bridge Domain Out of the ACI Fabric

Another way to extend the layer 2 domain beyond the ACI fabric is to create layer 2 outside connections. On the APIC GUI it is under menu Tenant→Networking→External Bridged Networks (Figure 53). A layer 2 outside connection is associated with a bridge domain and it is designed to extend the whole bridge domain (not an individual EPG under bridge domain) to the outside network.

Figure 53. Extending the Bridge Domain Out of the ACI Fabric



In Figure 53, we create a layer 2 outside connection for a bridge domain called "bd1". There are two EPGs (EPG APP and EPG WEB) under this bridge domain. The layer 2 outside connection extends the bridge domain to the Cisco Nexus switch and the hosts attached to the switch. The layer 2 outside connection configuration specifies that the bridge domain "bd1" is extended to the network connected to the border leaf on the far right. All the traffic for the "bd1" carries the VLAN tag 500 when it leaves the ACI fabric. For the traffic entering the ACI fabric, the border leaf assigns all traffic with VLAN 500 to an external EPG. The traffic flows between the external EPG and other EPGs in the ACI fabric are enforced with the configured contract.

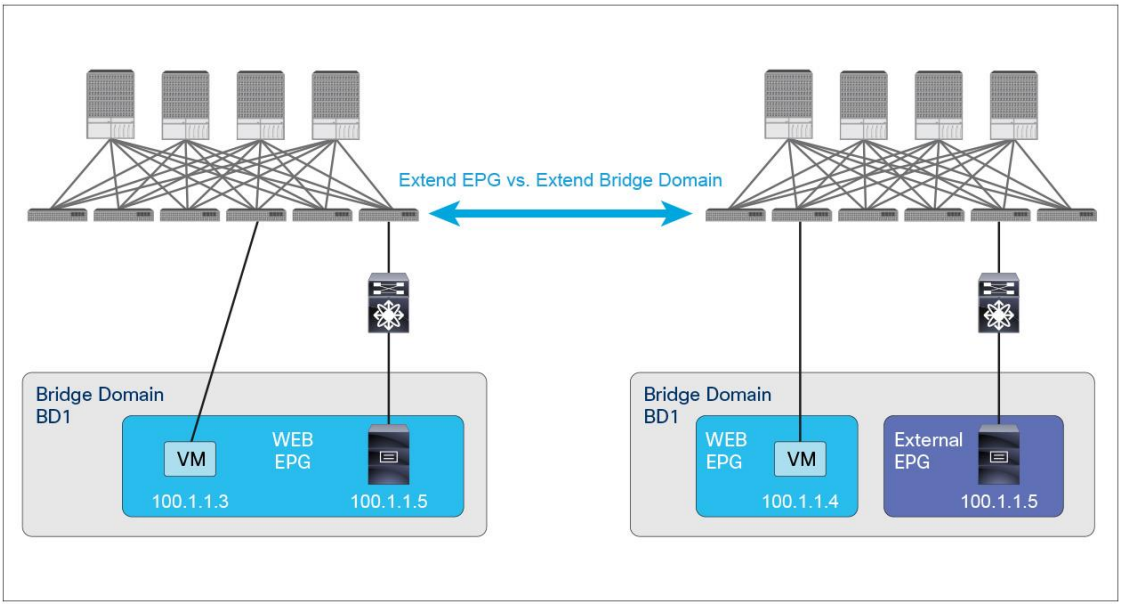
The bottom portion of Figure 53 illustrates the policy model with the layer 2 outside connection. Although the diagram only shows the contract between an external EPG with one inside EPG, there is no limitation about how many inside EPGs can talk to an external EPG. It is all driven by the policy configuration.

As shown in Figure 53, the layer 2 outside connection for a given bridge domain effectively extends the bridge domain beyond the ACI fabric. The endpoints in the outside network share all the characters and configurations for the bridge domain, such as the IP subnet assigned to the bridge domain and the default gateway. The endpoints in the outside network are also in the same flooding domain as the rest of the endpoints under the same bridge domain. Obviously, the ACI fabric can't control the flooding behavior for the outside network. In other words, even when a user disables the flooding of some unknown unicast under the bridge domain (which has been extended to the outside layer 2 network), the flooding of the unknown unicast still occurs on the outside network.

The ACI border leaf that connects to the outside layer 2 network learns the endpoint information. The learning behavior is the same as one explained in the section, "Extend EPG Out of the ACI Fabric."

On the surface, the layer 2 outside connection is similar to the way of extending an EPG by statically assigning a port plus VLAN tagging to an EPG. There are big differences between these two. Figure 52 explains the difference between these two methods by looking at the placement of outside endpoints and the policy model.

Figure 54. Differences between Layer 2 Outside Connection and Extending an EPG



On the left side of Figure 54, the user extends the EPG to the outside by assigning the port with VLAN to EPG WEB. In this case, the outside endpoint is directly placed in the WEB EPG. The same VLAN tagging used for the inside EPG is used for the outside EPG. Effectively, the same VLAN is stretched from the inside to the outside network.

On the right side, the user extends the bridge domain to the outside by creating a layer 2 outside connection for the bridge domain. An external EPG is created and placed under the bridge domain. The outside endpoint is assigned to the external EPG. It also implies that the outside endpoint comes from a different VLAN ID than the one used for the inside EPG under WEB EPG. However, the user has the choice to apply policy between WEB EPG and External EPG.

What they have in common is that in both cases, the outside endpoints are in the same bridge domain as the inside endpoints. Hence, they share same subnet and same default gateway. The learning behavior also remains the same between these two solutions. Table 1 summarizes the difference and common characteristics for these two.

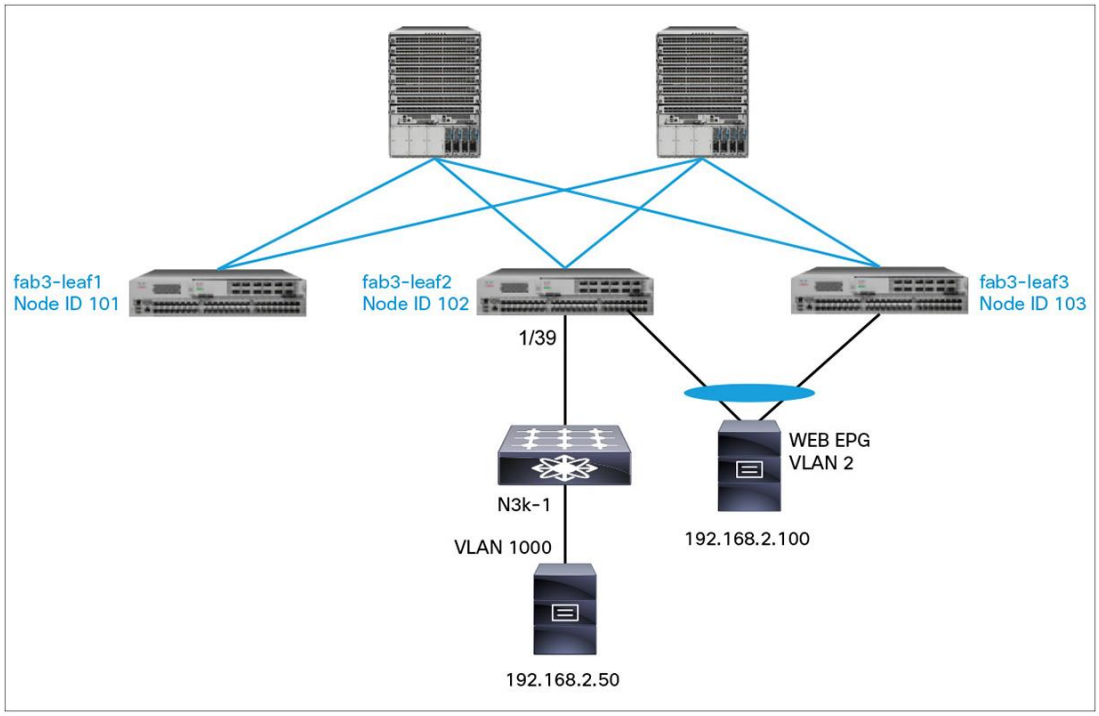
Table 1. Differences and Commonalities

	Extend EPG	Extend Bridge Domain
Use Cases	Extend EPG beyond ACI fabric; Extend VLAN represented by EPG out of ACI fabric	Extend bridge domain out of fabric. Extend tenant subnet of the bridge domain out of fabric
Configuration	Statically assign port to EPG (static binding under EPG)	Create external bridged networks (layer 2 outside connection)
External endpoints placement	In same EPG (VLAN) as directly attached endpoints	In different EPG (VLAN) but same bridge domain as directly attached endpoints
Policy model	Follow same policy as EPG	External endpoint is placed under external EPG; policy applied between internal and external EPG
Endpoint learning	Learning MAC and IP (with routing enabled for BD)	Learning MAC and IP (with routing enabled for BD)

