Multi-Pod

This chapter contains the following sections:

About Multi-Pod

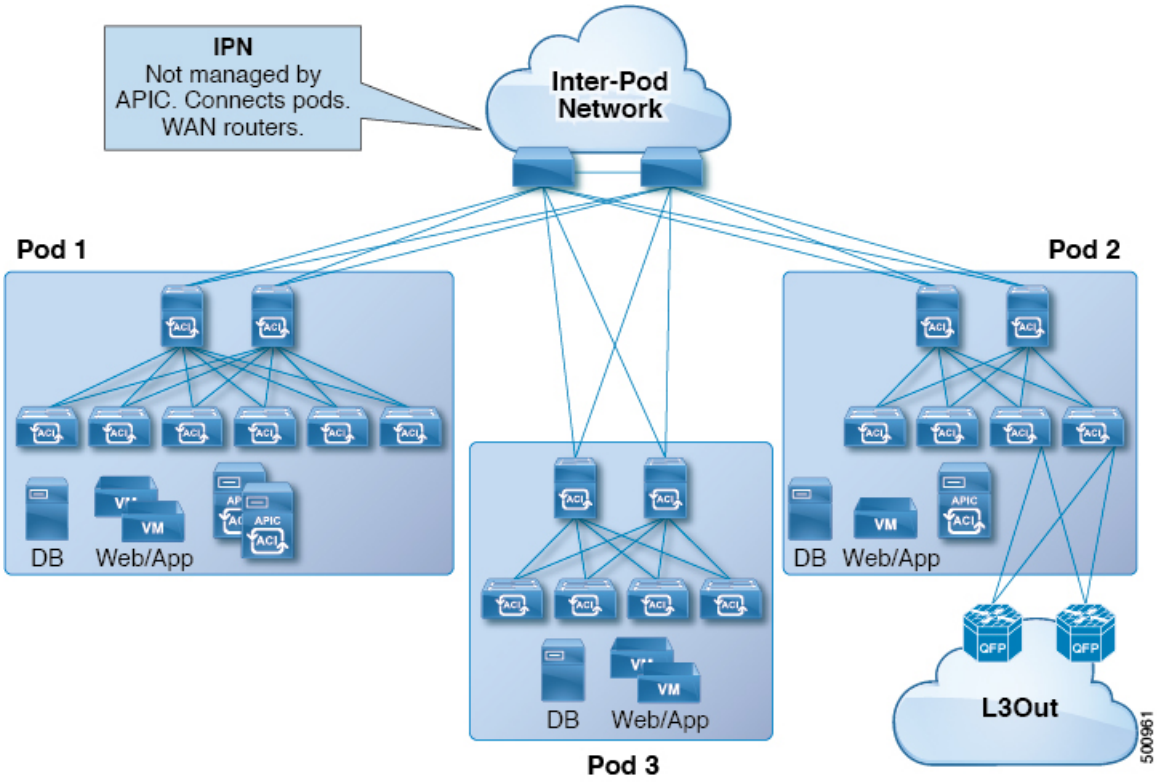
Multi-Pod enables provisioning a more fault-tolerant fabric comprised of multiple pods with isolated control plane protocols. Also, Multi-Pod provides more flexibility with regard to the full mesh cabling between leaf and spine switches. For example, if leaf switches are spread across different floors or different buildings, Multi-Pod enables provisioning multiple pods per floor or building and providing connectivity between pods through spine switches.

Multi-Pod uses MP-BGP EVPN as the control-plane communication protocol between the ACI spines in different pods.

WAN routers can be provisioned in the Inter-Pod Network (IPN), directly connected to spine switches, or connected to border leaf switches. Spine switches connected to the IPN are connected to at least one leaf switch in the pod.

Multi-Pod uses a single APIC cluster for all the pods; all the pods act as a single fabric. Individual APIC controllers are placed across the pods but they are all part of a single APIC cluster.

Figure 1: Multi-Pod Overview



Multi-Pod Provisioning

The IPN is not managed by the APIC. It must be preconfigured with the following information:

- Configure the interfaces connected to the spines of all pods. Use Layer 3 sub-interfaces tagging traffic with VLAN-4 and increase the MTU at least 50 bytes above the maximum MTU required for inter-site control plane and data plane traffic.

