Deploying Cisco Virtual APIC Using Nutanix

Revised: September 10, 2026

New and changed information

This table provides an overview of the significant changes up to this current release. The table does not provide an exhaustive list of all changes or of the new features up to this release.

Cisco APIC Release Version	Feature
6.1(4)	Support for deploying virtual APIC using Nutanix AHV.

Overview of Cisco virtual APIC deployment using Nutanix AHV

Beginning with Cisco APIC release 6.0(2), you can deploy a cluster wherein all the APICs in the cluster are virtual APICs. From Cisco APIC release 6.0(2), you can deploy a virtual APIC on AWS using the CloudFormation template or a virtual APIC on the ESX host using the OVF template in VMware vCenter. From Cisco APIC release 6.1(4), you can deploy a virtual APIC on an AHV host using the OVA template in Nutanix AHV (Acropolis Hypervisor). The virtual APIC can be deployed on an existing server on the customer premises.

This document provides details about deploying a virtual APIC on Nutanix AHV. For information about deploying virtual APIC using AWS, see the Deploying Cisco Virtual APIC Using AWS document and for information about deploying virtual APIC using VMware vCenter, see the Deploying Cisco Virtual APIC Using VMware vCenter document.

Modes of deployment

Two modes of deployment are supported:

- Layer 2—the AHV host, on which the virtual APIC(s) are deployed, is directly connected to the leaf switches of the ACI fabric. The AHV hosts hosting the APICs can be connected to the ACI leaf switches with active-backup uplinks.
- Layer 3—the AHV host, on which the virtual APIC(s) are deployed, is remotely attached to the ACI fabric, through an external network.

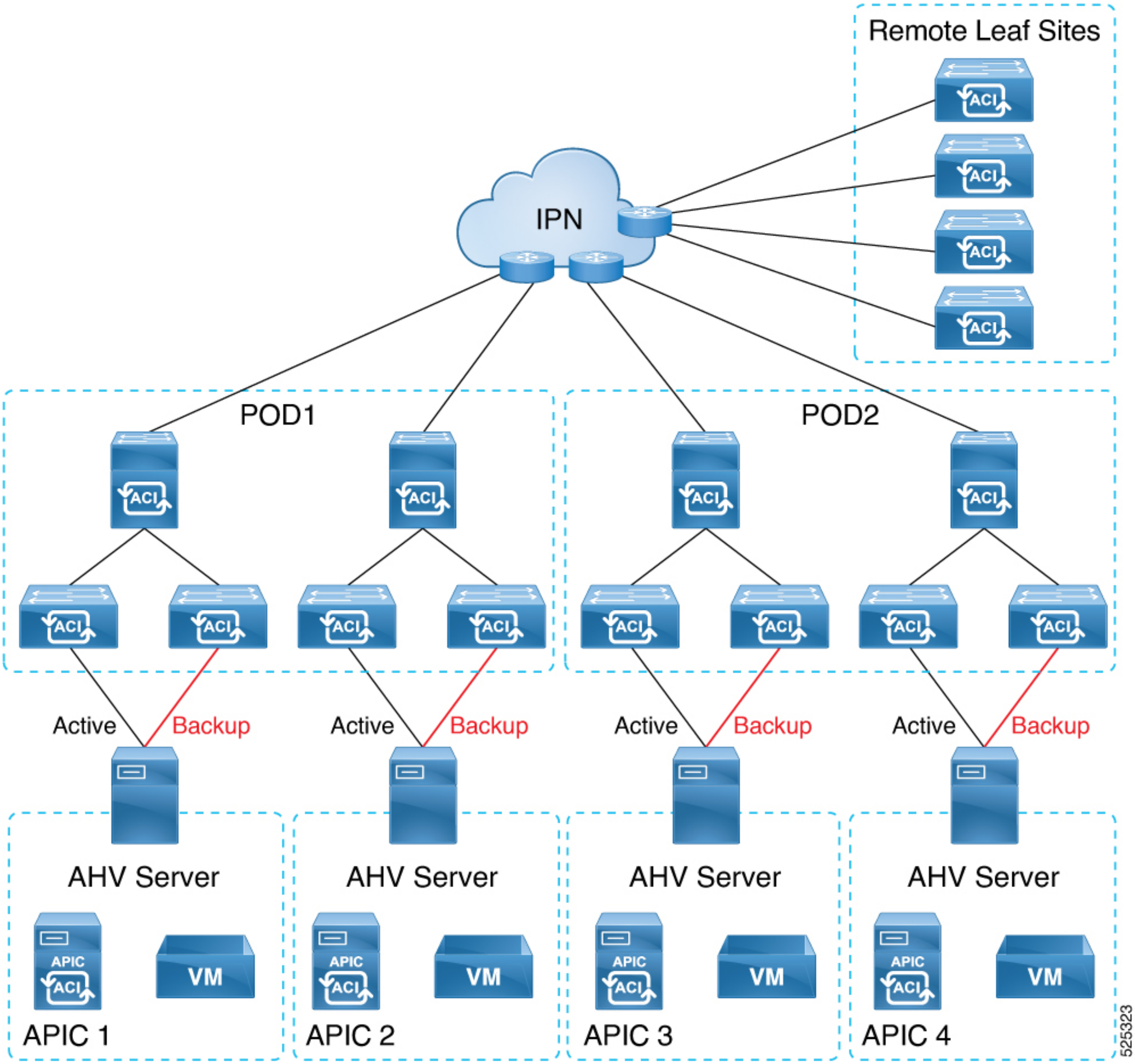
AHV host directly connected to the ACI fabric

In this mode, referred to as Layer 2 connected, wherein the AHV host is directly connected to the leaf switches of the ACI fabric. The AHV host(s) hosting APICs in Layer 2 mode must only use active-backup mode for uplinks.