Extend Bridge Domain with a Layer 2 Outside Connection Example

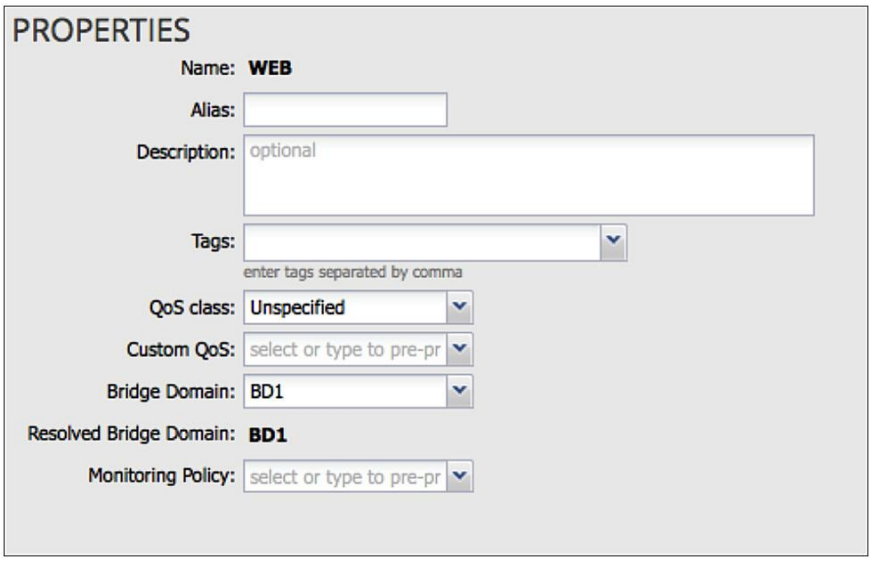
This section explores the steps to create a layer 2 outside connection for a bridge domain with an actual example (Figure 55).

Figure 55. Topology of Layer 2 Outside Connection Example



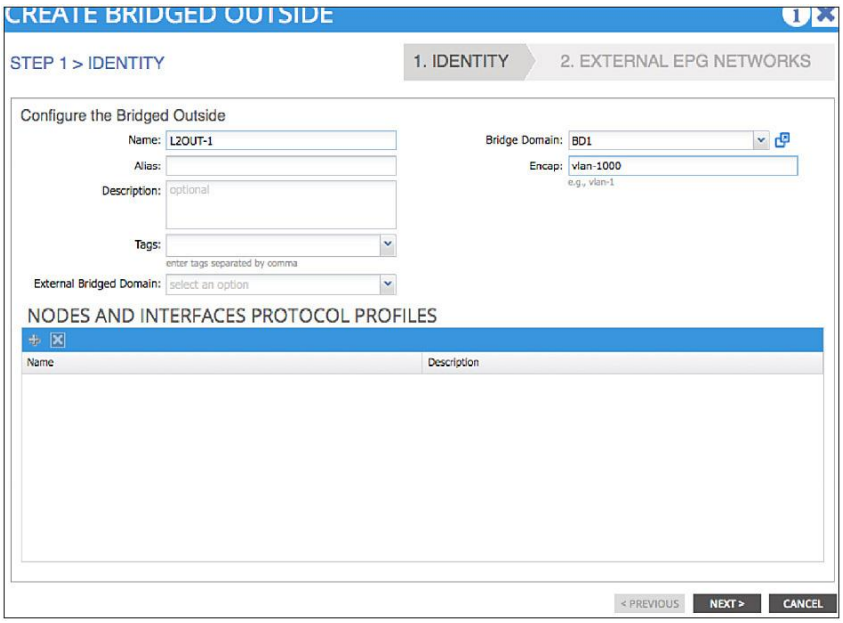
In Figure 55, the endpoint 192.168.2.100 is attached to the ACI leaf and the endpoint is in WEB EPG. And the WEB EPG is in the bridge domain, named “BD1”. Following are the steps to create a layer 2 outside connection for the bridge domain. With the layer 2 outside connection, it extends the bridge domain to the Nexus 3000-1 VLAN 1000 and places all hosts in VLAN 1000 in bridge domain “BD1”. Before the actual configuration, let’s check that WEB EPG is part of bridge domain “BD1”.

Figure 56. EPG and Bridge Domain Relationship



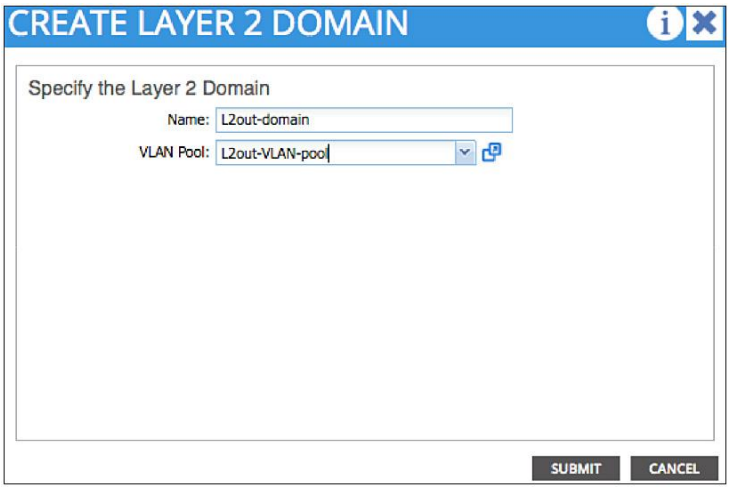
1. Start to create a layer 2 outside connection by going to menu Tenant→Networking→External Bridged Network. Click Action to start the process of adding a layer 2 outside connection. In Figure 54, associate the layer 2 outside connection with bridge domain “BD1”. Also choose a VLAN ID. This VLAN needs to be configured on the external layer 2 network. This layer 2 outside connection will put this VLAN and the “BD1” of the ACI fabric under the same layer 2 domain. Also, this VLAN ID has to be in the range of the VLAN pool used for the layer 2 outside connection. The VLAN pool will be configured in step 2.

Figure 57. Start to Create a Layer 2 Outside Connection



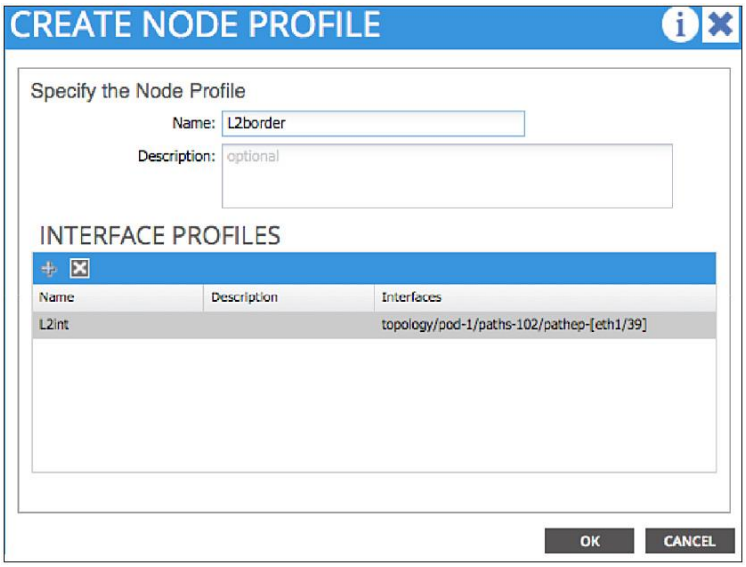
2. Click the drop-down menu of the “External Bridged Domain” to create a layer 2 domain named “L2out-domain”. Click the drop-down menu of the VLAN Pool to create a new VLAN Pool named “L2out-VLAN-pool” with a VLAN range from 1000 to 1200. This is essentially a mechanism to specify the range of the VLAN ID that will be used for creating a layer 2 outside connection (Figure 57). This helps avoid overlapping the VLAN range between VLANs used for an EPG and those for a layer 2 outside connection.