  If remote leaf switches are included in any pods, we strongly recommend that you deploy ACI software release 4.1(2) or later. A more complex configuration is required with earlier releases to connect the spines to the IPN, mandating the use of two sub-interfaces (with VLAN-4 and VLAN-5 tags) and a separate VRF on the IPN devices. For more information, see the Cisco ACI Remote Leaf Architecture White Paper.

- Enable OSPF on sub-interfaces with the correct area ID.
- Enable DHCP Relay on IPN interfaces connected to all spines.
- Enable PIM.
- Add bridge domain GIPO range as PIM Bidirectional (**bidir**) group range (default is 225.0.0.0/8).

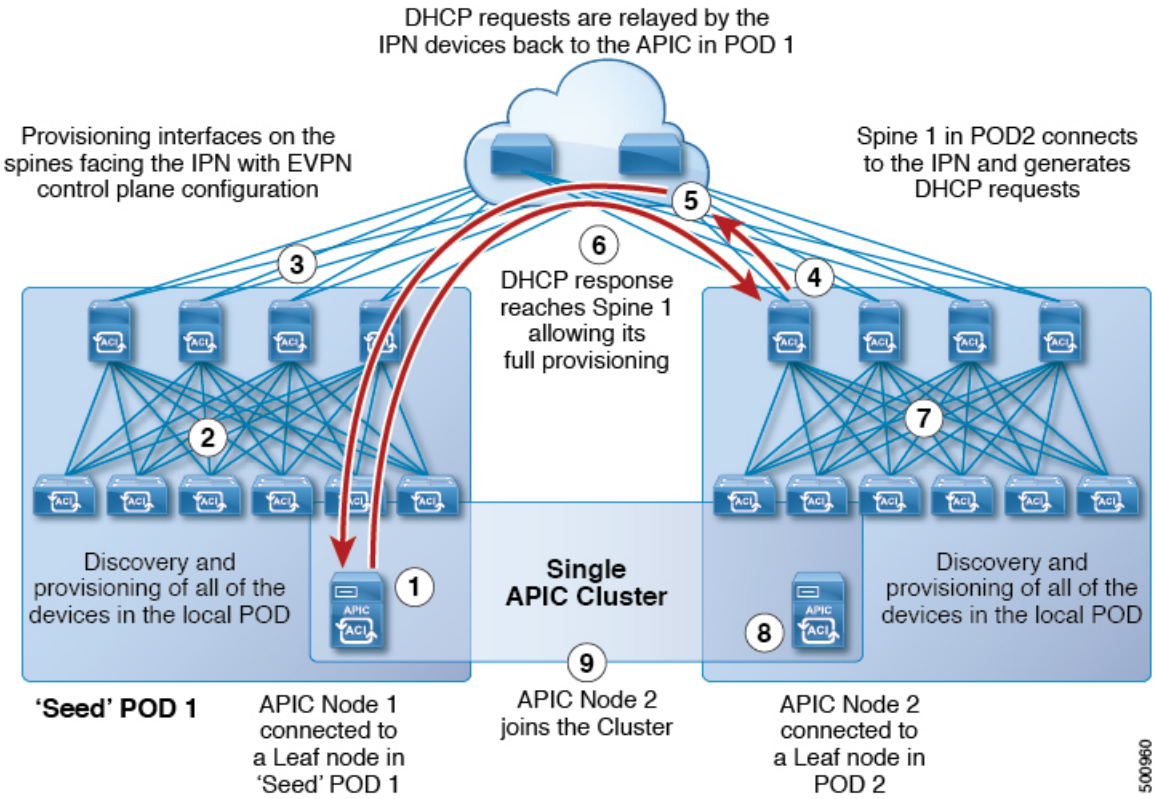
  A group in **bidir** mode has only shared tree forwarding capabilities.

- Add 239.255.255.240/28 as PIM **bidir** group range.

- Enable PIM and IGMP on the interfaces connected to all spines.