Virtual APICs with active-backup AHV uplinks

As shown in this image, the APIC VMs are hosted on AHV hosts which are directly connected to the leaf switch(es) of the ACI fabric. One of the AHV host uplinks connected to the leaf switch is in active mode, the other is in backup mode.

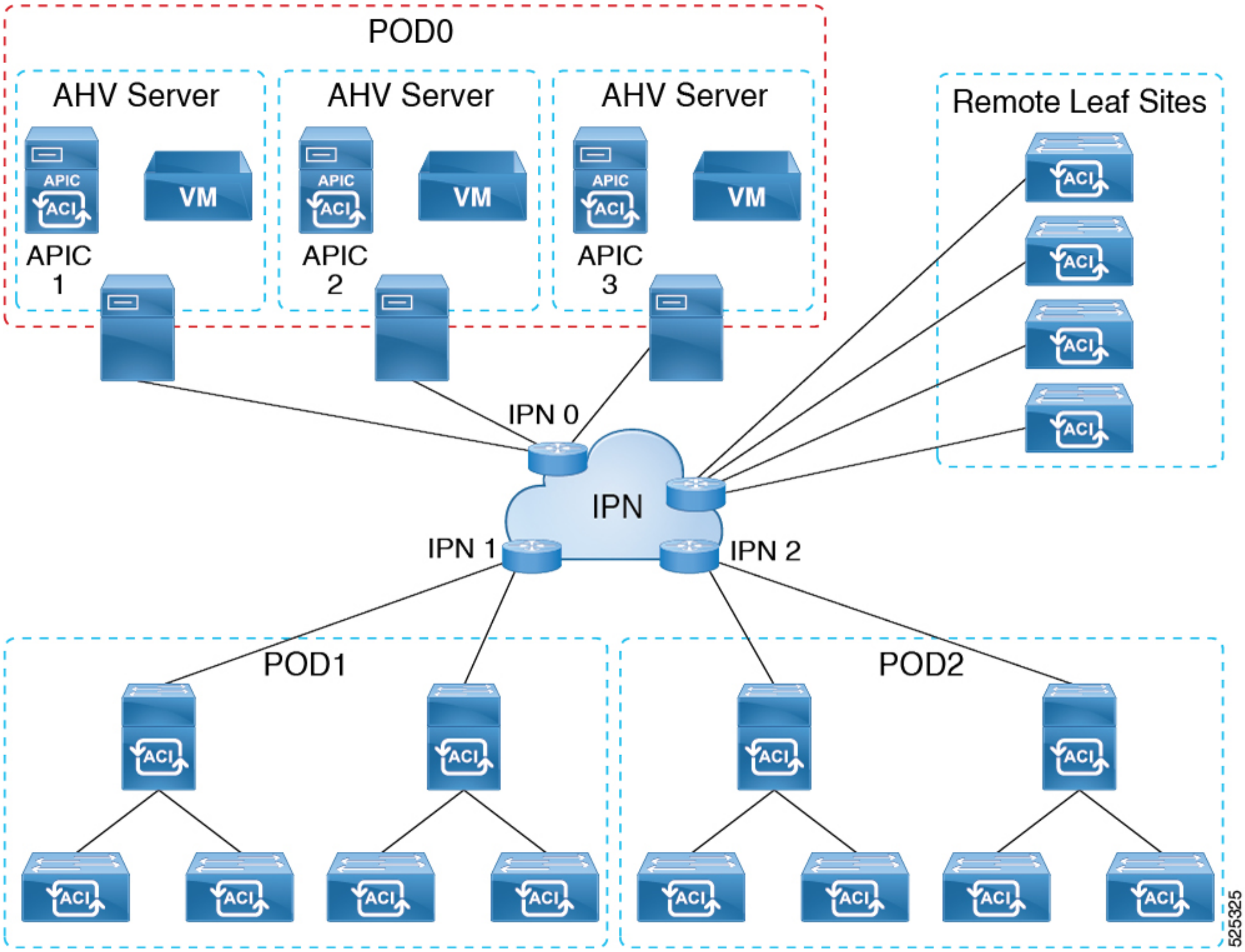
Figure 1: Virtual APICs directly connected to the ACI fabric with active-backup AHV uplinks



AHV host remotely attached to the ACI fabric

As shown in this image, the APIC VMs are hosted on the AHV hosts which are connected to an external network and remotely attached to the ACI fabric through IPN.

Figure 2: Virtual APICs remotely attached to the ACI fabric



Guidelines and limitations for deploying virtual APIC on AHV

These guidelines and limitations apply for deploying virtual APIC on AHV.

- Virtual APIC deployment on Nutanix is supported only through Nutanix Prism Central. Deployment through Nutanix Prism Element is not supported because Prism Element does not provide support for OVA deployment.
- UEFI BIOS boot configuration is not supported for virtual APIC.
- In releases prior to Cisco ACI 6.2(1), you cannot create clusters that mix physical and virtual APIC controllers. All nodes in an APIC cluster must use the same controller type—either all physical or all virtual. For example, if APIC1 and APIC2 are physical controllers, then APIC3 must also be a physical controller.

  Starting with Cisco ACI release 6.2(1), mixed mode clustering is supported based on the controller type. You can now use clusters with all virtual, all physical, or a mix of physical and virtual APICs (except AWS virtual controllers).

- Mixed-mode clusters with different PIDs are not supported. For example, a cluster combining the APIC-SERVER-NUTANIX-M1 PID and the APIC-SERVER-VMWARE-M1 PID is not allowed.
- Cluster and fabric security is provided using self-signed certificates.
- If multiple APICs are hosted on the same AHV host and fabric recovery is triggered, always begin the recovery process by starting with a cluster size of 1. After completing ID recovery, expand the cluster to 3 nodes.