Figure 58. Create Layer 2 Domain and VLAN Pool



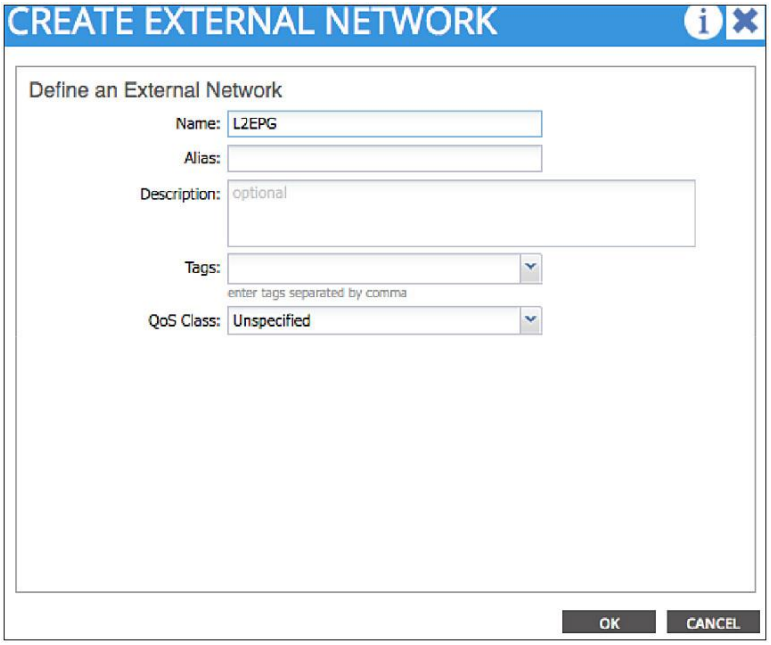
3. Add a layer 2 border leaf node and layer 2 interface for a layer 2 outside connection (Figure 59).

Figure 59. Add Border Leaf Node to Layer 2 Outside Connection



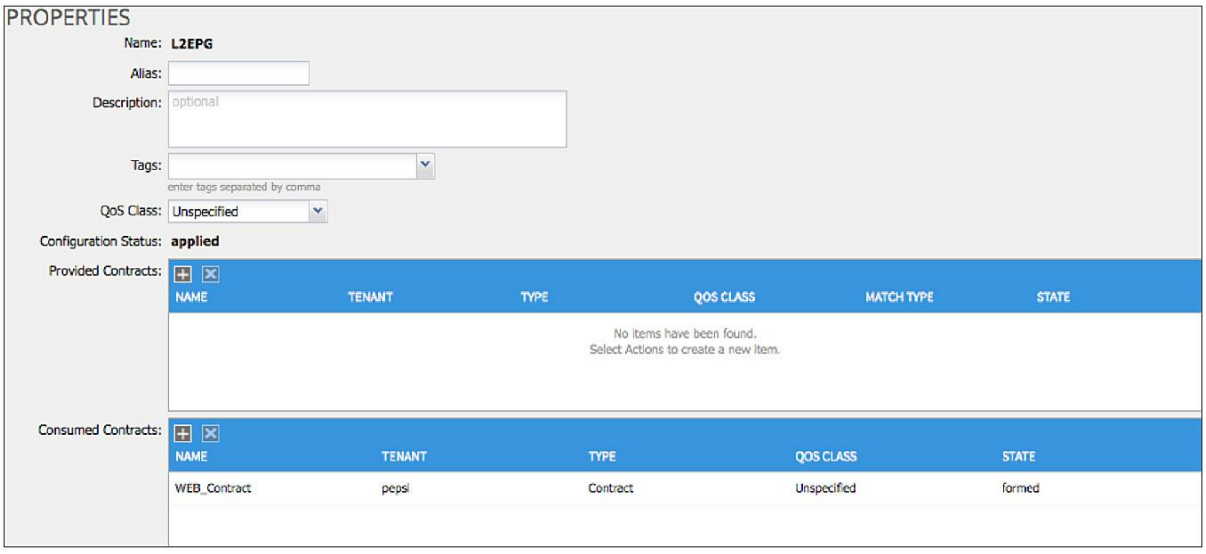
4. After adding a layer 2 border leaf and layer 2 interface, click “Next” to start creating a layer 2 EPG. Simply provide a name for the layer 2 EPG. All the traffic entering the ACI fabric with VLAN 1000 (the VLAN ID provided in step 1) will be classified into this layer 2 EPG.

Figure 60. Create External EPG for Layer 2 Outside Connection



5. Configure the contract between WEB EPG and the layer 2 external EPG. Go to menu External Bridged Networks→Networks to specify “WEB_contract” as the consumed contract. After specifying the “WEB_contract” as the provided contract for WEB EPG, the communication between this external layer 2 EPG and WEB EPG will be allowed.

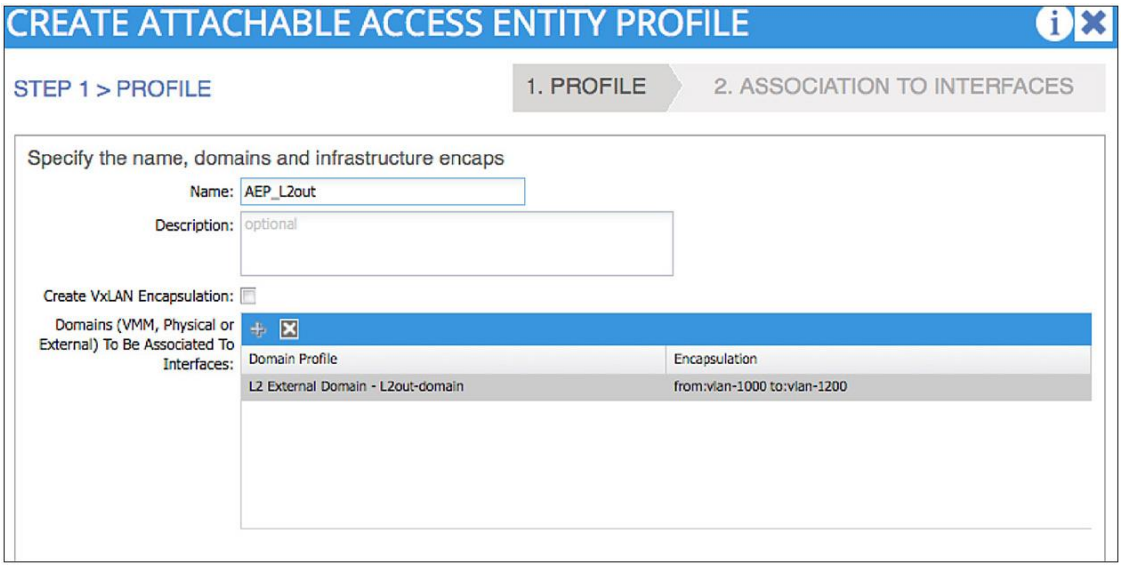
Figure 61. Specify Contract for External EPG



6. The last step of this configuration is to create the Access Attachable Entity Profile. On a high level, this is a mechanism to instruct the APIC to allow certain VLANs on selected ports. Follow these steps to create the attachable entity profile on the APIC GUI.