Note When deploying PIM **bidir**, at any given time it is only possible to have a single active RP (Rendezvous Point) for a given multicast group range. RP redundancy is hence achieved by leveraging a **Phantom RP** configuration. Because multicast source information is no longer available in Bidir, the Anycast or MSDP mechanism used to provide redundancy in sparse-mode is not an option for **bidir**.

Figure 2: Multi-Pod Provisioning



Guidelines for Setting Up a Cisco ACI Multi-Pod Fabric

To configure a Cisco ACI Multi-Pod fabric, follow these guidelines:

- Cisco ACI Multi-Pod is supported on the following:
  - All ACI-mode spine switches
  - All Cisco Nexus 9000 Series ACI-mode leaf switches
  - All of the Cisco Nexus 9500 platform ACI-mode switch line cards and fabric modules
- Create the associated node group and Layer 3 Outside (L3Out) policies.

- Before you make any changes to a spine switch, ensure that there is at least one operationally “up” external link that is participating in the Cisco ACI Multi-Pod topology. Failure to do so could bring down the Cisco ACI Multi-Pod connectivity.
- If you have to convert a Cisco ACI Multi-Pod setup to a single pod (containing only Pod 1), the Cisco Application Policy Infrastructure Controllers (APICs) connected to the pods that are decommissioned should be re-initialized and connected to the leaf switches in Pod 1, which will allow them to re-join the cluster after going through the initial setup script. See Moving an APIC from One Pod to Another Pod, on page 11 for those instructions. The TEP pool configuration should not be deleted.
- Cisco ACI GOLF (also known as Layer 3 EVPN Services for Fabric WAN) and Cisco ACI Multi-Pod can be deployed together over all the switches used in the Cisco ACI Multi-Pod and EVPN topologies.

  For more information on GOLF, see Cisco ACI GOLF.
- In a Cisco ACI Multi-Pod fabric, the Pod 1 configuration (with the associated TEP pool) must always exist on Cisco APIC, as the Cisco APIC nodes are always addressed from the Pod 1 TEP pool. This remains valid also in the scenario where the Pod 1 is physically decommissioned (which is a fully supported procedure) so that the original Pod 1 TEP pool is not re-assigned to other pods that may be added to the fabric.
- In a Cisco ACI Multi-Pod fabric setup, if a new spine switch is added to a pod, it must first be connected to at least one leaf switch in the pod. This enables the Cisco APIC to discover the spine switch and join it to the fabric.
- After a pod is created and nodes are added in the pod, deleting the pod results in stale entries from the pod that are active in the fabric. This occurs because the Cisco APIC uses open source DHCP, which creates some resources that the Cisco APIC cannot delete when a pod is deleted
- If you connect spine switches belonging to separate pods with direct back-to-back links, an OSPF neighborship might get established on the peer interface between the two spine switches. If there is a mismatch between the peer interfaces, with one of the peers having the Cisco ACI Multi-Pod direct flag disabled, the session won't be up and forwarding will not happen. Even though the system will throw a fault in this situation, this is expected behavior.
- Beginning with Cisco APIC release 5.2(3), the IPN underlay protocol can be external BGP (eBGP). Internal BGP (iBGP) is not supported as the underlay protocol.

  When preparing to migrate a Cisco ACI Multi-Pod fabric between OSPF and BGP as the IPN underlay, follow these guidelines:

  - A BGP underlay is not supported if the Cisco ACI fabric is connected to a cloud site or to a GOLF router.
  - A BGP underlay supports only an IPv4 address family, not an IPv6 address family.
- When deploying Cisco APIC cluster connectivity to the fabric over a Layer 3 network, which was introduced in Cisco APIC release 5.2(1), the IPN network can use OSPF as the underlay protocol, or a BGP underlay if the Cisco APIC connects to the fabric using the same network that provides Cisco ACI Multi-Pod or remote leaf switch connectivity.
- If you delete and recreate the Cisco ACI Multi-Pod L3Out, for example to change the name of a policy, a clean reload of some of the spine switches in the fabric must be performed. The deletion of the Cisco ACI Multi-Pod L3Out causes one or more of the spine switches in the fabric to lose connectivity to the Cisco APICs and these spine switches are unable to download the updated policy from the Cisco APIC. Which spine switches get into such a state depends upon the deployed topology. To recover from this

state, a clean reload must be performed on these spine switches. The reload is performed using the **setup-clean-config.sh** command, followed by the reload command on the spine switch.