Networking prerequisites

Networking prerequisites

Ensure that the network related prerequisites are met before deploying the virtual APIC on an AHV:

- **UCS hardware prerequisites:**
  - Ensure that LLDP and port-channel are disabled.
  - From vNIC interfaces, ensure that VLAN mode is set to Trunk mode.
- **Networking prerequisites:**
  - DVS refers to OVS in Nutanix and port-group refers to subnet in Nutanix.
  - VLAN configuration is not required on the subnet created on OVS.
    - If uplink is Access for management/OOB network, then VLAN 0 is configured.
    - If uplink is Trunk for management/OOB network, then the respective VLAN must be configured. For example, VLAN 100.
    - Infrastructure subnet is configured with infrastructure or inband VLANs. When infrastructure VLAN is provided, LLDP does not work as VLAN 0 is mandatory to bypass LLDP.

      NIC configuration for virtual APIC VM is updated with attachement type as **Trunked**. VLANs such as infrastructure or inband VLANs are allowed if configured.
  - Active-active uplink configuration for AHV hosts is not supported. Only active-backup uplink configuration is supported.
  - Network controller Managed subnets are not supported. Only VLAN basic subnets are supported for infrastructure subnet.
  - IPAM enabled subnet is not supported for the management and infra subnets.
  - (For Layer 2 virtual APIC) See Deploy virtual APIC using Nutanix, on page 7.

Virtual machine specifications

Ensure that the AHV hosts can support the VM specifications mentioned in these tables.

Table 1: Virtual machine specifications - medium scale (APIC-SERVER-NUTANIX-M1)

CPU	16 vCPU of 2.5 GHz or higher

Memory	96 GB of RAM
Storage	• Disk 1: SSD or NVMe – 120 GB (root disk) Disk 2: SSD or NVMe – 360 GB (empty data disk) I/O latency of 20ms
Network	Two interfaces. • OOB 1Gbps or higher • Infra or Inband 10Gbps or higher Latency tolerance between virtual APICs is up to 50 ms.

Table 2: Virtual machine specifications - small scale (APIC-SERVER-NUTANIX-S1)

CPU	8 CPU of 2.5 GHz or higher
Memory	32 GB of RAM
Storage	• Disk 1: SSD or NVMe – 120 GB (root disk) Disk 2: SSD or NVMe – 360 GB (empty data disk) I/O latency of 20ms
Network	Two interfaces: • OOB 1Gbps or higher • Infra or Inband 10Gbps or higher Latency tolerance between virtual APICs is up to 50 ms.

Table 3: Virtual machine specifications - large scale (APIC-SERVER-NUTANIX-L1)

CPU	32 vCPU of 2.5 GHz or higher
Memory	192 GB of RAM
Storage	• Disk 1: SSD or NVMe – 120 GB (root disk) Disk 2: SSD or NVMe – 900 GB (empty data disk) I/O latency of 20ms
Network	Two interfaces. • OOB 1Gbps or higher • Infra or Inband 10Gbps or higher Latency tolerance between virtual APICs is up to 50 ms.

Nutanix platform software versions

- Nutanix Prism Central - pc.2024.2.0.1
- Nutanix AOS - 6.10.1.5

Deploy virtual APIC using Nutanix

Follow these steps to deploy virtual APIC (Layer 2 and Layer 3) on an AHV host using Nutanix.

Before you begin

- See Networking prerequisites, on page 5
- See Virtual machine specifications, on page 5.

Procedure

Step 1 Upload the virtual APIC OVA file to Nutanix Prism Central.

a) In the **Compute & Storage** sub-menu on the left pane, choose **OVAs**.

The list of OVAs is displayed in the right pane.

b) Click **Upload OVA**.

The **Upload OVA** screen appears.

c) Choose a AHV cluster.
d) Choose **URL** option under **OVA Source** and enter the URL in the **OVA URL** field.
e) Click **Upload**.

The OVA file is uploaded to Nutanix Prism Central from the specified URL. The uploaded OVA file can be viewed by clicking **OVAs** in the left pane. The virtual APIC OVA file can now be used to deploy the virtual APIC.

If you do not choose the **URL** option, you can upload the OVA file through the Web browser. To do this, choose the default **OVA File** under **OVA Source**, and then choose **Select File** under **Choose OVA File**.

Step 2 Deploy the virtual APIC.

a) In the **Compute & Storage** sub-menu on the left pane, choose **OVAs**.

The list of OVAs is displayed in the right pane.

b) Right-click the downloaded OVA file and choose **Deploy as VM**.

The **Deploy as VM** wizard appears.

Step 3 In the **Configuration** screen of the wizard, specify these configurations.

a) Enter the name and optional description of the virtual APIC that will be deployed.
b) Enter the correct CPU and memory values based on your VM configurations.

The default values of CPU and Memory are provided for APIC-SERVER-NUTANIX-M1.

- APIC-SERVER-NUTANIX-S1 - Enter 8 in the CPU field and 32 in the Memory field.

- APIC-SERVER-NUTANIX-M1 - Enter 16 in the CPU field and 96 in the Memory field.
- APIC-SERVER-NUTANIX-L1 - Enter 32 in the CPU field and 192 in the Memory field.

c) Uncheck the **Enable Memory Overcommit** check box and click **Next**.

Step 4 In the **Resources** screen of the wizard, specify the resource configurations.

a) In the **Networks** area, edit the network.

The **Update NIC** dialog box appears.

b) From the **Subnet** drop-down list, choose a subnet.
c) From the **Attachment Type** drop-down list, choose the NIC mode.

The NIC mode must be **Trunked** for infrastructure network and **Access** for management network.

d) Click **Save**.
e) (Only for APIC-SERVER-NUTANIX-L1) In the **Disks** area, delete Disk2.

Disk1 is the root disk and must not be changed.