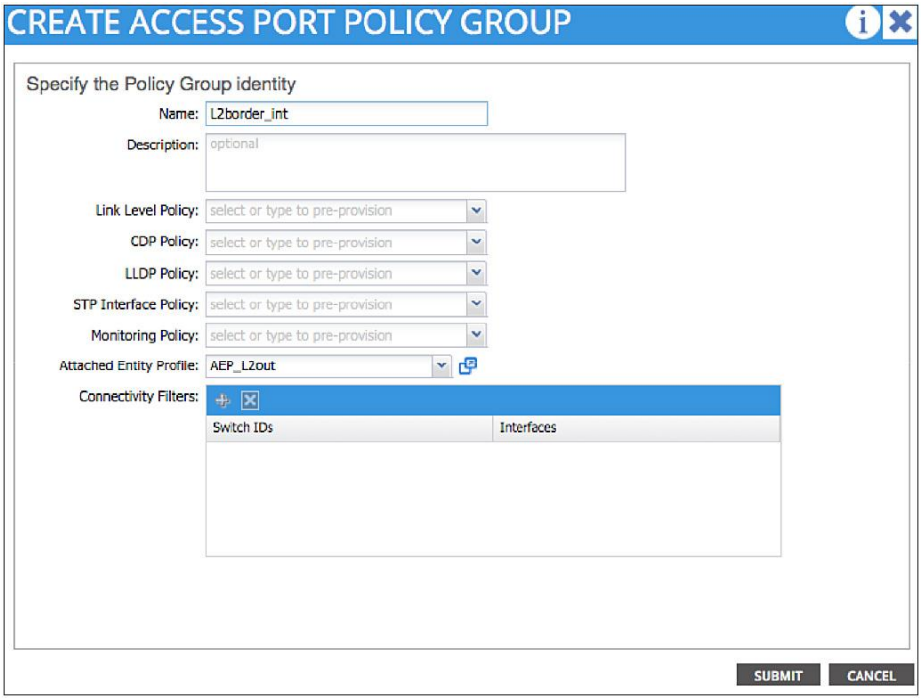
6.1 Create an attached entity profile by going to menu Fabric→Access Policies→Attachable Access Entity Profile (Figure 62). Click Actions on the right corner to add new attachable entity profile. Click the "+" sign to associate the profile with the layer 2 external domain, called "L2out-domain" (created in step 2).

Figure 62. Creating an Attachable Access Entity Profile



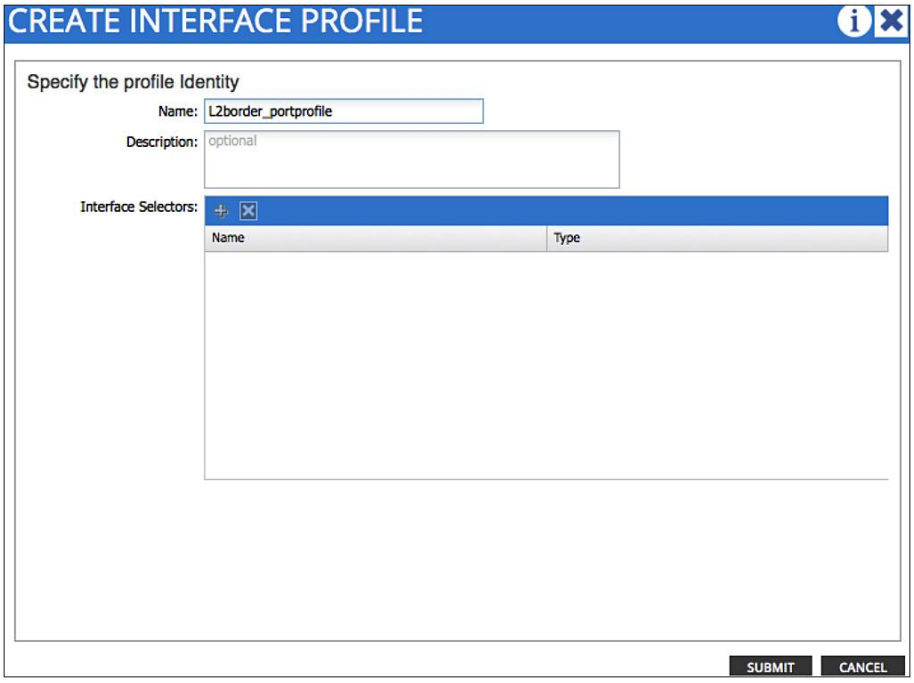
6.2 Create an Interface Policy Group by going to menu Fabric→Access Policies→Interface Policies→Policy Groups (Figure 63). Associate the policy group with an attachable entity profile, called "AEP_L2out".

Figure 63. Creating an Access Port Policy Group



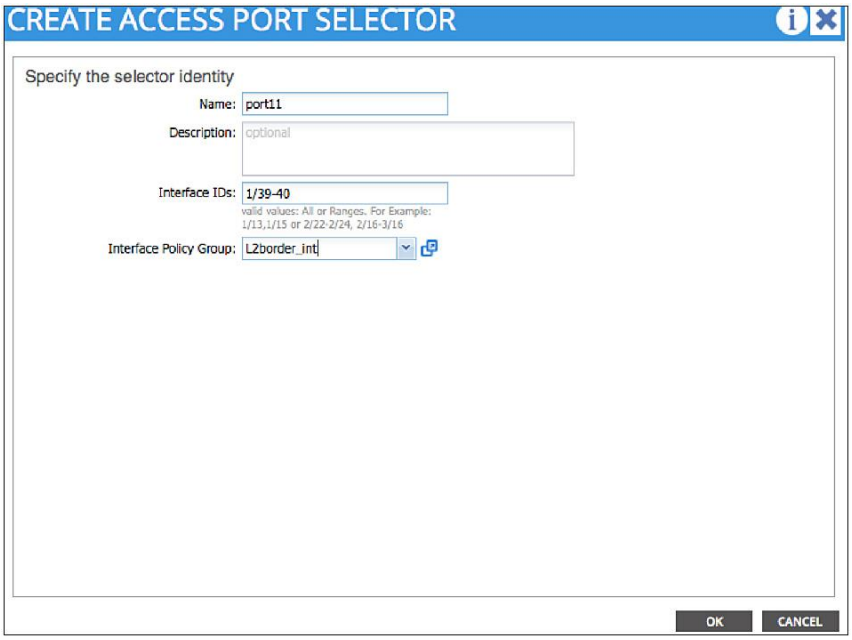
6.3 Create an interface profile named “L2border_portpfile” (Figure 64).

Figure 64. Creating an Interface Profile



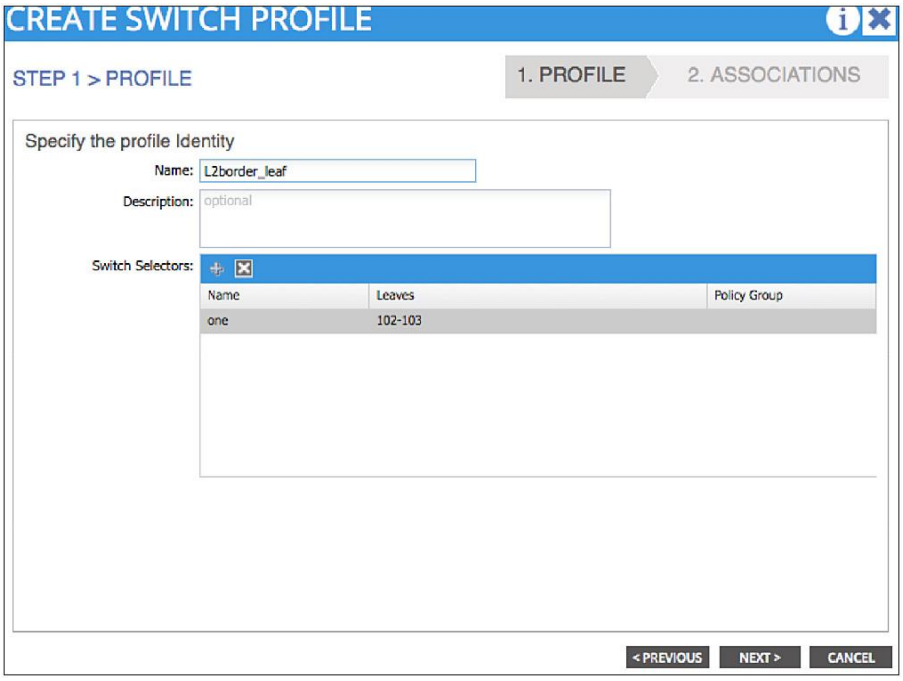
Click the “+” sign next to Interface Selectors to add interfaces to this interface profile (Figure 65). Add interface eth1/39 and eth1/40 to the profile and select the policy group created in step 6.2 - “L2border_int” - for this profile.