Note Cisco ACI does not support IP fragmentation. Therefore, when you configure Layer 3 Outside (L3Out) connections to external routers, or Multi-Pod connections through an Inter-Pod Network (IPN), it is recommended that the interface MTU is set appropriately on both ends of a link. On some platforms, such as Cisco ACI, Cisco NX-OS, and Cisco IOS, the configurable MTU value does not take into account the Ethernet headers (matching IP MTU, and excluding the 14-18 Ethernet header size), while other platforms, such as IOS-XR, include the Ethernet header in the configured MTU value. A configured value of 9000 results in a max IP packet size of 9000 bytes in Cisco ACI, Cisco NX-OS, and Cisco IOS, but results in a max IP packet size of 8986 bytes for an IOS-XR untagged interface.

For the appropriate MTU values for each platform, see the relevant configuration guides.

We highly recommend that you test the MTU using CLI-based commands. For example, on the Cisco NX-OS CLI, use a command such as `ping 1.1.1.1 df-bit packet-size 9000 source-interface ethernet 1/1`.

Note Cisco APIC will always establish a TCP connection to fabric switches with an MTU of 1496 bytes (TCP MSS 1456) regardless of the CP-MTU setting. The IPN network for remote pods and remote leaf switches must support at least 1500 byte MTU for fabric discovery.

- You can set the global MTU for control plane (CP) packets sent by the nodes (Cisco APIC and the switches) in the fabric at **System** > **System Settings** > **Control Plane MTU**.
- In a Cisco ACI Multi-Pod topology, the MTU set for the fabric external ports must be greater than or equal to the CP MTU value set. Otherwise, the fabric external ports might drop the CP MTU packets.
- If you change the IPN or CP MTU, we recommend changing the CP MTU value first, then changing the MTU value on the spine of the remote pod. This reduces the risk of losing connectivity between the pods due to MTU mismatch. This is to ensure that the MTU across all the interfaces of the IPN devices between the pods is large enough for both control plane and VXLAN data plane traffic at any given time. For data traffic, keep in mind the extra 50 bytes due to VXLAN.
- To decommission a pod, decommission all the nodes in the pod. For instructions, see *Decommissioning and Recommissioning a Pod* in *Cisco APIC Troubleshooting Guide*.

Setting Up the Multi-Pod Fabric

In Cisco Application Policy Infrastructure Controller (APIC) 4.0(1) and later, a wizard was added to the GUI to simplify Multi-Pod configuration. To configure Multi-Pod using the GUI, follow the procedures in this section.

Setting up Multi-Pod between two physical pods involves preparing an existing physical pod to communicate over the interpod network (IPN) with the new pod. You then add the physical pod, and Cisco Cisco APIC creates the Multi-Pod fabric.

You can also configure Multi-Pod using the NX-OS style CLI and REST API. See the sections Setting Up Multi-Pod Fabric Using the NX-OS CLI and Setting Up Multi-Pod Fabric Using the REST API in this guide for instructions.

Note You can also use the GUI wizard to add a Cisco Application Centric Infrastructure (ACI) Virtual Pod (vPod) as a remote extension of the Cisco ACI fabric. For information about Cisco ACI vPod, see the Cisco ACI vPod documentation.

Preparing the Pod for IPN Connectivity

Before you create a new pod, you first must ensure that the existing physical pod can communicate with it.

Procedure

Step 1 Log in to the Cisco APIC.

Step 2 Choose **Fabric** > **Inventory**.

Step 3 Expand **Quick Start** and click **Add Pod**.

Step 4 In the work pane, click **Add Pod**.

Step 5 In the **Configure Interpod Connectivity STEP 1 > Overview** panel, review the tasks that are required to configure interpod network (IPN) connectivity, and then click **Get Started**.

Step 6 In the **Configure Interpod Connectivity STEP 2 > IP Connectivity** dialog box, complete the following steps:

a) If you see a **Name** field in an **L3 Outside Configuration** area, choose an existing fabric external routing profile from the **Name** drop-down list.

b) Using the **Spine ID** selector, choose the spine.