For APIC-SERVER-NUTANIX-S1 and APIC-SERVER-NUTANIX-M1, storage need not be changed.

f) (Only for APIC-SERVER-NUTANIX-L1) Click **Attach Disk** to add a new disk with capacity of 900 GB.

The **Attach Disk** dialog box appears.

g) (Only for APIC-SERVER-NUTANIX-L1) Enter 900 in the **Capacity** field and click **Save**.
h) In the **Boot Configuration** area, choose **Legacy BIOS Mode** as the boot configuration option.

UEFI BIOS mode is not supported.

i) Click **Confirm** in the **Change Boot Configuration** confirmation dialog box and click **Next**.

Step 5 In the **Management** screen of the wizard, ensure that UTC timezone is selected in the **Timezone** drop-down list and click **Next**.

Note
We recommend to deploy the vAPIC using UTC as the initial time zone and then configure the desired time zone by setting up NTP from APIC configuration settings. Otherwise, this could result in potential issues when attempting to join the vAPIC node to the vAPIC cluster. Similar issues may arise during configuration imports or ID recovery scenarios. For more information, see Nutanix KB article.

Step 6 In the **Review** screen of the wizard, review the information and click **Create VM**.

Step 7 Turn on the VM and specify admin password and OOB IP address.

a) Choose **VMs** in the left pane.

The list of VMs is displayed in the right pane.

b) Right-click the created VM and choose **Power On**.
c) After the VM is powered on, select the VM and click the **Console** tab to configure the admin password and out-of-band network.
d) Wait for the console prompt to start the initial configuration of the first virtual APIC.
e) Enter the admin password of the virtual APIC and confirm the same.
f) Enter the OOB IP address of the virtual APIC and IP address of the default gateway.

The first virtual APIC is ready for bootstrapping.

Step 8 Repeat this procedure for each virtual APIC based on the cluster size. For example, if you are building a 3-node cluster, you need to repeat this procedure for each virtual APIC deployment. Ensure that all the VMs are powered on and same admin password is provided as the first virtual APIC before proceeding to bootstrapping and cluster bringup procedure. See the detailed Bringing up the Cisco APIC Cluster Using the GUI procedure in the *Cisco APIC Getting Started Guide*. Use the IP address of APIC 1 to access the APIC Cluster Bringup GUI.

Step 9 (Only for Layer 2 virtual APIC) Perform LLDP passthrough to discover the other virtual APICs in the cluster. See Perform LLDP passthrough using Nutanix, on page 9.

Perform LLDP passthrough using Nutanix

(Only for Layer 2 virtual APIC) Follow these steps to perfom LLDP passthrough using Nutanix. LLDP passthrough ensures that LLDP packets can pass from the Layer 2 virtual APIC to the leaf switch.

Perform these steps either from any CVM or Nutanix Prism Element. Repeat this procedure for each virtual APIC based on the cluster size.

Procedure

Step 1 Use the `acli vm.get vapic_name | grep -E 'mac_addr|network_name|host_name|bios_uuid'` command to determine the vAPIC UUID from the vAPIC name, host name or host IP address from the vAPIC UUID, and MAC address and infra subnet from the vAPIC UUID.

Example:

```
nutanix@NTNX-WZP28479L64-A-CVM:10.193.147.74:~$ acli vm.get vapic1 | grep  -E
'mac_addr|network_name|host_name|bios_uuid'
      bios_uuid: "db7aaa17-6ae3-43e5-967e-402147def3cf" <---- this is vAPIC UUID ----
      mac_addr: "50:6b:8d:fe:a1:5a"
      network_name: "mgmt"
      mac_addr: "50:6b:8d:c0:72:15" <-- MAC address of infra subnet connected vnic---
      network_name: “infra_nw1” <--- Infra subnet associated to vAPIC ---
      host_name: "10.193.147.88" <--- Host IP / Host name
```

Step 2 Use the `acli -o json-pretty net.list | grep -w infra_subnet -A 5` command to retrieve the corresponding infra subnet and associated Virtual Switch names.

Example:

```
nutanix@NTNX-WZP28479L64-A-CVM:10.193.147.74:~$ acli -o json-pretty  net.list | grep -w infra_nw1 -A
 5
      "name": "infra_nw1",
      "subnet": null,
      "type": "kBridged",
      "uuid": "e5b7322c-dfaf-4c21-9543-ec8cb998f498",
      "virtual_switch": "OVS1" <-- Virtual Switch name learned from infra subnet infra_nw1---
```

Step 3 Use the `acli net.list_virtual_switch` command to retrieve the corresponding bridge name.

Example:

```
nutanix@NTNX-WZP28479L64-A-CVM:10.193.147.74:~$ acli net.list_virtual_switch
VS Name  VS UUID                               MTU   Bond Type     Bridge  IGMP Snooping
OVS1     a0c77076-636c-4551-82cd-3af4e02394b3  1500  kActiveBackup br7     Disabled  <-- br7 is the
```

```
 bridge name for infra subnet infra_nw1 which is on OVS1 virtial switch ---
OVS2     74676c08-e664-4ab1-a395-e511bf0ad482  1500  kActiveBackup  br8     Disabled
OVS3     e2a577b0-4f4e-4302-9739-b2b09e172632  1500  kNoBond        br9     Disabled
vs0      a992fcd8-2edb-47ed-9a8c-e4ad6f6be428  1500  kActiveBackup  br0     Disabled
```

Step 4 Use the `manage_ovs --bridge_name` *`bridge_name`* `show_uplinks` command to get the associated uplink interfaces from bridge names.

Example:

```
nutanix@NTNX-WZP28479L6A-A-CVM:10.193.147.74:~$ manage_ovs --bridge_name br7 show_uplinks
  Bond: br7-up
    bond_mode: active-backup
    interfaces: eth2 eth3 <--uplink to be captured for br7
    lacp: off
    lacp-fallback: false
    lacp_speed: slow
    lacp_status: off
```

Step 5 Use the `ssh root@`*`host-ip`* `virsh domiflist` *`vapic-uuid`* command (where *host-ip* is collected in step 1) to get the tap interface attached to the infra network from MAC address.