Figure 65. Add Interfaces to Port Selector



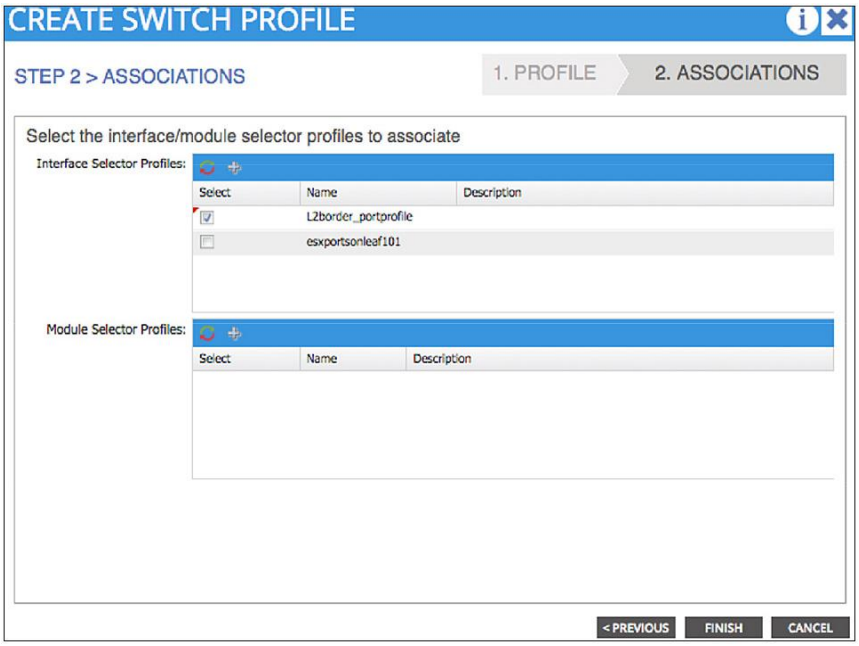
6.4 Create a switch profile by going to menu Fabric→Access Policies→Switch Policies→Profiles. Under “switch selectors” add leaf nodes 102 and 103 (Figure 66).

Figure 66. Creating a Switch Profile



6.5 Click Next to associate the interface profile with this switch profile (Figure 67). Select “L2border_portprofile” (created in step 6.3).

Figure 67. Associating the Interface Profile with the Switch Profile



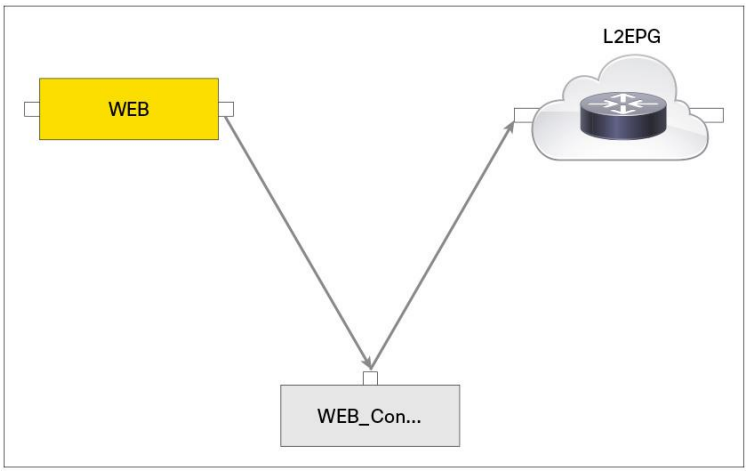
6.6 Click “Finish” to complete the process.

The attachable entity profile can be created by posting the following XML file to URI: http://apicip/api/policymgr/mo/uni.xml.

```
<polUni>
 <infraInfra>

<!—Associate attachable entity with L2 domain  -->
   <infraAttEntityP name="AEP_L2out">
      <infraRsDomP tDn="uni/l2dom-L2out-domain"/>

        <!-- Functions -->
      <infraProvAcc name="provfunc"/>
   </infraAttEntityP>

<!-- Policy Group, i.e. a bunch of configurations bundled together -->
   <infraFuncP>
         <infraAccPortGrp name="L2border_int">
            <infraRsAttEntP tDn="uni/infra/attentp-AEP_L2out" />
         </infraAccPortGrp>
   </infraFuncP>

<!-- Bundle ports together and reference the policy group -->
   <infraAccPortP name="L2border_portprofile">
      <infraHPortS name="port11" type="range">
         <infraPortBlk name="block0" fromPort="39" toPort="40" />
         <infraRsAccBaseGrp tDn="uni/infra/funcprof/accportgrp-L2border_int" />
      </infraHPortS>
   </infraAccPortP>
<!-- Bundle nodes together and reference the port selector -->
   <infraNodeP name="L2border_leaf">
      <infraLeafS name="one" type="range">
         <infraNodeBlk name="block0" from_="102" to_="103" />
      </infraLeafS>
      <infraRsAccPortP tDn="uni/infra/accportprof-L2border_portprofile" />
   </infraNodeP>

 </infraInfra>
</polUni>
```

After completing the configuration of the attachable entity profile the connection between the external endpoint behind the Nexus 3000 Switch and the endpoints in WEB EPG is established. As shown in the Application profile, the contract “WEB_contract” applies to the traffic flow between the WEB EPG and L2EPG (the external EPG for the layer 2 outside connection).

Figure 68. Contract between Internal EPG and External EPG



ACI Interaction with Spanning Tree Protocol (STP)

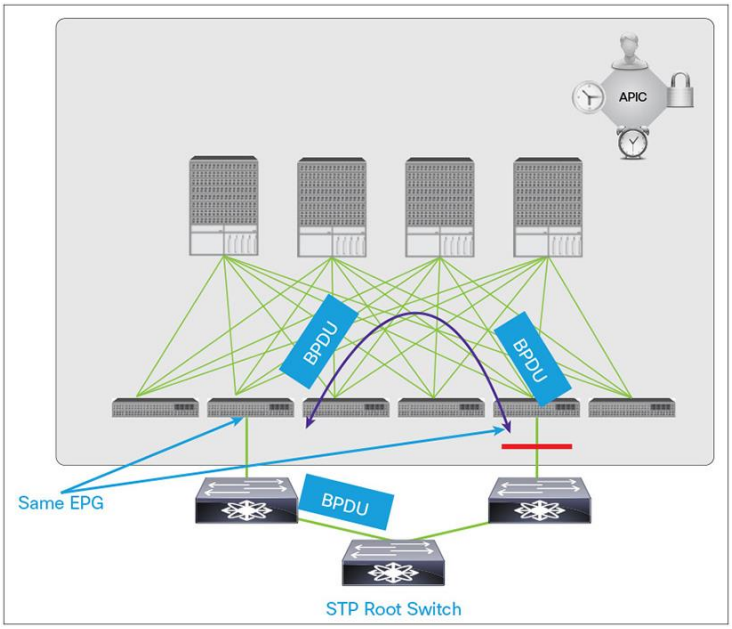
This section explores how ACI interacts with STP when connecting to an external layer 2 network.

Bridge Protocol Data Unit (BPDU) Flooding Behavior in the ACI Fabric

The ACI fabric is an IP-based fabric that implements an integrated overlay, allowing any subnet to be placed anywhere in the fabric and supports a fabric-wide mobility domain for virtualized workloads. STP is not required within the ACI fabric and leaf. The spine and APIC don't run STP instances.

When connecting to an outside layer 2 network, the ACI fabric floods the STP BPDU frame within the boundary of the EPG. External switches are expected to break any potential loop upon receiving the flooded BPDU from the ACI fabric. Figure 69 depicts this process.