Click the + (plus) icon to add the IDs of more spines.

c) In the **Interfaces** area, in the **Interface** field, enter the spine switch interface (slot and port) used to connect to the IPN.

Click the + (plus) icon to add more interfaces.

d) In the **IPV4 Address** field, enter the IPv4 gateway address and network mask for the interface.

e) From the **MTU (bytes)** drop-down list, choose a value for the maximum transmit unit of the external network.

The range is 1500 to 9216.

f) Click **Next**.

Step 7 **Configure Interpod Connectivity STEP 3 > Routing Protocols** dialog box, in the **OSPF** area, complete the following steps:

a) Leave the **Use Defaults** checked or uncheck it.

When the **Use Defaults** check box is checked, the GUI conceals the optional fields for configuring Open Shortest Path (OSPF). When it is unchecked, it displays all the fields. The check box is checked by default.

b) In the **Area ID** field, enter the OSPF area ID.

c) In the **Area Type** area, choose an OSPF area type.

You can choose **NSSA area**, **Regular area** (the default), or **Stub area**.

d) (Optional) With the **Area Cost** selector, choose an appropriate OSPF area cost value.

e) From the **Interface Policy** drop-down list, choose or configure an OSPF interface policy.

You can choose an existing policy, or you can create one with the **Create OSPF Interface Policy** dialog box.

Step 8 In the **Configure Interpod Connectivity STEP 3 > Routing Protocols** dialog box, in the **BGP** area, complete the following steps:

a) Leave the **Use Defaults** checked or uncheck it.

When the **Use Defaults** check box is checked, the GUI conceals the fields for configuring Border Gateway Protocol (BGP). When it is unchecked, it displays all the fields. The check box is checked by default.

b) In the **Community** field, enter the community name.

We recommend that you use the default community name. If you use a different name, follow the same format as the default.

c) In the **Peering Type** field, choose either **Full Mesh** or **Route Reflector** for the route peering type.

If you choose **Route Reflector** in the **Peering Type** field and you later want to remove the spine switch from the controller, you must first disable **Route Reflector** in the *BGP Route Reflector* page. Not doing so results in an error.

To disable a route reflector, right-click on the appropriate route reflector in the **Route Reflector Nodes** area in the **BGP Route Reflector** page and select **Delete**. See the section "Configuring an MP-BGP Route Reflector Using the GUI" in the chapter "MP-BGP Route Reflectors" in the *Cisco APIC Layer 3 Networking Configuration Guide*.

d) In the **Peer Password**, field, enter the BGP peer password.

e) In the **Confirm Password** field, reenter the BGP peer password.

f) In the **External Route Reflector Nodes** area, click the + (plus) icon to add nodes.

For redundancy purposes, more than one spine is configured as a route reflector node: one primary reflector and one secondary reflector. It is best practice to deploy at least one external route reflector per pod for redundancy purposes.

The **External Route Reflector Nodes** fields appear only if you chose **Route Reflector** as the peering type.

g) Click **Next**.

Step 9 In the **Configure Interpod Connectivity STEP 4 > External TEP** dialog box, complete the following steps:

a) Leave the **Use Defaults** checked or uncheck it.

When the **Use Defaults** check box is checked, the GUI conceals the optional fields for configuring the external TEP pool. When it is unchecked, it displays all the fields. The check box is checked by default.

b) Note the nonconfigurable values in the **Pod** and **Internal TEP Pool** fields.

c) In the **External TEP Pool** field, enter the external TEP pool for the physical pod.

The external TEP pool must not overlap the internal TEP pool or external TEP pools belonging to other pods.

d) In the **Dataplane TEP Pool** field, accept the default, which is generated when you configure the **External TEP Pool**; if you enter another address, it must be outside of the external TEP pool.
e) (Optional) In the **Router ID** field, enter the IPN router IP address.
f) (Optional) In the **Loopback Address** field, enter the IPN router loopback IP address.

If you uncheck the **Use Defaults**, the Cisco APIC displays the nonconfigurable **Unicast TEP IP** and **Spine ID** fields.

g) Click **Finish**.
The **Summary** panel appears, displaying details of the IPN configuration. You can also click **View JSON** to view the REST API for the configuration. You can save the REST API for later use.