Example:

```
nutanix@NTNX-WZP28479L64-A-CVM:10.193.147.74:~$ ssh root@10.193.147.88 virsh domiflist
db7aaa17-6ae3-43e5-967e-402147def3cf
Nutanix AHV
 Interface   Type       Source   Model    MAC
-------------------------------------------------------------
 tap0        ethernet   -        virtio   50:6b:8d:fe:a1:5a
 tap2        ethernet   -        virtio   50:6b:8d:c0:72:15 <--- got tap2 from mac address --
```

Step 6 Configure these commands using the collected information such as bridge name, host IP or name, VM name, MAC address, uplinks, and tap information to perform LLDP passthrough.

```
/usr/local/nutanix/bin/acli vm.nic_update vapic_vm_name vapic_infra_subnet_name  type=kDirectNic
ssh root@host-ip "lldpcli configure ports uplink1_name,uplink2_name lldp status rx-only"
ssh root@host-ip "ovs-ofctl add-flow bridge_name
"in_port=tap_interface_name,dl_type=0x88cc,actions=output:uplink1_name,uplink2_name""
ssh root@host-ip "ovs-ofctl add-flow bridge_name
"in_port=uplink1_name,dl_type=0x88cc,actions=out:tap_interface_name""
ssh root@host-ip "ovs-ofctl add-flow bridge_name
"in_port=uplink2_name,dl_type=0x88cc,actions=output:tap_interface_name""
```

Example:

```
/usr/local/nutanix/bin/acli vm.nic_update vapic1 50:6b:8d:c0:72:15 type=kDirectNic
ssh root@10.193.147.88 "lldpcli configure ports eth2,eth3 lldp status rx-only"
ssh root@10.193.147.88 "ovs-ofctl add-flow br7 "in_port=tap2,dl_type=0x88cc,actions=output:eth2,eth3""
ssh root@10.193.147.88 "ovs-ofctl add-flow br7 "in_port=eth2,dl_type=0x88cc,actions=output:tap2””
ssh root@10.193.147.88 "ovs-ofctl add-flow br7 "in_port=eth3,dl_type=0x88cc,actions=output:tap2””
```

Note

The LLDP passthrough procedure must also be repeated in these scenarios:

- AHV maintenance
- AHV reload
- Virtual APIC reload from Prism Central or Prism Element

- Virtual APIC migration

Create an ACI network with a layer 3 connected APIC cluster

The procedures in this section establish the connectivity between the created virtual APIC cluster, and the remote ACI fabric. The layer 3 connected APIC cluster is able to discover the fabric nodes using DHCP relay and an OSPF or BGP underlay provided by the IPN.

This list outlines the steps to deploy an ACI network with a layer 3 connected virtual APIC cluster:

Procedure

Step 1 Configure the IPN as described is these procedures — Provision the fabric-facing IPN device, on page 13 and Provision the APIC cluster-facing IPN device, on page 12.

Step 2 Bring up the APIC cluster using the *Bringing up the APIC Cluster* procedure desribed in the *Cisco APIC Getting Started Guide*.

Step 3 Configure a layer 3 connection for the APIC cluster to communicate over the IPN with the fabric pod. See Prepare connectivity to the fabric pod, on page 14.

Step 4 Bring up the fabric pod. The fabric will be discovered by the APIC cluster over the layer 3 connection as described in Summary of fabric discovery and registration, on page 17.

What to do next

You can connect additional fabric pods and remote leaf sites to the layer 3 connected APIC cluster in a similar manner.

Guidelines and restrictions to deploy APIC cluster connectivity to the fabric over a layer 3 network

When deploying a virtual layer 3-connected APIC cluster, follow these guidelines and limitations.

- Ensure that the APIC connected subnet is configured for the APIC Infrastructure Layer 3 Network VLAN. If in-band management is used, the subnet must be configured as a trunk allowing both the Infrastructure Layer 3 Network VLAN and the in-band management VLAN.
- All APIC cluster sizes are supported in a layer 3 connected APIC pod.
- APICs in a layer 3 connected APIC pod cannot form a cluster with APICs within the fabric pod. In this topology, there should be no APICs in the fabric pod.
- The layer 3 connected APICs can be in the same subnet or in different subnets. In case of same subnet, configure "no ip redirects" in APIC-connected IPN interface.
- The layer 3 connected APICs can be geographically distributed from each other provided that the latency between APICs and with the fabric pod does not exceed 50 milliseconds round-trip time (RTT), which translates approximately to a geographical distance of up to 2,500 miles.

- Although any device that can meet the IPN network requirements can be used as an IPN device, we recommend to deploy, when possible, switches of the Cisco Nexus 9300 Cloud Scale family. These are the devices most commonly found in production and also the devices more frequently validated in Cisco internal testing. For further information about IPN device requirements, see "Inter-Pod Connectivity Deployment Considerations" in the ACI Multi-Pod White Paper.
- The APIC subnets must be advertised to the spines as either OSPF or BGP routes. Both OSPF and BGP are supported as underlay protocols between the APIC nodes and the IPN devices
- As all control plane traffic between the APIC cluster and the fabric pod traverses the IPN, we recommend configuring QoS for this traffic. See the Configuring QoS section in this guide.
- APIC Cluster Connectivity to the Fabric Over a Layer 3 Network does not support the following:
    - ACI CNI for Kubernetes (Redhat Openshift, SUSE/Rancher RKE, Upstream Kubernetes on Ubuntu)
    - ACI ML2 for Openstack (Redhat Openstack, Canonical Openstack)
- APIC cluster connectivity to the fabric over a Layer 3 network supports strict mode. In strict mode, you must approve the controller explicitly.