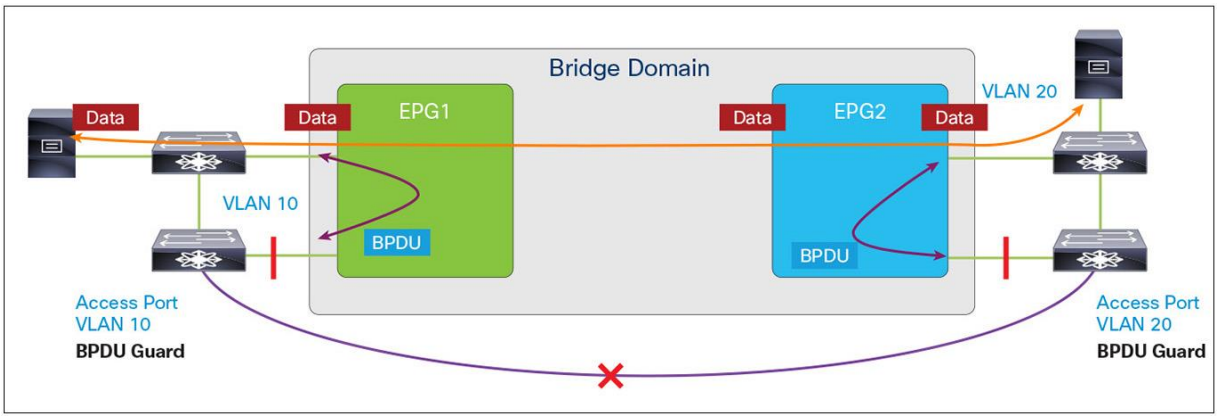
Figure 69. BPDU Flooding Behavior Process



The ACI leaf floods the BPDU frame within the EPG by using the VXLAN network identifier (VNID) assigned for the EPG when it encapsulates the BPDU in VXLAN format. The flooding scope of the BPDU is different than the one for data traffic. The unknown unicast traffic and broadcast traffic are flooded within the bridge domain. On the outside layer 2 network, STP instances are aligned with the VLAN boundary. To keep it consistent, the ACI fabric maintains the STP boundary by flooding the BPDU within the scope of the EPG.

The ACI fabric design is flexible; it allows multiple EPGs to be placed under the same bridge domain. This flexibility can be used to stitch different VLANs together or can be used as VLAN translation. Figure 70 illustrates this use case.

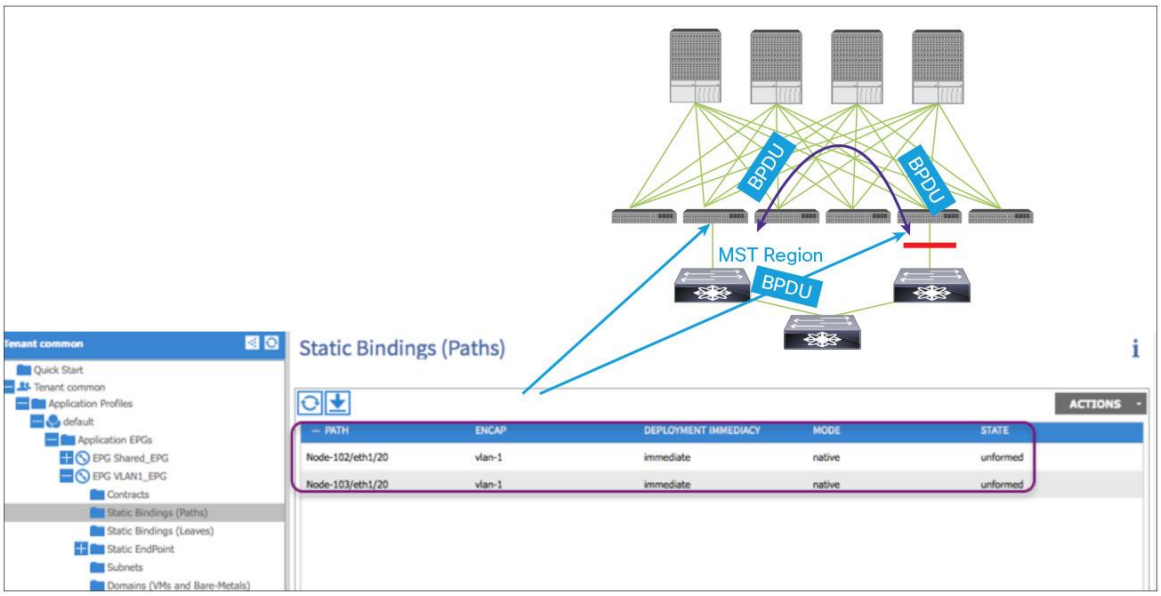
Figure 70. ACI Fabric Design Flexibility



In this example, VLAN 10 and 20 from the outside network are stitched together by the ACI fabric. The ACI fabric provides layer 2 bridging for traffic between these two VLANs. And they are in the same flooding domain. From the STP's perspective, the ACI fabric floods the BPDU within the EPG. When the ACI leaf receives the BPDU on EPG1 it floods to all leaf ports in EPG1 and it does not send the BPDU frame to ports in other EPGs. As a result, this flooding behavior can break the potential loop within the EPG (VLAN 10 and VLAN 20, respectively). It is important to make sure VLAN 10 and VLAN 20 do not have any physical connections other than the one provided by the ACI fabric. It is critical to turn on the BPDU guard feature on the access ports of outside switches. This way, if someone mistakenly connects the leaf switch to the right switch, the BPDU guard can disable the port and break the loop.

Additional configuration is required in order for Multiple Spanning Tree (MST) BPDU to be flooded properly. The BPDU frame for Per-VLAN Spanning Tree (PVST) and Rapid Per-VLAN Spanning Tree (RPVST) have a VLAN tag. The ACI leaf can identify which EPG the BPDU needs to be flooded based on the VLAN tag in the frame. However, for MST (802.1s), BPDU frames don't carry a VLAN tag and they 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 ACI fabric. By default there is no native VLAN enabled on the ACI fabric. To accept traffic for any VLAN, the VLAN needs to be previsioned, either by a statically assigned port and VLAN to an EPG, or by the EPG being provisioned dynamically by the APIC when there is integration between the APIC and Virtual Machine Manager (VMM), (such as vCenter or Microsoft SCVMM). As a result, to ensure MST BPDU is flooded to the desired ports, the user needs to create an EPG to carry the BPDU. As shown in Figure 71, the mode needs to be "native" given that the BPDU frame is untagged.

Figure 71. Assign Port to an EPG Using Native Mode



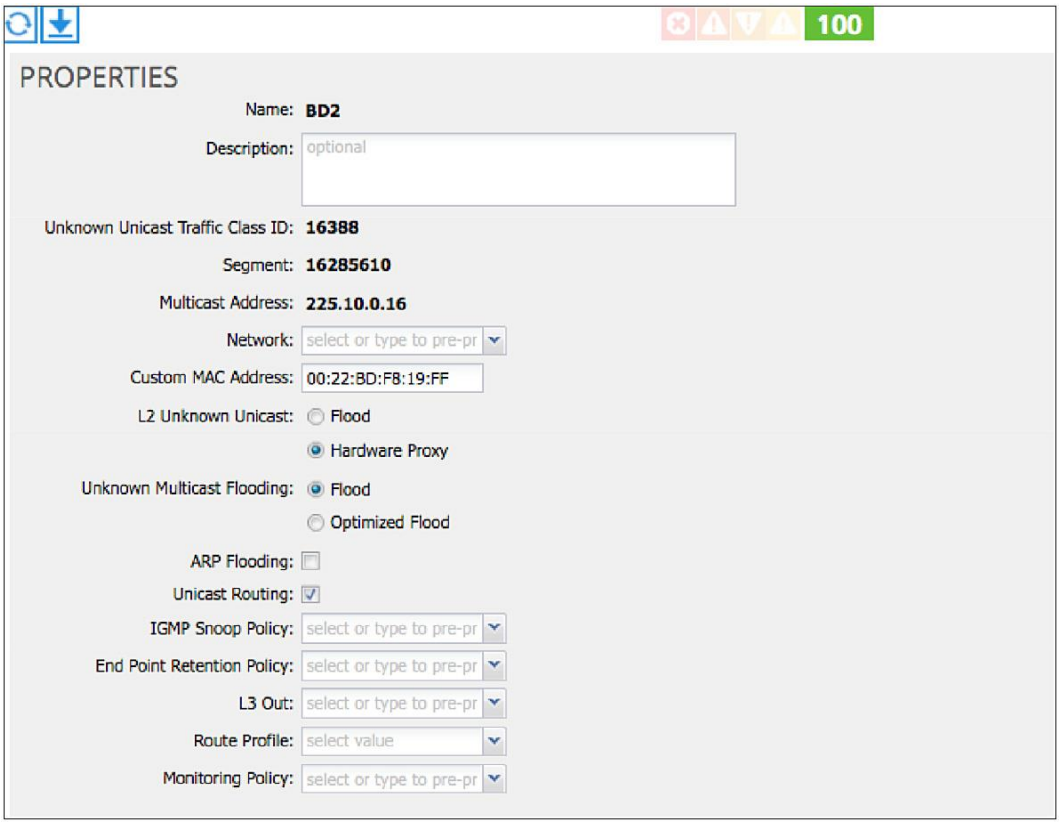
In addition to the configuration tasks highlighted in Figure 71, the user also needs to create a physical domain and associated VLAN pool (which includes VLAN 1 in this example), and the attachable access entity profile to allow VLAN 1 to be used for these ports.