What to do next

Take one of the following actions:

- You can proceed directly with adding a pod, continuing with the procedure Adding a Pod to Create a Multi-Pod Fabric, on page 8 in this guide.
- Close the **Configure Interpod Connectivity** dialog box and add the pod later, returning to the procedure Adding a Pod to Create a Multi-Pod Fabric, on page 8 in this guide.

Adding a Pod to Create a Multi-Pod Fabric

The **Add Physical Pod** dialog enables you to set up a Multi-Pod environment. You define a new physical pod ID and tunnel endpoint (TEP) pool. You also configure the new pod network settings and the subinterfaces for the physical spines.

Before you begin

You have performed the following tasks:

- Created the node group and L3Out policies.
- Configured the interpod network (IPN). For a sample configuration, see Sample IPN Configuration for Multi-Pod For Cisco Nexus 9000 Series Switches, on page 10 in this guide.
- Prepared an existing pod to communicate with the new pod over the IPN. See the procedure Preparing the Pod for IPN Connectivity, on page 6 in this guide.
- Made sure that the spine switch that connects to the IPN also connects to at least one leaf switch in the pod.
- Created a tunnel endpoint (TEP) pool. See the procedure Preparing the Pod for IPN Connectivity, on page 6 in this guide.

Procedure

Step 1 Log in to Cisco Application Policy Infrastructure Controller (APIC).
Step 2 Take one of the following actions:

- If you completed the procedure Preparing the Pod for IPN Connectivity, on page 6 and have not closed the **Configure Interpod Connectivity** dialog box, skip Step 3 through Step 5, and resume this procedure at Step 6.
- If you have completed the procedure Preparing the Pod for IPN Connectivity, on page 6 and have closed the **Configure Interpod Connectivity** dialog box, proceed to Step 3 in this procedure.

Step 3 Choose **Fabric** > **Inventory**.

Step 4 Click **Quick Start** and click **Add Pod**.

Step 5 In the work pane, click **Add Pod**.

Step 6 In the **Add Physical Pod STEP 2 > Pod Fabric** dialog box, complete the following steps:

a) In the **Pod ID** field, choose the pod ID.

The pod ID can be any positive integer; however, it must be unique in the Cisco ACI fabric.

b) In the **Pod TEP Pool** field, enter the pool address and subnet.

The pod TEP pool represents a range of traffic encapsulation identifiers and is a shared resource and can be consumed by multiple domains.

c) With the **Spine ID** selector, choose the spine ID.

Choose more spine IDs by clicking the + (plus) icon.

d) In the **Interfaces** area, in the **Interface** field, enter the spine switch interface (slot and port) that is used to connect to the interpod network (IPN).

e) In the **IPv4 Address** field, enter the IPv4 gateway address and network mask for the interface.

f) In the **MTU (bytes)** field, choose a value for the maximum transmit unit (MTU) of the external network.

You can configure another interface by clicking the + (plus) icon.

Step 7 In the **Add Physical Pod STEP 3 > External TEP** dialog box, complete the following steps:

a) Leave the **Use Defaults** check box checked or uncheck it to display the optional fields to configure an external TEP pool.

b) Note the values in the **Pod** and **Internal TEP Pool** fields, which are already configured.

c) In the **External TEP Pool** field, enter the external TEP pool for the physical pod.

The external TEP pool must not overlap the internal TEP pool.

d) In the **Dataplane TEP IP** field, enter the address that is used to route traffic between pods.

e) (Optional) In the **Unicast TEP IP** field, enter the unicast TEP IP address.

Cisco APIC automatically configures the unicast TEP IP address when you enter the data plane TEP IP address.

f) (Optional) Note the value in the nonconfigurable **Node** field.

g) (Optional) In the **Router ID** field, enter the IPN router IP address.

Cisco APIC automatically configures the router IP address when you enter the data plane TEP address.

h) In the **Loopback Address** field, enter the router loopback IP address.

Leave the **Loopback Address** blank if you use a router IP address.

i) Click **Finish**.