Provision the APIC cluster-facing IPN device

This section describes the configuration of the IPN device connected to Pod 0, the APIC cluster pod. With reference to the topology for Layer 3 in the AHV host remotely attached to the ACI fabric, on page 3 section, the cluster-facing IPN device is shown as IPN0. As a recommended practice, the IPN0 comprises two devices for redundancy. The fabric interface of each APIC is dual-homed to the two devices. In this configuration example, two Cisco Nexus 9000 series switches (IPN0a and IPN0b) are configured with these choices:

- VLAN 1500 is used as interface VLAN for the APICs.
- The switch interfaces are configured as layer 2 trunk ports. As an alternative, the interfaces could be access ports if the APIC fabric interface is configured to use VLAN 0 during APIC setup.
- Both switches are configured using HSRP to share a single IP address that serves as the APIC subnet default gateway address.
- APIC subnets are advertised to the spines using OSPF as the underlay protocol. As an alternative, a BGP underlay could be deployed.

```
# Example configuration of IPN0a:

interface Vlan1500
  no shutdown
  vrf member IPN
  no ip redirects
  ip address 172.16.0.252/24
  ip ospf passive-interface
  ip router ospf 1 area 0.0.0.0
  hsrp version 2
  hsrp 1500
    ip 172.16.0.1

interface Ethernet1/1
  switchport mode trunk
  switchport tru allowed vlan 1500
  spanning-tree port type edge trunk
```

```
# Example configuration of IPN0b:

interface Vlan1500
  no shutdown
  vrf member IPN
  no ip redirects
  ip address 172.16.0.253/24
  ip ospf passive-interface
  ip router ospf 1 area 0.0.0.0
  hsrp version 2
  hsrp 1500
    ip 172.16.0.1

interface Ethernet1/1
  switchport mode trunk
  switchport tru allowed vlan 1500
  spanning-tree port type edge trunk
```

Provision the fabric-facing IPN device

This section describes the configuration of the MPod IPN, which is the IPN device connected to a fabric pod. The IPN is not managed by the APIC. It must be preconfigured with this information:

- Configure the interfaces connected to the spines of the fabric pod. 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.
- Enable OSPF (or BGP) on the sub-interface specifying the OSPF process and area ID.
- Enable DHCP Relay on the IPN interfaces connected to spines.
- Enable PIM.
- Add bridge domain GIPo range as PIM Bidirectional (**bidir**) group range (default is 225.0.0.0/15).

  A group in **bidir** mode has only shared tree forwarding capabilities.
- Add 239.255.255.240/28 as PIM **bidir** group range.
- Enable PIM on the interfaces connected to all spines.

Note Multicast is not required for a single pod fabric with a layer 3-connected APIC cluster, but it is required between pods in a multi-pod fabric.

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**.

This switch configuration example is for a switch deployed as the MPod IPN. The DHCP relay configuration allows the fabric to be discovered by the APIC cluster. The deployment of a dedicated VRF in the IPN for inter-pod connectivity is optional, but is a best practice recommendation. As an alternative, you can use a global routing domain.

```
Example: OSPF as the underlay protocol
feature dhcp
feature pim
service dhcp
ip dhcp relay

# Create a new VRF.
vrf context overlay-1
  ip pim rp-address 12.1.1.1 group-list 225.0.0.0/15 bidir
  ip pim rp-address 12.1.1.1 group-list 239.255.255.240/28 bidir

interface Ethernet1/54.4    #spine connected interface
  mtu 9150
  encapsulation dot1q 4
  vrf member overlay-1
  ip address 192.168.0.1/30
  ip ospf network point-to-point
  ip router ospf infra area 0.0.0.0
  ip dhcp relay address 172.16.0.2  #infra address of APIC 1
  ip dhcp relay address 172.16.0.3  #infra address of APIC 2
  ip dhcp relay address 172.16.0.4  #infra address of APIC 3
  no shutdown

interface loopback29
  vrf member overlay-1
  ip address 12.1.1.2/30

router ospf infra
  vrf overlay-1
    router-id 29.29.29.29


Example: BGP as the underlay protocol
router bgp 65010
  vrf IPN
    neighbor 192.168.0.2 remote-as 65001
      address-family ipv4 unicast
        disable-peer-as-check
```

In the BGP configuration, the `disable-peer-as-check` command is needed for multi-pod because each pod uses the same ASN.

Prepare connectivity to the fabric pod

Before bringing up the fabric pod (Pod 1), you first must pre-configure the layer 3-connected APIC cluster (Pod 0) for connectivity through the IPN to a spine in the fabric pod. This is necessary for automatic fabric discovery.

Before you begin

- If the layer 3 conncted virtual APIC cluster is deployed in a separate security zone from the fabric, configure the firewall to allow any necessary protocols and ports.
- Configure the inter-pod network (IPN) device that is connected to the fabric pod spines.
- Configure a fabric external routing profile.
- Configure an OSPF interface policy if you are using OSPF as the underlay protocol.

Procedure

Step 1 Log in to one of the APICs in the layer 3 connected cluster.

Step 2 Choose **Fabric** > **Inventory** > **Pod Fabric Setup Policy**.

Step 3 In the work pane, click the **+** symbol in the **Pod Fabric Setup Policy** page.

The **Set Up Pod TEP Pool** dialog box opens.

Step 4 In the **Set Up Pod TEP Pool** dialog box, complete the following steps:

a) Using the **Pod ID** selector, choose Pod 1.
b) In the **TEP Pool** field, enter the TEP pool of the fabric pod.
c) Click **Submit**.

Step 5 In the navigation pane, expand **Quick Start** and click **Add Pod**.

Step 6 In the work pane, click **Add Pod**.