It is recommended to use vPC to connect the outside layer 2 network to ACI so that there is no loop in the network. Additionally, it drastically reduces the chances of a topology change and reduces the impact to the network.

STP Topology Change Notification (TCN) Snooping

Although the ACI fabric control plan doesn't run STP, it does intercept the STP TCN frame so that it can flush out MAC address entries to avoid black-holding traffic when there is a STP topology change on the outside layer 2 network. Upon receiving an STP BPDU TCN frame, the APIC flushes the MAC addresses for the corresponding EPG that experienced the STP topology change. This does have an impact on the choice of how the ACI fabric forwards layer 2 unknown unicast. By default, the ACI leaf forwards the layer 2 unknown unicast traffic to a spine proxy for further lookup. The spine node will drop the packet if the MAC address doesn't exist in the proxy database. This option is called "Hardware Proxy," and it is the default option. Another option is a flood node, like a standard layer 2 switch. When the "Hardware Proxy" option is selected and the fabric detects an STP topology change, the MAC addresses for the EPG will be flushed in the fabric. The layer 2 traffic will disappear until the MAC addresses are learned again. To prevent this from happening as much as possible, it is recommended to use vPC to connect to the ACI leaf, and also to turn on the "peer-switch" feature to avoid a root-bridge change. Alternatively, a user can turn on the unknown unicast flooding to reduce the traffic disruption during an STP topology change.

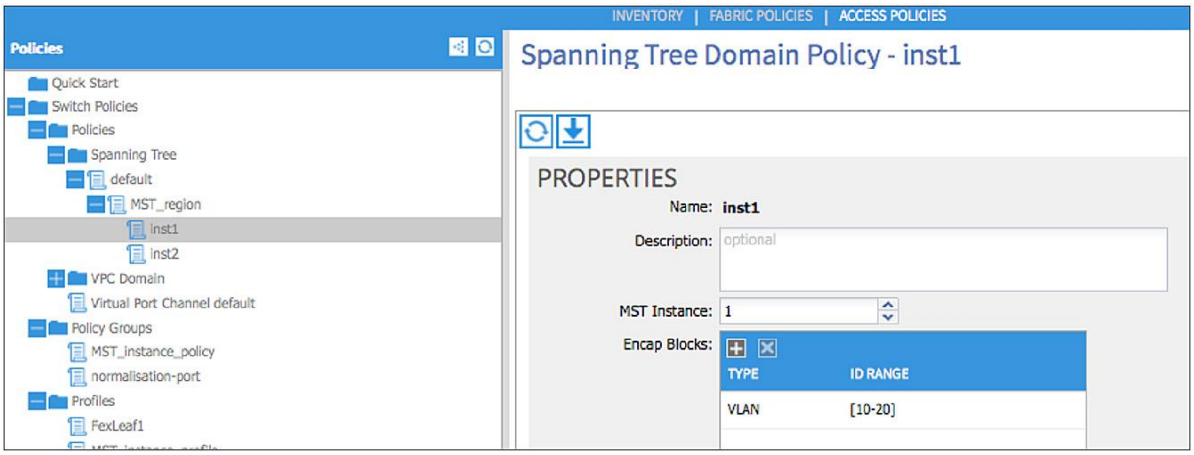
Figure 72. Forwarding Mode Configuration for Bridge Domain



With PVST and RPVST, the VLAN tag in the BPDU TCN frame indicates the VLAN that had a topology change. The APIC flushes the MAC addresses for the EPG that maps to the outside VLAN. With MST, the BPDU frame only indicates the instance ID that had the topology change. In order for the APIC to identify the EPGs for which the MAC entries need to be flushed, the user needs to configure a policy to create an STP instance to VLAN mapping on the APIC.

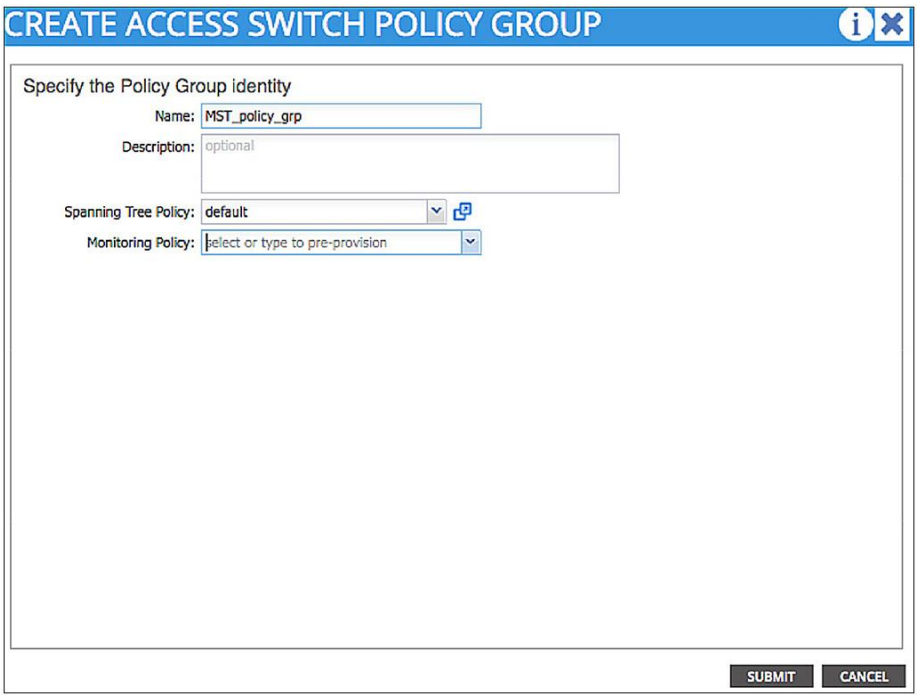
There are three major steps to follow in order to create the STP instance to VLAN mapping on the APIC. First, create a spanning tree policy under menu Fabric→Access Policies→Switch Polices→Policies→Spanning Tree. There is a policy named “default” created already. Under this policy, configure the MST region name, create an instance to VLAN mapping, and make sure they match with the MST configuration on the outside network. In the example in Figure 66, there are two instances: instance one has VLAN 10 to 20, and instance two has VLAN 21 to 30 (not shown in the Figure 73 screen capture).