Sample IPN Configuration for Multi-Pod For Cisco Nexus 9000 Series Switches

Note

- The deployment of a dedicated VRF in the IPN for Inter-Pod connectivity is optional, but is a best practice recommendation. You can also use a global routing domain as an alternative.
- For the area of the sample configuration that shows `ip dhcp relay address 10.0.0.1`, this configuration is valid based on the assumption that the TEP pool of Pod 1 is 10.0.0.0/x.

Procedure

Sample configuration:

Example:

```
Sample IPN configuration for Cisco Nexus 9000 series switches:
================================

  (pod1-spine1)-----2/7[ IPN-N9K ]2/9-----(pod2-spine1)


feature dhcp
feature pim

service dhcp
ip dhcp relay
ip pim ssm range 232.0.0.0/8

# Create a new VRF for Multipod.
vrf context fabric-mpod
  ip pim rp-address 12.1.1.1 group-list 225.0.0.0/8 bidir
  ip pim rp-address 12.1.1.1 group-list 239.255.255.240/28 bidir
  ip pim ssm range 232.0.0.0/8

interface Ethernet2/7
  no switchport
  mtu 9150
  no shutdown

interface Ethernet2/7.4
  description pod1-spine1
  mtu 9150
  encapsulation dot1q 4
  vrf member fabric-mpod
  ip address 201.1.2.2/30
  ip router ospf a1 area 0.0.0.0
  ip pim sparse-mode
  ip dhcp relay address 10.0.0.1
  ip dhcp relay address 10.0.0.2
  ip dhcp relay address 10.0.0.3
  no shutdown
```

```
interface Ethernet2/9
  no switchport
  mtu 9150
  no shutdown

interface Ethernet2/9.4
  description to pod2-spine1
  mtu 9150
  encapsulation dot1q 4
  vrf member fabric-mpod
  ip address 203.1.2.2/30
  ip router ospf a1 area 0.0.0.0
  ip pim sparse-mode
  ip dhcp relay address 10.0.0.1
  ip dhcp relay address 10.0.0.2
  ip dhcp relay address 10.0.0.3
  no shutdown

interface loopback29
  vrf member fabric-mpod
  ip address 12.1.1.1/32

router ospf a1
  vrf fabric-mpod
    router-id 29.29.29.29
```

Moving an APIC from One Pod to Another Pod

Use this procedure to move an APIC from one pod to another pod in an Multi-Pod setup.

Procedure

Step 1 Decommission the APIC in the cluster.

a) On the menu bar, choose **System** > **Controllers**.
b) In the **Navigation** pane, expand **Controllers** > **apic_controller_name** > **Cluster as Seen by Node**.
c) In the **Navigation** pane, click an **apic_controller_name** that is within the cluster and not the controller that is being decommissioned.
d) In the **Work** pane, verify that the **Health State** in the **Active Controllers** summary table indicates the cluster is **Fully Fit** before continuing.
e) In the **Work** pane, click **Actions** > **Decommission**.
f) Click **Yes**.
The decommissioned controller displays **Unregistered** in the **Operational State** column. The controller is then taken out of service and no longer visible in the **Work** pane.

Step 2 Move the decommissioned APIC to the desired pod.

Step 3 Enter the following commands to reboot the APIC.

```
apic1# acidiag touch setup
apic1# acidiag reboot
```

Step 4 In the APIC setup script, specify the pod ID where the APIC node has been moved.

a) Log in to Cisco Integrated Management Controller (CIMC).
b) In the pod ID prompt, enter the pod ID.

Note Do not modify the **TEP Pool** address information.

Step 5 Recommission the APIC.

a) From the menu bar, choose **SYSTEM** > **Controllers**.
b) In the **Navigation** pane, expand **Controllers** > **apic_controller_name** > **Cluster as Seen by Node**.
c) From the **Work** pane, verify in the **Active Controllers** summary table that the cluster **Health State** is **Fully Fit** before continuing.
d) From the **Work** pane, click the decommissioned controller that displaying **Unregistered** in the **Operational State** column.
e) From the **Work** pane, click **Actions** > **Commission**.
f) In the **Confirmation** dialog box, click **Yes**.
g) Verify that the commissioned Cisco APIC controller is in the operational state and the health state is **Fully Fit**.