Step 7 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 8 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** field, enter the Node ID allocated for the spine switch that will be the initial spine to communicate with APIC 1 in Pod 0.
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 MTU should be 9150 (the default). This value should also be configured on the IPN interface.

f) Click **Next**.

Step 9 In the **Configure Interpod Connectivity STEP 3 > Routing Protocols** dialog box, in the **OSPF** area, complete the following steps to configure OSPF for the spine to IPN interface:

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** or **Regular area** (the default). **Stub area** is not supported.

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. An example is shown in the following table:

Table 4: OSPF Interface Policy Example

Property	Setting
Name	ospfIfPol
Network Type	Point-to-point
Priority	1
Cost of Interface	unspecified
Interface Controls	none checked
Hello Interval (sec)	10
Dead Interval (sec)	40
Retransmit Interval (sec)	5

Step 10 In the **Configure Interpod Connectivity STEP 3 > Routing Protocols** dialog box, in the **BGP** area, leave the **Use Defaults** checked or uncheck it.

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

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

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

e) In the **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.

   Also, enter the **Peer Address** and **Remote AS** to use for BGP underlay.

Step 11 Click **Next**.

Step 12 In the **Configure Interpod Connectivity STEP 4 > External TEP** dialog box, complete the following steps:

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

   The **Use Defaults** check box is checked by default. When the 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.

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 **Data Plane TEP IP** 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) Enter a **Router ID** for the spine node.

Step 13 Click **Next**.
The **Summary** panel appears, displaying a list of policies created by this wizard. You can change the names of these policies here.

Step 14 Click **Finish**.

What to do next

Monitor the discovery and registration of the fabric nodes by APIC, as summarized in the following section.

Summary of fabric discovery and registration

This is a summary of the switch registration and discovery process.

- The IPN, acting as a DHCP relay agent, forwards DHCP requests from the spine to the APIC.
- Spine switch is now visible to the APIC and appears the under **Nodes Pending Registration** in the **Fabric Membership** screen. Navigate to **Fabric** > **Inventory** > **Fabric Membership**.
- Register the spine switch on the APIC .
- The APIC allocates an IP address for the spine interface to the IPN and a TEP IP from the TEP pool configured for the fabric pod containing the spine. At this point, the spine joins the ACI fabric.
- The spine advertises the TEP subnet to the IPN through OSPF or BGP so that the IPN learns this subnet.
- The spine acts as a DHCP relay agent for its connected leaf switches, and forwards the requests to the APIC.
- The leaf switches are then visible to the APIC and appear in the section of **Nodes Pending Registration** in **Fabric > Inventory > Fabric Membership**.
- You must manually register these leaf switches on the APIC.
- After a leaf switch is registered, the APIC forwards the TEP IP address and DHCP configuration information to the leaf, which then joins the ACI fabric. Through this discovery process, the layer 3-connected APIC cluster discovers all switches in the fabric pod.

Configure QoS for the layer 3 connected APIC cluster

As all traffic between the APIC cluster and the fabric traverses the IPN, we recommend configuring QoS for this traffic. Specific recommendations are as follows:

Note The configuration examples in this section are for a Cisco Nexus 9000 series switch which can be used as an IPN device. The configuration may differ when using a different platform.

- Enable a DSCP class CoS translation policy.
- Retain end-to-end DCSP values for all IPN traffic. When the DSCP class CoS translation policy is enabled, the spine switches will set the DSCP value in packets sent to the IPN per this policy. Traffic destined to the APICs will use the Policy Plane class. Inter-pod control plane traffic (that is, OSPF and MP-BGP packets) will use the Control Plane class. Ensure that the Policy Plane class is configured for Expedited Forwarding (EF) and the Control Plane class is configured for CS4. The IPN network can be configured to prioritize these two classes to ensure that the policy plane and control plane remain stable during times of congestion.

  This example shows the configuration for setting QoS for policy plane traffic on a Cisco Nexus 9000 series switch used as an IPN device.

```
interface Vlan1500
  no shutdown
  vrf member overlay-1
  ip address 172.16.0.2/24
  ip ospf passive-interface
  ip router ospf 1 area 0.0.0.0
  hsrp version 2
  hsrp 100
    ip 172.16.0.1

ip access-list APIC_Cluster
  10 permit ip 172.16.0.0/24 any

class-map type qos match-all APIC-Class
  match access-group name APIC_Cluster
policy-map type qos APIC_Policy
  class APIC-Class
    set dscp 46

interface Ethernet1/1
  switchport mode trunk
  switchport access vlan 1500
service-policy type qos input APIC_Policy

interface Ethernet1/2
  switchport mode trunk
  switchport access vlan 1500
service-policy type qos input APIC_Policy

interface Ethernet1/3
  switchport mode trunk
  switchport access vlan 1500
service-policy type qos input APIC_Policy
```

- We also recommend that you rate limit non-policy-plane traffic to the APIC to prevent drops on policy plane traffic. Deploy a policer on the cluster-facing IPN to limit traffic with a DSCP value other than 46 (EF) to 4 Gbps with burst of 60 Mbps. This example shows a rate limiting configuration on a Cisco Nexus 9000 series switch used as an IPN device.

```
class-map type qos match-all no_rate_limit
  match dscp 46

policy-map type qos APIC_Rate_Limit
```