Figure 73. Creating the STP Instance to VLAN Mapping



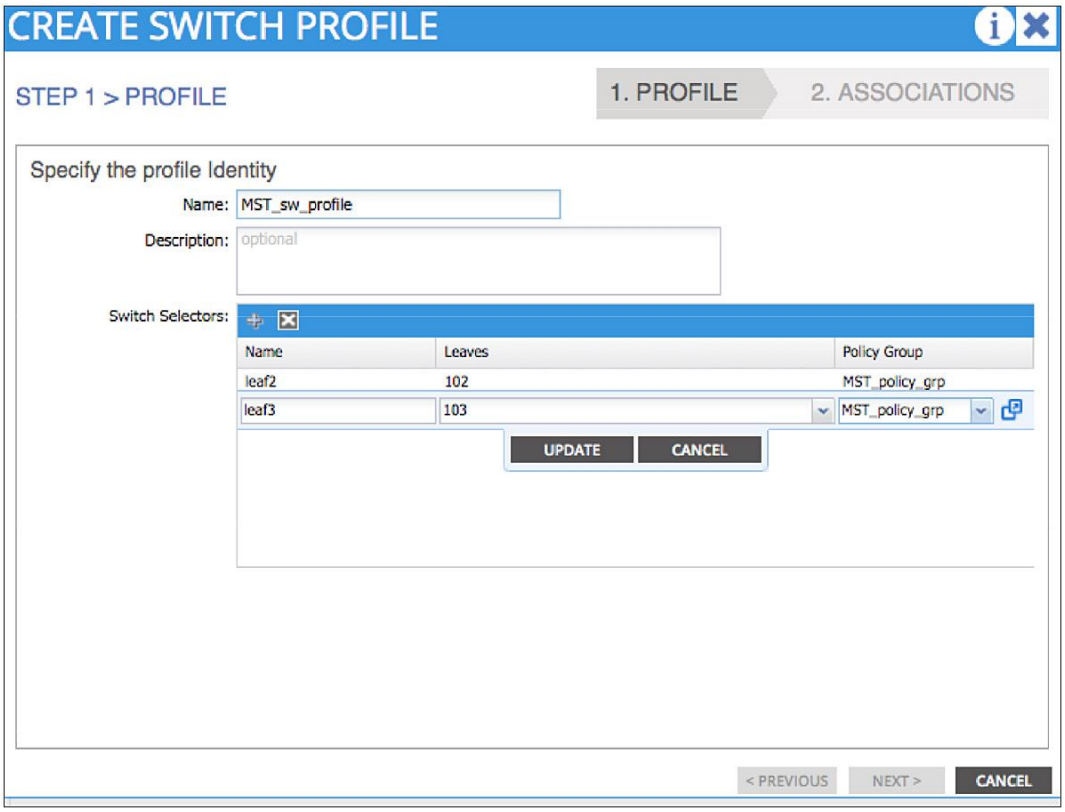
Second, create a policy group under menu Fabric→Access Policies→Switch Polices→Policy Groups and choose "default" for the spanning tree policy (Figure 74).

Figure 74. Creating a Policy Group that Includes Spanning Tree Policy



Last, apply this policy group to the selected leaf switch by creating switch profiles. In the example in Figure 68, the policy group "MST_policy_grp" created in the previous step is applied to leaf nodes 102 and 103.

Figure 75. Applying the Policy Group to Leaf Nodes



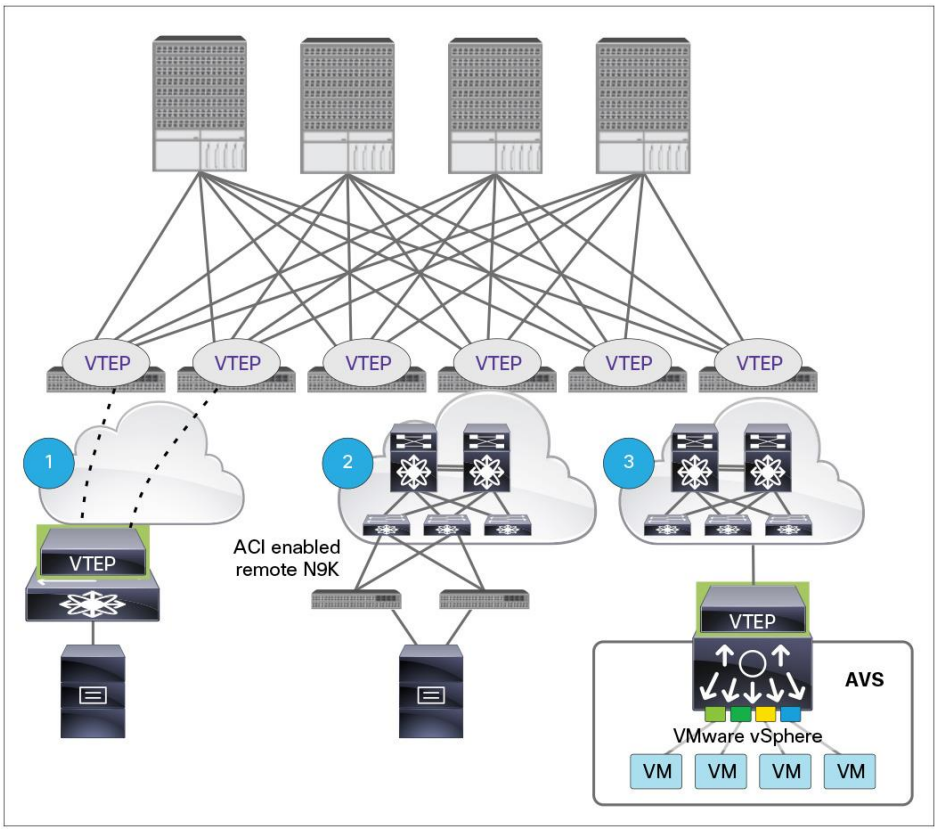
Ignore the Interface Selector and Module Selector in the configuration process.

Remote VXLAN Tunnel Endpoint (VTEP)

In the previous two sections we explored how to provide a layer 2 extension for the ACI fabric by extending an EPG or by extending a bridge domain. In both cases the traffic entered the ACI fabric with VLAN tagging. In this section, we briefly explain how to extend the ACI fabric with VXLAN by connecting a remote VTEP to the ACI fabric. Note that this section describes the software features that are planned for future release. We recommend users check with a Cisco sales representative or later documents for this feature availability.

With remote VTEP, a user can extend the ACI fabric to remote hosts that are connected to the ACI fabric through a layer 2 or IP network (Figure 76).

Figure 76. Extending ACI Fabric with Remote VTEP



The first option connects a remote switch that supports the VXLAN to the ACI fabric. The remote switch can be any VXLAN switch. Given that it doesn't support the ACI fabric, the southbound API OpFlex APIC can't manage this remote switch and the VXLAN-related configuration needs to be applied manually. With the remote VTEP in option 1, a user can extend an EPG or bridge domain similar to the one explained in the previous two sections. With remote VTEP, the ACI leaf identifies the incoming traffic with a VNID (instead of VLAN ID), and based on the configuration leaf switch, assign incoming traffic to an EPG and bridge domain. The ACI leaf also provides the default gateway function and policy enforcement for remote endpoints behind the remote VTEP. The ACI leaf can learn the remote endpoints and create MAC/IP to the remote VTEP and VNID mapping.

Option 2 extends the ACI fabric to a remote network by deploying a remote Nexus 9000 switch running ACI-enabled software. This remote switch is functioning as a remote leaf node. It is managed by the APIC and provides similar functions as a locally-attached ACI leaf, such as a distributed gateway, policy enforcement, learning endpoints, or application visibility (atomic counters, health score, etc.). The existing network infrastructure provides IP transportation between the remote Nexus 9000 switch and the ACI fabric. Compared to option 1, the remote Nexus 9000 switch in this design can be centrally managed by the APIC. It also provides local forwarding between EPGs while enforcing policies.

In the situation where remote hosts are virtualized and the customer is open to upgrading the virtual switch, option 3 can be deployed. Option 3 installs an Application Virtual Switch (AVS) on the remote hosts and extends the ACI fabric to remote hosts that are attached to existing network infrastructure. The AVS supports OpFlex and can be managed by the APIC. It brings the ACI forwarding, policy model, policy enforcement, and application visibility to remote hosts without touching existing network infrastructure. The AVS software version supports full switching functionality, and provides local switching for traffic within an EPG, cross-EPG, and traffic between a bare-metal server and virtualized server. This is a natural choice for customers who want to take advantage of existing network infrastructure and consider investment protection a high priority.

Conclusion

The ACI fabric solution provides many options to allow users to seamlessly connect to existing network infrastructure through layer 2 or layer 3 technology. These options accommodate the requirements for WAN connection, layer 2/layer 3 DCI, brownfield field data center deployment, migration, and investment protection. The ACI fabric connections to outside networks not only provide connectivity, but they also extend policy models to cover communication between outside endpoints and directly attached endpoints.

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 www.cisco.com/go/offices.

Printed in USA C07-732033-00 08/14