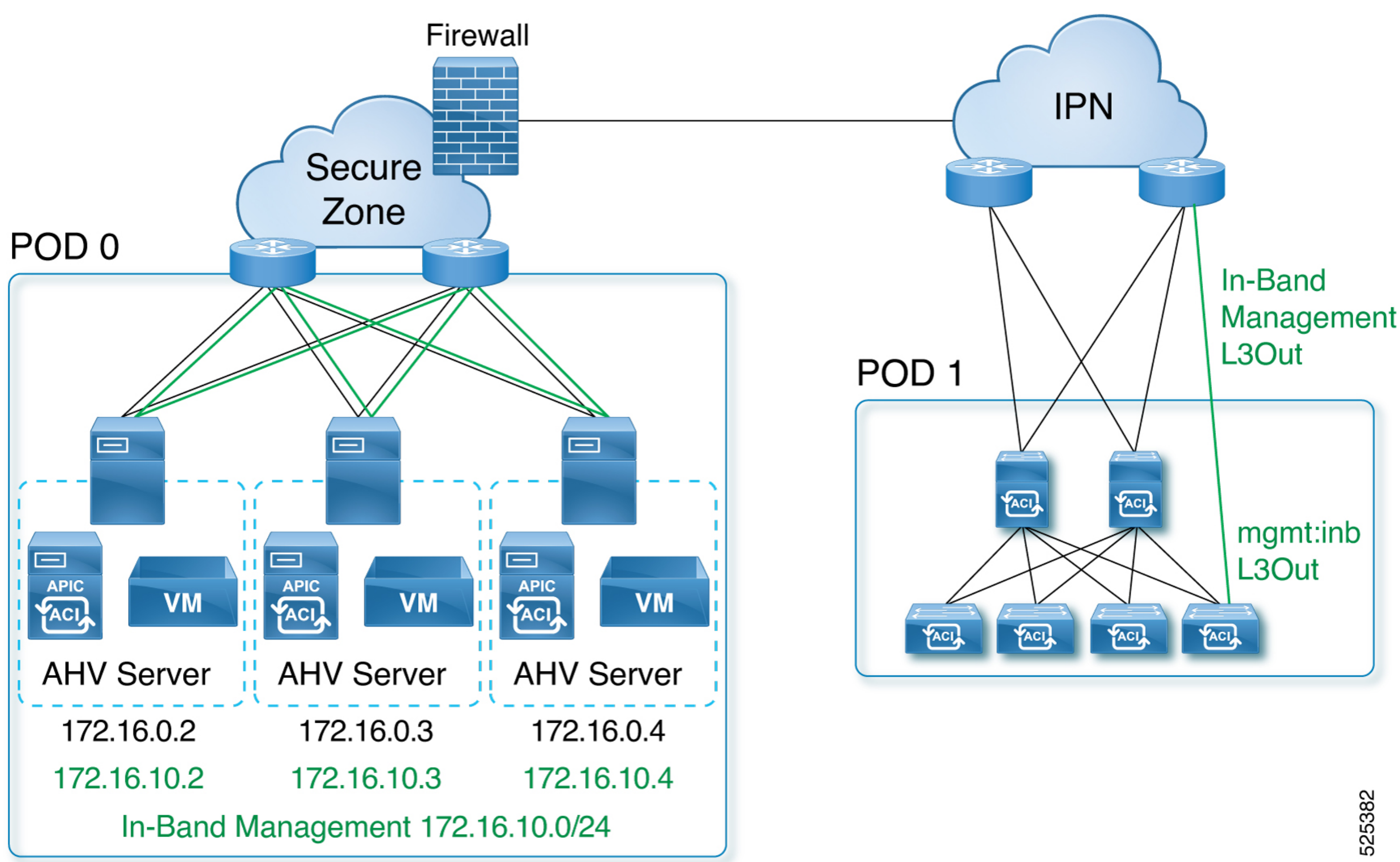
```
    class no_rate_limit
    class class-default
      police cir 4 gbps bc 7500 kbytes conform transmit violate drop

  interface Ethernet1/1-3
    switchport mode trunk
    switchport access vlan 1500
    service-policy type qos input APIC_Policy
    service-policy type qos output APIC_Rate_Limit
```

Configure in-band management

When you deploy in-band management with a layer 3 connected APIC cluster, the in-band management VRF (mgmt:inb) is not routed through the spines to the IPN. Connectivity to the in-band management interfaces on the APIC must be routed from an L3Out configured from one of the border leaf switches. This L3Out should be configured in the **mgmt** tenant for the in-band management VRF.

Figure 3: In-band management



Use this procedure to configure in-band management for the layer 3 connected APIC cluster.

Before you begin

Configure the in-band VLAN in the APIC-connected subnet in OVS.

Procedure

Step 1 Configure the APIC upstream switches.

Example:

```
interface Vlan100
  no shutdown
  vrf member IPN
  ip address 172.16.0.252/24
  ip ospf passive-interface
  ip router ospf 1 area 0.0.0.0
  hsrp version 2
  hsrp 100
    ip 172.16.0.1

interface Vlan101
  no shutdown
  vrf member IPN
  ip address 172.16.10.252/24
  ip ospf passive-interface
  ip router ospf 1 area 0.0.0.0
  hsrp version 2
  hsrp 101
    ip 172.16.10.1
```

You can route the in-band management network using the IPN VRF or you can choose a different VRF.

Step 2 Configure the APICs for in-band management using the normal in-band management configuration procedure.

a) Configure the Static Node Management Addresses for each APIC.

b) Configure the In-Band EPG. For the encapsulation VLAN, you can use any VLAN except for the infrastructure VLAN ID specified during the initial cluster bringup.

For additional information about configuring in-band management, see the "Static In-band Management" chapter of the Cisco APIC and Static Management Access tech note.

Step 3 Configure the APIC interfaces on the upstream switches.

The APIC-connected interface must be a trunk interface. If the APICs were initialized with an infrastructure VLAN other than 0, you can configure the interface as in the following example.

```
interface Ethernet1/1
  switchport mode trunk
  switchport trunk allowed vlan 100-101
```

Note
You cannot change the APIC infrastructure VLAN ID after the cluster is formed.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA 95134-1706
USA

Asia Pacific Headquarters
CiscoSystems(USA)Pte.Ltd.
Singapore

Europe Headquarters
CiscoSystemsInternationalBV
Amsterdam,TheNetherlands